Debiased Causal Mediation Analysis in Ultra-High-Dimensional Settings in the Presence of Interaction Effects

Shi Bo, AmirEmad Ghassami†, and Debarghya Mukherjee†
Department of Mathematics and Statistics, Boston University

First Version: December 11, 2024; Current Version: August 11, 2026

Abstract

Mediation analysis is a crucial tool for uncovering the mechanisms through which a treatment affects an outcome, providing deeper causal insights and guiding effective interventions. Despite substantial advances in mediation analysis with fixed- or low-dimensional mediators and covariates, estimation and inference for mediation functionals remain limited when both mediators and covariates are ultra-high-dimensional. In this paper, we propose an estimator for the mediation functional in a high-dimensional setting that accommodates the treatment–covariate interactions in the mediator model, as well as treatment–covariate and treatment–mediator interactions in the outcome model. As the parameter of interest involves high-dimensional components from different treatment arms and regression equations, existing debiasing approaches are not directly applicable, motivating our multi-step debiasing technique for handling such structurally complex functionals. We establish that the proposed estimator is $\sqrt{n}$-consistent and asymptotically normal, enabling valid inference for natural direct and indirect effects. We evaluate our proposed methodology through extensive simulation studies and apply it to the TCGA lung cancer dataset to estimate the effect of smoking, mediated by DNA methylation, on the survival time of lung cancer patients.

Keywords— Causal Mediation Analysis; Debiased Estimation; Ultra-High-Dimensional Models; Interaction Effects

† Corresponding authors.

Contents

1 Introduction **3**
- 1.1 Related Work . . . 4

2 Notations and Problem Setting **5**

3 Debiased Estimator for the Mediation Functional **7**

4 Bias Analysis and Connection to Fixed-Dimensional Setting **11**
- 4.1 Bias Analysis . . . 11
- 4.2 Connection to Fixed-Dimensional Setting . . . 11

5 Theoretical Analysis **13**
- 5.1 Assumptions . . . 14
- 5.2 Consistency and Asymptotic Normality of $\hat{\theta}_0^{\text{cf}}$ . . . 16

6 Simulation and Real Data Analysis **19**
- 6.1 Simulation Designs . . . 19
- 6.2 Real Data Experiments . . . 22

7 Conclusion **24**

A Proof of Proposition 4.1 **28**

B Proof of Results from Section 5 **29**
- B.1 Proof of Lemma 5.8 . . . 29
- B.2 Proof of Theorem 5.9 . . . 33
- B.3 Proof of Corollary 5.11 . . . 37
- B.4 Proof of Corollary 5.12 . . . 39
- B.5 Proof of Theorem 5.13 . . . 41
- B.6 Proof of Theorem 5.14 . . . 44

C Proof of Additional Lemmas **52**
- C.1 A Lemma on Sample Restricted Eigenvalue Conditions . . . 52
- C.2 Proof of Lemma B.1 . . . 53
- C.3 Proof of Lemma B.2 . . . 56
- C.4 Calculation of Effects . . . 60

D Connection with the Existing Literatures **60**

E Phenotype Features **71**

F More on Simulation Study and Sensitivity Analysis **72**

1 Introduction

The most commonly targeted quantity in causal inference is the total causal effect, which captures the effect of a treatment, action, or policy on an outcome variable of interest. Causal mediation analysis goes one step further by identifying the mechanisms through which treatment influences the outcome variable. Hence, it plays a crucial role in deepening our understanding of the intricate causal relationships within the variables of the system under study. Mediation analysis is widely used in many fields of science, including social science, behavioral science, economics, decision-making, epidemiology, and neuroscience. In the past three decades, the topic has also received much attention in the statistical literature (see, e.g., (Albert, 2008; Lindquist, 2012; Baron and Kenny, 1986; Richiardi et al., 2013; Ten Have and Joffe, 2012)).

The advancement of technology has enabled the collection of an extensive array of variables in many scientific domains, including brain imaging, genetics, epidemiology, and public health studies. Consequently, mediation analysis is increasingly carried out in high-dimensional settings, where the number of potential mediators can be comparable to or even larger than the sample size. This challenge is particularly pronounced in environmental epigenetics and other omics applications, where numerous molecular measurements may serve as mediating mechanisms. Early approaches to high-dimensional mediation analysis often relied on heuristic screening, dimension reduction, or pathway selection, with limited theoretical guarantees. Furthermore, the existing theory typically does not simultaneously address ultra-high-dimensional covariates, mediators, and treatment interactions in both the mediator and outcome models, which are important components in many real-world applications. Developing a methodology that provides valid inference for natural direct and indirect effects in this full, interaction-rich, ultra-high-dimensional setting, therefore, addresses an important gap in the literature and is the focus of the current paper.

Our contribution: In this paper, we propose a new methodology for estimating mediation functionals in the presence of ultra-high-dimensional (UHD) covariates and mediators, where the ambient dimensions may be exponentially large compared to the sample size. We consider a structural linear model in which the response depends on both UHD covariates and UHD mediators, and the mediators themselves are generated from a linear model in the UHD covariates. Unlike the majority of the existing literature on high-dimensional mediation analysis, our setup allows for interaction with binary exposure for both data-generating processes. To the best of our knowledge, this is the first work to provide a methodology for estimating direct and indirect treatment effects in the presence of both UHD covariates and mediators, along with their interaction with a binary exposure variable.

Our method of estimating the mediation functional essentially combines two high-dimensional regressions with two carefully designed balancing steps. We first regress the response on both covariates and mediators using treatment observations. Rather than estimating the entire high-dimensional relationship between the mediators and covariates, which would require a large number of high-dimensional regressions, we use control observations to construct balancing weights that directly approximate only the component of this relationship needed for estimating the parameter of interest. Because the mediation functional involves parameters from different treatment arms and different regression equations, the resulting target is not a standard linear functional of a single high-dimensional regression coefficient, nor a standard bilinear functional or inner product between two regression coefficients; it contains composite terms involving both UHD vectors and UHD matrices (where both the number of rows and columns may be exponentially large in sample size). Consequently, existing debiasing methods for linear functionals Javanmard and Montanari (2014); Van de Geer et al. (2014); Zhang and Zhang (2014) and bilinear or inner-product functionals Guo et al. (2019); Cai et al. (2021) are not directly applicable; see Remark 3.2 for more details. To overcome this difficulty, we develop a multi-stage debiasing strategy; in particular, we introduce another carefully designed high-dimensional regression, coupled with the preceding balancing step, to remove the first-order bias introduced by regularized estimation and by the intermediate approximation used in constructing the mediation functional. We show that, under fairly general sparsity, regularity, and restricted-eigenvalue-type conditions, the proposed estimator of the mediation functional (and consequently the natural direct and indirect effects) is $\sqrt{n}$-consistent and asymptotically normal. We also construct a consistent estimator of its asymptotic variance, which leads to asymptotically valid confidence intervals for the mediation functional, and for the natural direct and indirect effects.

The rest of the paper is organized as follows. After introducing the problem setting in Section 2, we describe our debiased estimation procedure in Section 3. In Section 4, we provide the bias analysis of our estimator and its connection to the fixed-dimensional setting. In Section 5, we provide a theoretical

justification for our approach under high-dimensional asymptotics. The proof of the main theorem, along with technical lemmas, can be found in the supplementary material. In Section 6, we conduct some simulation experiments and a real data experiment on the TCGA lung cancer dataset. A software implementation in R is available at UHDmedi[1]. We conclude in Section 7 and discuss future work.

1.1 Related Work

Early approaches to high-dimensional mediation analysis often relied on screening, dimension reduction, or pathway selection. For example, Zhang et al. (2016) and Perera et al. (2022) proposed to first screen mediators that are strongly associated with the outcome and then conduct standard mediation analysis on the selected low-dimensional set. Huang and Pan (2016) and Chén et al. (2018) suggested dimension-reduction approaches that map high-dimensional mediators to low-dimensional representations, while Zhang et al. (2018) developed a distance-based approach that aggregates individually weak signals. Zhao and Luo (2016) proposed a pathwise Lasso method for selecting sparse mediation pathways in the presence of high-dimensional mediators. These methods provide useful tools for empirical analysis, but none of them provide full inferential theory for natural direct and indirect effects in ultra-high-dimensional settings. More recent work has developed theoretical guarantees for high-dimensional mediation inference. In particular, Guo et al. (2022, 2023, 2024) study estimation and testing for mediation models with high-dimensional mediators and fixed-dimensional exposures or observed covariates, including extensions to generalized mediation models. However, our work differs in several key aspects; we not only allow both UHD covariates and mediators in the response generation model, but also both treatment-covariate and treatment-mediator interactions in the outcome generation model and treatment-covariate interaction in the mediator generation model, which makes our model significantly more general and makes the analysis substantially more complicated. Moreover, our theoretical analysis avoids the strong beta-min/support-recovery-type requirements used in that line of work and is established under weaker high-dimensional assumptions; see Section D of the supplementary material for more details. Lin et al. (2025) proposed a testing procedure for high-dimensional mediation effects that allows both the dimensions of mediators and exposures to grow with the sample size, but does not accommodate UHD covariates or interactions between treatment variable and covariates/mediators. Rakshit and Guo (2024) developed high-dimensional Poisson regression inference with applications to mediation analysis for count outcomes, allowing high-dimensional mediators and treatment–mediator interactions. While their framework addresses an important nonlinear outcome setting, our work studies a different source of difficulty under a linear link: the simultaneous presence of ultra-high-dimensional covariates and mediators, together with their interactions with the treatment variable, which leads to a substantially more complex parameter of interest. In addition, Jones et al. (2025) studied causal mediation analysis from a post-selection inference perspective with fixed-dimensional mediators, as opposed to the high-dimensional setup considered in this paper. Chakrabortty et al. (2018) and Shi and Li (2022) investigated high-dimensional mediation problems under sparse linear structural equation or graphical models, with a focus on causal discovery and individual or controlled mediation effects, as opposed to natural direct/indirect effects considered in this paper. Furthermore, they allow the dimension to grow polynomially with n, whereas our setup allows an exponentially large ambient dimension. Loh et al. (2022) proposed an outcome-regression-based averaging approach for estimating interventional direct/indirect effects with high-dimensional mediators, but their causal targets differ from the natural direct/indirect effects considered in our work. Moreover, their approach does not provide theoretical guarantees and relies on plug-in outcome-regression estimators (which can be potentially biased in high dimension), whereas our method develops a debiased estimator with valid asymptotic inference in the ultra-high-dimensional regime. Taken together, these studies make important contributions to high-dimensional mediation analysis, while their main focus is complementary to the one pursued in this work.

Our debiasing approach is primarily motivated by the large body of literature on debiasing techniques for estimating a contrast of high-dimensional parameters in a linear model (see, e.g., (Javanmard and Montanari, 2014; Van de Geer et al., 2014; Zhang and Zhang, 2014)) or bilinear/quadraric functionals (Guo et al., 2019; Cai et al., 2021). However, these existing tools are not directly applicable to our setting, because the mediation functional depends on nuisance components learned across different treatment arms and different structural equations, and further involves both an unknown UHD matrix and an unknown UHD vector; see Remark 3.2

[1] https://github.com/shibo769/UHDmedi/tree/main

and Section D.4 of the supplementary material for more details. The use of debiasing techniques in causal inference was also recently employed by Athey et al. (2018), where the authors developed a methodology to estimate the total effect of a binary treatment on an outcome in the presence of high-dimensional covariates. Our work is a generalization of their work towards estimating both the natural direct and indirect causal effects separately in the presence of high-dimensional mediators and interaction with the treatment variable. Related balancing/Riesz-representer-based approaches, e.g., (Hirshberg and Wager, 2021; Chernozhukov et al., 2022a; Ben-Michael et al., 2021), provide general tools for debiasing linear functionals; however, in our setting, these balancing steps are only intermediate components of a larger multi-stage debiasing procedure. Our proposed method can also be viewed as a high-dimensional counterpart of the influence function-based approach of Tchetgen Tchetgen and Shpitser (2012). In particular, we demonstrate that our estimator exhibits a second-order bias in product form, similar to the influence function-based approach in the fixed-dimensional setting, as established by Liu and Ghassami (2024). However, our framework is different from this line of work, as the framework of (Tchetgen Tchetgen and Shpitser, 2012; Liu and Ghassami, 2024) is fundamentally fixed-dimensional and relies on nuisance objects such as the conditional mediator density, which are not directly estimable in our UHD regime without severely restrictive assumptions. Kindly see Section D of the supplementary material for more details.

2 Notations and Problem Setting

For a random variable/vector X, we denote $\mathbb{E}[X]$ and Σ_X as the expectation and the variance/covariance matrix of X, respectively. For a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, we use X_i (respectively $X_{.,i}$) to denote its i^{th} row (respectively i^{th} column). We also use the notation $\|\mathbf{X}\|_{p,q}$ to denote $(\sum_{i=1}^{m} \|X_i\|_p^q)^{1/q}$, i.e. ℓ_q norm of the $\{\|X_i\|_p : 1 \leq i \leq m\}$. We also use the notation $\|\mathbf{X}\|_{\text{op}}$ and $\|\mathbf{X}\|_F$ to denote its operator norm and the Frobenius norm, respectively. The notations $\mathbf{1}_q$ and $\mathbf{0}_q$ are used to denote the vectors consisting of all ones and all zeros, respectively, in $\mathbb{R}^q$. For a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$, we say that it admits an asymptotically deterministic equivalent sequence if there exists a deterministic sequence $\{x_n\}_{n\in\mathbb{N}}$ such that $X_n - x_n = o_p(1)$.

We consider a setting with a binary treatment variable $A \in \{0, 1\}$ (also called *exposure* or *action variable*), a set of pre-treatment covariates $X = [X_1 \cdots X_p]^\top \in \mathbb{R}^p$, an outcome variable of interest, denoted by Y, and a set of post-treatment, pre-outcome variables $M = [M_1 \cdots M_q]^\top \in \mathbb{R}^q$. We consider the challenging setting where $p \wedge q \gg n$. Using the potential outcome notations (Rubin, 1974), let $Y^{(a,m)}$ denote the potential outcome of Y, had the treatment and mediator variables been set to values $A = a$ and $M = m$. Similarly, we define $M^{(a)}$ as the potential outcome of M had the treatment variables been set to the value $A = a$. Based on variables $Y^{(a,m)}$ and $M^{(a)}$, we define $Y^{(a)} = Y^{(a,M^{(a)})}$. An i.i.d. sample $(X_i, A_i, M_i, Y_i)_{i=1}^n$ of size n is given, where we assume

- $Y = Y^{(a,m)}$ if $A = a$, $M = m$, and
- $M = M^{(a)}$ if $A = a$,

which is referred to as the *consistency assumption*, requiring that the realized outcome is equal to the potential outcome corresponding to the observed treatment and mediator, and the realized mediator is equal to the potential mediator corresponding to the observed treatment. We also assume that the probability of receiving treatment and control and any value of the mediator are bounded away from zero, i.e., for some $e > 0$, we have

$$0 < e < p(a \mid x) \;\; \forall (a, x), \quad \text{and} \quad 0 < e < p(m \mid a, x) \;\; \forall (a, x, m).$$

which is referred to as the *positivity assumption*. The average treatment effect (ATE) is defined as $\mathbb{E}[Y^{(a=1)} - Y^{(a=0)}]$. This parameter captures the difference in the potential outcome mean if all units received treatment $A = 1$ and if all units received treatment $A = 0$ (e.g., placebo treatment). In mediation analysis, the focus is on the fact that this causal effect is partly mediated through the variable M, and it is of interest to quantify the direct and indirect portions of the causal effect by partitioning the ATE as follows (Robins and Greenland, 1992; Pearl, 2022).

$$\begin{aligned} \mathbb{E}\left[Y^{(1)} - Y^{(0)}\right] &= \mathbb{E}\left[Y^{\left(1,M^{(1)}\right)} - Y^{\left(0,M^{(0)}\right)}\right] \\ &= \mathbb{E}\left[Y^{\left(1,M^{(1)}\right)} - Y^{\left(1,M^{(0)}\right)}\right] + \mathbb{E}\left[Y^{\left(1,M^{(0)}\right)} - Y^{\left(0,M^{(0)}\right)}\right]. \end{aligned} \tag{2.1}$$

The first and the second terms in the last expression are called the total indirect effect and the pure direct effect, respectively by Robins and Greenland (1992), and are called the natural indirect effect (NIE) and the natural direct effect (NDE) of the treatment on the outcome, respectively, by Pearl (2022). We will use the latter terminology in this work. NIE captures the change in the expectation of the response variable in a hypothetical scenario where the value of the treatment variable is fixed at $A = 1$, while the mediator behaves as if the treatment had been changed from value 0 to 1. NDE captures the change in the expectation of the outcome in a hypothetical scenario where the value of the treatment variable is changed from value 0 to 1, while the mediator behaves as if the treatment is fixed at 0. Note that if one can identify and estimate the parameter $\mathbb{E}[Y^{(a',M^{(a)})}]$ for $a, a' \in \{0, 1\}$, then based on Equation (2.1), NIE, NDE and ATE can all be estimated. Henceforth, we focus on the parameter $\mathbb{E}[Y^{(1,M^{(0)})}]$ in this paper.

We operate under the assumption of sequential exchangeability, which implies there are no unobserved confounders for the treatment-outcome, mediator-outcome, and treatment-mediator relationships (Imai et al., 2010), formalized as follows.

Assumption 2.1 (Sequential exchangeability)**.** *Let $X_1 \perp\!\!\!\perp X_2 \mid X_3$ indicate that the random variables X_1 and X_2 are conditionally independent given the random variable X_3.*

1. $Y^{(a,m)} \perp\!\!\!\perp \{A, M\} \mid X$, *for all* a, m,
2. $M^{(a)} \perp\!\!\!\perp A \mid X$, *for all* a,
3. $Y^{(a,m)} \perp\!\!\!\perp M^{(a')} \mid X$, *for all* a, a', m.

As established in (Imai et al., 2010), under Assumption 2.1, the parameter $\mathbb{E}[Y^{(1,M^{(0)})}]$ can be identified as

$$\mathbb{E}[Y^{(1,M^{(0)})}] = \iint \mathbb{E}[Y \mid X = x, M = m, A = 1] f(m \mid X = x, A = 0) f(x) dm dx, \tag{2.2}$$

where we use f to represent densities over variables. The right hand side is known as the *mediation functional* in the literature, and will be our *parameter of interest* in this work.

The estimation of the mediation effect is a well-studied problem in statistics in both parametric and non-parametric setups. Yet, in the presence of UHD variables with both p and q larger than the sample size, the preference often leans toward simple linear models instead of intricate non-linear models due to the developed theory as well as the interpretability of the model. Hence, we focus on the following data generating mechanisms for the outcome and mediator variables:

$$\begin{aligned} Y &= (1 - A)\left(X^\top \beta_0 + M^\top \gamma_0 + \epsilon'\right) + A\left(X^\top \beta_1 + M^\top \gamma_1 + \epsilon\right), \\ M &= (1 - A)\left(\mathbf{B}_0 X + U\right) + A\left(\mathbf{B}_1 X + U'\right), \end{aligned} \tag{2.3}$$

where (ϵ, ϵ') and (U, U') are centered random shocks and the matrices $\mathbf{B}_1, \mathbf{B}_0 \in \mathbb{R}^{q \times p}$, and vectors $\beta_1, \beta_0 \in \mathbb{R}^p$, $\gamma_1, \gamma_0 \in \mathbb{R}^q$ contain unknown regression coefficients that encode the relationships among the exposures, mediators and outcome. Note that this data generating process (DGP) is flexible in the sense that

1. it allows for UHD mediators M and UHD pre-treatment covariates X,
2. it allows for $A - X$ and $A - M$ interactions (although it requires no $X - M$ interaction), and
3. it does not require any particular modeling assumptions or restrictions on the propensity score, the conditional distribution of A given X.

This DGP implies the following.

$$\mu_1(X, M) := \mathbb{E}[Y \mid X, M, A = 1] = X^\top \beta_1 + M^\top \gamma_1, \tag{2.4}$$

$$\mu_{10}(X) := \mathbb{E}[\mu_1(X, M) \mid X, A = 0] = X^\top \beta_1 + X^\top \mathbf{B}_0^\top \gamma_1, \tag{2.5}$$

for all $X \in \mathbb{R}^p, M \in \mathbb{R}^q$, where we absorbed the intercepts in the means. Therefore, the parameter of interest in Equation (2.2) reduces to

$$\mathbb{E}[\mu_{10}(X)] = \mu_X^\top (\beta_1 + \mathbf{B}_0^\top \gamma_1), \quad \mu_X = \mathbb{E}[X]. \tag{2.6}$$

Henceforth, we denote the number of treated units by n_t, and the number of control units by n_c. We use the notations $\mathbf{X}_c$, $\mathbf{X}_t$, and $\mathbf{M}_c$, $\mathbf{M}_t$ to refer to the feature matrices corresponding to control and treated units, respectively, with rows corresponding to observations from different individuals. Similar to Athey et al. (2018), we primarily focus on the sample version of the parameter of interest, which for the mediation functional, will be

$$\theta_0 \equiv \theta_{10} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mu_{10}(X_i) = \bar{X}^\top(\beta_1 + \mathbf{B}_0^\top \gamma_1), \tag{2.7}$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Note that, under Assumption 2.1, the expected values of $Y^{(1,M(1))} \equiv Y^{(1)}$ and $Y^{(0,M(0))} \equiv Y^{(0)}$ can be expressed similarly as

$$\begin{aligned}
\theta_{11} &:= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{E}[Y \mid X_i, M, A=1] \mid X_i, A=1] = \bar{X}^\top(\beta_1 + \mathbf{B}_1^\top \gamma_1), \\
\theta_{00} &:= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbb{E}[Y \mid X_i, M, A=0] \mid X_i, A=0] = \bar{X}^\top(\beta_0 + \mathbf{B}_0^\top \gamma_0).
\end{aligned} \tag{2.8}$$

Using these quantities, the (sample average) natural indirect effect (NIE) can be written as $\theta_{\text{NIE}} = \theta_{11} - \theta_{10}$ whereas the natural direct effect (NDE) can be written as $\theta_{\text{NDE}} = \theta_{10} - \theta_{00}$. We use the superscript "pop" to denote corresponding population versions, i.e. $\theta_{10}^{\text{pop}} = \mu_X^\top(\beta_1 + \mathbf{B}_0^\top\gamma_1)$, $\theta_{11}^{\text{pop}} = \mu_X^\top(\beta_1 + \mathbf{B}_1^\top\gamma_1)$ and $\theta_{00}^{\text{pop}} = \mu_X^\top(\beta_0 + \mathbf{B}_0^\top\gamma_0)$. As a consequence, we have $\theta_{\text{NDE}}^{\text{pop}} = \theta_{10}^{\text{pop}} - \theta_{00}^{\text{pop}}$ and $\theta_{\text{NIE}}^{\text{pop}} = \theta_{11}^{\text{pop}} - \theta_{10}^{\text{pop}}$. The estimation of θ_{10} (with the same estimator also applicable to θ_{10}^{pop}), is substantially more involved than that of θ_{11} and θ_{00}. This is because θ_{10} involves a cross-arm interaction between parameters learned from different treatment groups (i.e., $\mathbf{B}_0$ from the control arm and γ_1 from the treatment arm), whereas θ_{aa} depends only on parameters associated with arm a, for $a \in \{0,1\}$. Accordingly, in the next section, we first present our proposed methodology for estimating θ_{10} (which we denote by θ_0 for notational simplicity). We then describe how this estimator can be combined with estimators of $\{\theta_{aa} : a \in \{0,1\}\}$ to construct estimators of θ_{NDE} and θ_{NIE}.

3 Debiased Estimator for the Mediation Functional

In this section, we present our methodology for estimating the mediation functional $\theta_{10} = \bar{X}^\top\left(\beta_1 + \mathbf{B}_0^\top\gamma_1\right)$. A naive approach for estimating this parameter is a simple plug-in-based approach: in the first step, we estimate (β_1, γ_1) by regressing Y on (X, M) using the treatment observations, and then estimate $\mathbf{B}_0$ by regressing M on X coordinate-wise using the control observations. The resulting estimator is $\hat{\theta}^{\text{plug-in}} = \bar{X}^\top(\hat{\beta}_1 + \hat{\mathbf{B}}_0^\top\hat{\gamma}_1)$. As the setting under consideration is high-dimensional, penalized regression methods, such as the Lasso, are needed to estimate the regression parameters. However, directly plugging the resulting regularized estimators into the mediation functional generally introduces a substantial regularization bias. Hence, our goal is to devise a debiasing procedure for estimating θ_0 such that the refined estimator is $\sqrt{n}$-consistent and asymptotically normal.

To motivate our debiasing procedure, we first consider a simpler scenario in which $\mathbf{B}_0$ is known. In that case, we can debias $(\hat{\beta}_1, \hat{\gamma}_1)$ in the direction $(\bar{X}, \ \mathbf{B}_0\bar{X})$ to obtain a debiased estimator. Using tools from the literature on standard high-dimensional debiased Lasso estimators, one can show that the resulting estimator is $\sqrt{n}$-consistent and asymptotically normal. However, this debiasing procedure is not directly applicable when $\mathbf{B}_0$ is unknown. Importantly, estimation of the full UHD matrix $\mathbf{B}_0$ is not necessary for estimating θ_0; only the vector $\mathbf{B}_0\bar{X}$ enters the direction along which $(\hat{\beta}_1, \hat{\gamma}_1)$ needs to be debiased. Motivated by this observation, we directly approximate $\mathbf{B}_0\bar{X}$ with respect to the ℓ_∞ norm through a balancing procedure using the control observations. Note that $\mathbf{B}_0\bar{X}$ can be completely dense under our general setup (see Remark 3.2 for more details), in which case consistent estimation of the entire vector with respect to the ℓ_p for $p \in [1, \infty)$ norm is not possible in the UHD regime. However, as we establish later, it suffices to estimate $\mathbf{B}_0\bar{X}$ in the ℓ_∞ norm, which can be achieved under much weaker assumptions. We now present the details of our methodology, which is summarized in Algorithm 1 and Figure 1. Our method essentially comprises six key steps. We first divide the dataset into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$. Then in Step 1, using the control observations in $\mathcal{D}_1$, we construct a balancing weight $\hat{\tau}_2^{(1)}$ such that $\mathbf{X}_{c,1}^\top\hat{\tau}_2^{(1)}$ approximates $\bar{X}$ by solving optimization problem (3.1), and then set $\hat{m}_{c,1} = \mathbf{M}_{c,1}^\top\hat{\tau}_2^{(1)}$. As per our DGP (Equation (2.3)), $M_i = \mathbf{B}_0 X_i + U_i$ on the

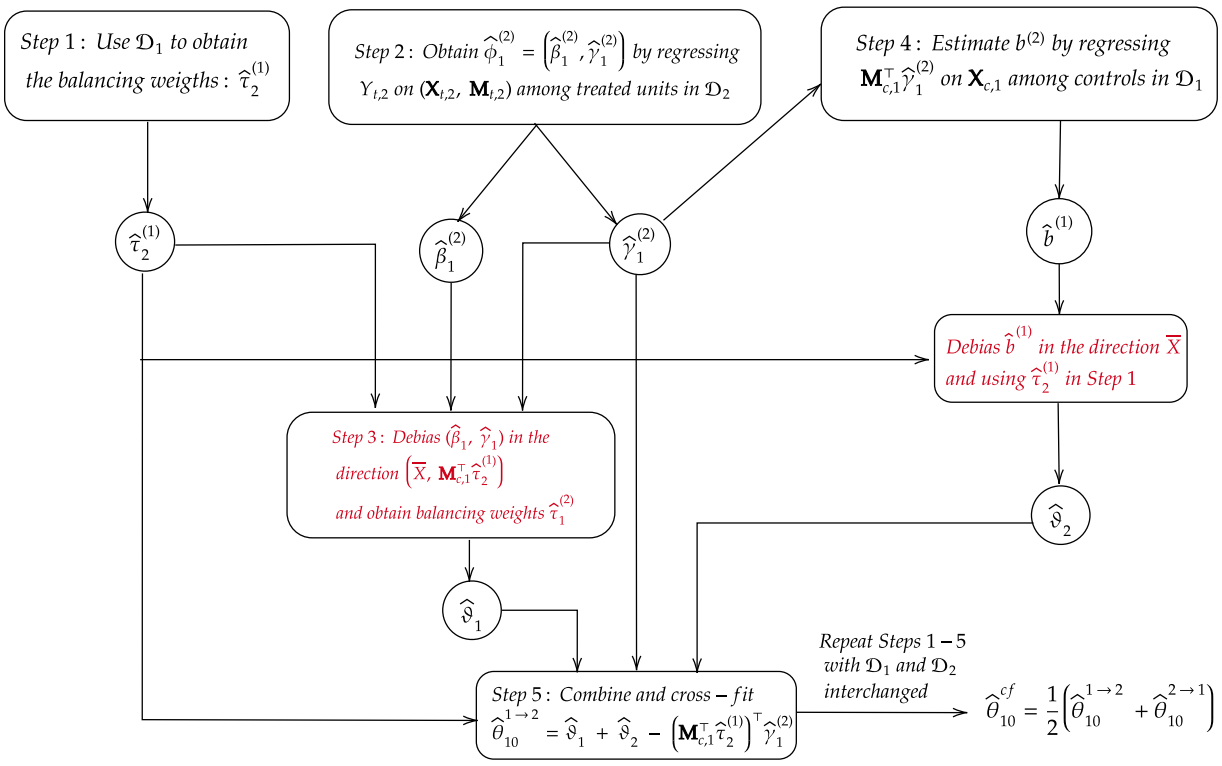


Figure 1: Schematic illustration of Algorithm 1.

control arm (i.e., $\mathbf{M}_{c,1} = \mathbf{X}_{c,1}\mathbf{B}_0^\top + \mathbf{U}_{c,1}$ in matrix form), it is immediate that the vector $\hat{m}_{c,1}$ approximates $\mathbf{B}_0\bar{X}$. In Step 2, we use the treated observations in $\mathcal{D}_2$ to estimate $\phi_1 = (\beta_1^\top \ \gamma_1^\top)^\top$ by regressing Y on (X, M) along with an ℓ_1 penalty (call this estimator $\hat{\phi}_1^{(2)} = (\hat{\beta}_1^{(2)}, \hat{\gamma}_1^{(2)})$).

In Step 3, we debias this estimator in the direction $a^{(1)} = (\bar{X}^\top \ \ \hat{m}_{c,1}^\top)^\top$ to obtain:

$$\hat{\vartheta}_1 = (a^{(1)})^\top \hat{\phi}_1^{(2)} + (\hat{\tau}_1^{(2)})^\top (\mathbf{y}_{t,2} - \mathbf{W}_{t,2}\hat{\phi}_1^{(2)})$$

Here $\hat{\tau}_1^{(2)}$ is the debiasing/correction weights obtained by solving Equation (3.2). This step can be viewed as a natural extension of the standard debiasing procedure that would be used if $\mathbf{B}_0$ were known, as in that case we would debias $\hat{\phi}_1$ in the direction $(\bar{X}^\top \ \ (\mathbf{B}_0\bar{X})^\top)^\top$. However, since $\mathbf{B}_0\bar{X}$ is unknown, we replace it by its approximation $\hat{m}_{c,1}$ and then debias $\hat{\phi}_1$ in the approximated direction $(\bar{X}^\top \ \ \hat{m}_{c,1}^\top)^\top$. However, this approximated debiasing direction introduces an additional layer of bias, because $\hat{m}_{c,1}$ is generally a biased estimator of $\mathbf{B}_0\bar{X}$. Therefore, although the correction in Step 3 reduces the regularization bias associated with $\hat{\phi}_1^{(2)}$, additional corrections are needed to remove the bias introduced by replacing $\mathbf{B}_0\bar{X}$ by $\hat{m}_{c,1}$.

Towards that end, in Step 4, we regress $M^\top \hat{\gamma}^{(2)}$ on X along with an ℓ_1 penalty using the control observations in $\mathcal{D}_1$. Since

$$\mathbb{E}[M^\top \hat{\gamma}_1^{(2)} \mid X, A = 0] = X^\top \mathbf{B}_0^\top \hat{\gamma}_1^{(2)},$$

this regression estimates $b^{(2)} = \mathbf{B}_0^\top \hat{\gamma}_1^{(2)}$. As we used the control observations in $\mathcal{D}_1$ to estimate $b^{(2)}$, we call this estimator $\hat{b}^{(1)}$. We then debias the resulting estimator $\hat{b}^{(1)}$ in the direction $\bar{X}$ using the same control-side balancing weight $\hat{\tau}_2^{(1)}$ that was used to construct $\hat{m}_{c,1}$ to obtain

$$\hat{\vartheta}_2 = \bar{X}^\top \hat{b}^{(1)} + (\hat{\tau}_2^{(1)})^\top (\mathbf{M}_{c,1}\hat{\gamma}_1^{(2)} - \mathbf{X}_{c,1}\hat{b}^{(1)})$$

This coupling (using the same $\hat{\tau}_2^{(1)}$ in Step 1 and Step 4) is crucial for removing the first-order bias introduced by the estimated direction and leads to a remaining bias of product form.

Algorithm 1 Estimation Procedure for Mediation Functional under Cross-fitting

Step 0: Randomly split the sample into two parts $(\mathcal{D}_1, \mathcal{D}_2)$ with $n/2$ samples in each.

Step 1: Use $\mathcal{D}_1$ to obtain the following balancing weights:

$$\hat{\tau}_2^{(1)} = \arg\min_\tau \left\{ \frac{1}{2}\|\tau\|_2^2 : \ \|\bar{X} - \mathbf{X}_{c,1}^\top \tau\|_\infty \le K_2 \sqrt{\frac{\log p}{n_{c,1}}}, \ \|\tau\|_\infty \le n_{c,1}^{-\alpha} \right\}. \tag{3.1}$$

Step 2: Use $\mathcal{D}_2$ to estimate $\phi_1 = (\beta_1^\top \quad \gamma_1^\top)^\top$ by regressing $\mathbf{y}_{t,2}$ on $\mathbf{W}_{t,2} = (\mathbf{X}_{t,2} \quad \mathbf{M}_{t,2})$ along with a ℓ_1 penalty

$$\hat{\phi}_1^{(2)} = ((\hat{\beta}_1^{(2)})^\top \quad (\hat{\gamma}_1^{(2)})^\top)^\top = \arg\min_\phi \left\{ \frac{1}{2n_{t,2}} \|\mathbf{y}_{t,2} - \mathbf{W}_{t,2}\phi\|_2^2 + \lambda_{n,1}\|\phi\|_1 \right\}.$$

Step 3: Compute bias correction weights with contrast $a^{(1)} = \left(\bar{X}^\top \quad \hat{m}_{c,1}^\top\right)^\top$, where $\hat{m}_{c,1} = \mathbf{M}_{c,1}^\top \hat{\tau}_2^{(1)}$ and tuning parameter K_1 using $\mathcal{D}_2$,

$$\hat{\tau}_1^{(2)} = \operatorname{argmin}_\tau \left\{ \frac{1}{2}\|\tau\|_2^2 : \ \ \|a^{(1)} - \mathbf{W}_{t,2}^\top \tau\|_\infty \leqslant K_1 \sqrt{\frac{\log(p+q)}{n_{t,2}}}, \ \ \|\tau\|_\infty \le n_{t,2}^{-\alpha} \right\}, \tag{3.2}$$

and set $\hat{\vartheta}_1 = (a^{(1)})^\top \hat{\phi}_1^{(2)} + (\hat{\tau}_1^{(2)})^\top (\mathbf{y}_{t,2} - \mathbf{W}_{t,2}\hat{\phi}_1^{(2)})$.

Step 4: On $\mathcal{D}_1$, approximate $b^{(2)} = \mathbf{B}_0^\top \hat{\gamma}_1^{(2)}$ by regressing $\mathbf{M}_{c,1}\hat{\gamma}_1^{(2)}$ on $\mathbf{X}_{c,1}$ along with a ℓ_1 penalty:

$$\hat{b}^{(1)} = \arg\min_b \left\{ \frac{1}{2n_{c,1}} \|\mathbf{M}_{c,1}\hat{\gamma}_1^{(2)} - \mathbf{X}_{c,1} b\|_2^2 + \lambda_{n,2}\|b\|_1 \right\}.$$

Set $\hat{\vartheta}_2 = \bar{X}^\top \hat{b}^{(1)} + (\hat{\tau}_2^{(1)})^\top (\mathbf{M}_{c,1}\hat{\gamma}_1^{(2)} - \mathbf{X}_{c,1}\hat{b}^{(1)})$.

Step 5: Compute the estimator as: $\hat{\theta}_{10}^{1\to 2} = \hat{\vartheta}_1 + \hat{\vartheta}_2 - \hat{m}_{c,1}^\top \hat{\gamma}_1^{(2)}$.

Step 6: Interchange the role of $\mathcal{D}_1$ and $\mathcal{D}_2$ and perform Step 1 – 5 to obtain analogous estimator $\hat{\theta}_{10}^{2\to 1}$.

Return: The final cross-fitted estimator $\hat{\theta}_{10}^{\text{cf}} = (\hat{\theta}_{10}^{1\to 2} + \hat{\theta}_{10}^{2\to 1})/2$.

In Step 5, we combine the two intermediate estimators to obtain a fold-specific estimator of θ_{10}. Thus, the estimator consists of a regression-based initial estimator together with two correction terms constructed using the treatment and control observations, respectively. As will be established in Proposition 4.1, the use of the same balancing weight $\hat{\tau}_2^{(1)}$ in the construction of $\hat{m}_{c,1}$ and in the second correction term ensures that the remaining bias has a product form and is asymptotically negligible. Finally, in Step 6, we interchange the roles of $\mathcal{D}_1$ and $\mathcal{D}_2$ and average the two fold-specific estimators to obtain the final cross-fitted estimator $\hat{\theta}_{10}^{\text{cf}}$.

Estimation of NDE and NIE: So far, we have described our procedure for estimating θ_{10}, as defined in Equation (2.7). As mentioned previously, estimating θ_{00} and θ_{11} is simpler than estimating θ_{10}, since these parameters do not involve quantities from different treatment arms. For any $a \in \{0,1\}$, recall that $\theta_{aa} = \bar{X}^\top(\beta_a + \mathbf{B}_a^\top \gamma_a)$ (Equation (2.8)). Moreover, it follows directly from the DGP (Equation (2.3)) that $\mathbb{E}[Y \mid X, A = a] = X^\top(\beta_a + \mathbf{B}_a^\top \gamma_a)$.

This representation naturally suggests a simple Lasso-based debiasing procedure for estimating θ_{aa}. Specifically, using the observations from treatment arm a, we regress Y on X to estimate the reduced-form coefficient $\beta_a + \mathbf{B}_a^\top \gamma_a$, and then debias the resulting estimator in the direction of $\bar{X}$. To ensure compatibility with our estimator of θ_{10}, we use an analogous data-splitting and cross-fitting procedure to construct $\hat{\theta}_{aa}^{\text{cf}}$ for each $a \in \{0,1\}$. We then set our estimators for NDE and NIE as $\hat{\theta}_{\text{NDE}}^{\text{cf}} = \hat{\theta}_{10}^{\text{cf}} - \hat{\theta}_{00}^{\text{cf}}$ and $\hat{\theta}_{\text{NIE}}^{\text{cf}} = \hat{\theta}_{11}^{\text{cf}} - \hat{\theta}_{10}^{\text{cf}}$.

Algorithm 2 Estimation Procedure for NDE

Step 0: Start with same split $\mathcal{D}_1$ and $\mathcal{D}_2$ as in Algorithm 1.

Step 1: Regress $\mathbf{y}_{c,1}$ on $\mathbf{X}_{c,1}$ using $\mathcal{D}_1$:

$$\hat{\rho}_c^{(1)} = \arg\min_\rho \left\{ \tfrac{1}{n_{c,1}} \|\mathbf{y}_{c,1} - \mathbf{X}_{c,1}\rho\|_2^2 + \lambda_{n,3}\|\rho\|_1 \right\}$$

Step 2: Use $\hat{\tau}_2^{(1)}$ obtained from the Step 1 of Algorithm 1 to obtain:

$$\hat{\theta}_{00}^{(1)} = \bar{X}^\top \hat{\rho}_c^{(1)} + (\hat{\tau}_2^{(1)})^\top (\mathbf{y}_{c,1} - \mathbf{X}_{c,1}\hat{\rho}_c^{(1)}) .$$

Step 3: Swap the role of the dataset; use the control observations from $\mathcal{D}_2$ to obtain $\hat{\rho}_c^{(2)}$ using Step 1 with $(\mathbf{y}_{c,2}, \mathbf{X}_{c,2})$ and use $\hat{\tau}_2^{(2)}$ obtained in Step 6 of Algorithm 1 to obtain $\hat{\theta}_{00}^{(2)}$.

Return: The estimator $\hat{\theta}^{\text{cf}}_{\text{NDE}} = \hat{\theta}^{\text{cf}}_{10} - (\hat{\theta}_{00}^{(1)} + \hat{\theta}_{00}^{(2)})/2$, where $\hat{\theta}^{\text{cf}}_{10}$ is obtained from Algorithm 1.

Algorithm 3 Estimation Procedure for NIE

Step 0: Start with same split $\mathcal{D}_1$ and $\mathcal{D}_2$ as in Algorithm 1.

Step 1: On $\mathcal{D}_1$ use the treatment observation; regress $\mathbf{y}_{t,1}$ on $\mathbf{X}_{t,1}$ to obtain $\hat{\rho}_t^{(1)}$:

$$\hat{\rho}_t^{(1)} = \arg\min_\rho \left\{ \tfrac{1}{n_{t,1}} \|\mathbf{y}_{t,1} - \mathbf{X}_{t,1}\rho\|_2^2 + \lambda_{n,4}\|\rho\|_1 \right\}$$

Step 2: Construct the balancing weight for debiasing:

$$\hat{\omega}^{(1)} = \arg\min \left\{ \tfrac{1}{2}\|\omega\|_2^2, \quad \|\bar{X} - \mathbf{X}_{t,1}^\top \omega\|_\infty \le K_3 \sqrt{\tfrac{\log p}{n_{t,1}}}, \quad \|\omega\|_\infty \le n_{t,1}^{-\alpha} \right\} \tag{3.3}$$

and set $\hat{\theta}_{11}^{(1)}$ as:

$$\hat{\theta}_{11}^{(1)} = \bar{X}^\top \hat{\rho}_t^{(1)} + (\hat{\omega}^{(1)})^\top (\mathbf{y}_{t,1} - \mathbf{X}_{t,1}\hat{\rho}_t^{(1)}) .$$

Step 3: Apply Step 1 and Step 2 to the treatment observations in $\mathcal{D}_2$ to obtain $\hat{\theta}_{11}^{(2)}$ analogously.

Return: The estimator $\hat{\theta}^{\text{cf}}_{\text{NIE}} = (\hat{\theta}_{11}^{(1)} + \hat{\theta}_{11}^{(2)})/2 - \hat{\theta}^{\text{cf}}_{10}$, where $\hat{\theta}^{\text{cf}}_{10}$ is obtained from Algorithm 1.

The complete estimation procedures for the NDE and NIE are summarized in Algorithms 2 and 3, respectively. A detailed theoretical analysis of these estimators is provided in Section B.5 and Section B.6 of the supplementary document, where we establish that $\hat{\theta}^{\text{cf}}_{\text{NDE}}$ and $\hat{\theta}^{\text{cf}}_{\text{NIE}}$ are $\sqrt{n}$-CAN estimators of θ_{NDE} and θ_{NIE} (as well as their population counterparts), respectively.

Remark 3.1. *The choice of α in Equation (3.1) and (3.2) of Algorithm 1 and in Equation (3.3) of Algorithm 3 is quite flexible; one can choose any $\alpha \in (1/2, 1)$ as long as $n^{1-\alpha} \gg \sqrt{\log(n(p+q))}$. The condition $\alpha < 1$ is sufficient to guarantee feasibility of the optimization problems with high probability, while the lower bound $\alpha > 1/2$ is needed in the asymptotic normality argument (see the proof of Lemma 5.8 and Theorem 5.9 for more details).*

Remark 3.2 (Relation to debiased inference for high-dimensional functionals)**.** *Our debiasing technique, as presented in Algorithm 1, is structurally distinct from existing debiasing approaches for high-dimensional linear and bilinear functionals. In particular, (Cai et al., 2021; Javanmard and Montanari, 2014; Van de Geer et al., 2014; Zhang and Zhang, 2014) study inference for linear functionals of the form $x^\top \beta$, while Guo et al. (2019) develops inference for quadratic and bilinear functionals such as $\beta^\top \gamma$ in high-dimensional linear models. In contrast, our parameter of interest contains the term $\bar{X}^\top \mathbf{B}_0^\top \gamma_1 = (\mathbf{B}_0 \bar{X})^\top \gamma_1$, which involves a UHD matrix $\mathbf{B}_0$ and a UHD vector γ_1 that are associated with the control and treatment arms, respectively.*

Moreover, although $\mathbf{B}_0$ *is assumed to be row-sparse, the vector* $\mathbf{B}_0\bar{X}$ *need not be sparse and can be completely dense under our general setup. Consequently, without imposing additional structural assumptions,* $\mathbf{B}_0\bar{X}$ *cannot generally be consistently estimated as a full UHD vector with respect to the* ℓ_1/ℓ_2 *norm, and consequently the method of Guo et al. (2019) cannot be directly applied. Our procedure simply obtains an approximation of* $\mathbf{B}_0\bar{X}$ *with respect to the* ℓ_∞ *norm along with careful debiasing of the bias introduced by this approximation. Therefore, our problem cannot be reduced to estimating an inner product between two separately* ℓ_2*-consistently estimable UHD vectors. This, together with the fact that the two components are associated with different treatment arms and different regression equations, prevents the techniques developed in Guo et al. (2019) from being directly applicable. A detailed explanation is provided in Section D.4 of the supplementary document.*

4 Bias Analysis and Connection to Fixed-Dimensional Setting

4.1 Bias Analysis

In this section, we present the bias structure of the debiased estimator $\hat{\theta}_{10}^{\text{cf}}$ and show that the bias is asymptotically negligible, i.e., it decays at a faster rate than $n^{-1/2}$. We would like to point out that *debiasing* a high-dimensional parameter does not make it *unbiased*, i.e., it does not remove the bias completely, but rather reduces the bias sufficiently to achieve $\sqrt{n}$-consistency (e.g., see (Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014)). Recall the definition of $b = \mathbf{B}_0^\top\hat{\gamma}_1$, $\hat{\phi}_1 = (\hat{\beta}_1, \hat{\gamma}_1)$ and $a = (\bar{X}^\top \;\; \hat{m}_c^\top)^\top$. For notational simplicity, throughout this subsection, we suppress the superscript/subscript j used in the algorithms to index the data fold, with all quantities understood to correspond to a generic fold. The following proposition establishes the exact bias structure of $\hat{\theta}_{10}^{\text{cf}}$.

Proposition 4.1. *The estimation error of* $\hat{\theta}_{10}^{\text{cf}}$*, obtained via Algorithm 1 can be decomposed as* $\hat{\theta}_{10}^{\text{cf}} - \theta_{10} = \Delta_n^{\text{cf}} + V_n^{\text{cf}}$ *where*

$$V_n^{\text{cf}} = \tfrac{1}{2}\textstyle\sum_{j=1}^{2}\left[\sum_{i\in\mathcal{D}_j, A_i=1}\hat{\tau}_{1,i}^{(j)}\epsilon_i + \sum_{i\in\mathcal{D}_j, A_i=0}\hat{\tau}_{2,i}^{(j)}U_i^\top\gamma_1\right],$$

and $\Delta_n^{\text{cf}} = (\Delta_n^{(1)} + \Delta_n^{(2)})/2$ *with*

$$\begin{aligned}\Delta_n^{(j)} = {} & (a^{(j)} - \mathbf{W}_{t,j'}^\top\hat{\tau}_1^{(j')})^\top(\hat{\phi}^{(j')} - \phi_1) + (\bar{X} - \mathbf{X}_{c,j}^\top\hat{\tau}_2^{(j)})^\top(\hat{b}^{(j)} - b^{(j')}) \\ & + (\bar{X} - \mathbf{X}_{c,j}^\top\hat{\tau}_2^{(j)})^\top\mathbf{B}_0^\top(\hat{\gamma}_1^{(j')} - \gamma_1)\end{aligned} \tag{4.1}$$

for $j \in \{1,2\}$ *and* $j' = 3 - j$*. Consequently, we have, for* $j \in \{1,2\}$*:*

$$\begin{aligned}|\Delta_n^{(j)}| \le {} & \|\mathbf{B}_0(\bar{X} - \mathbf{X}_{c,j}^\top\hat{\tau}_2^{(j)})\|_\infty\|\hat{\gamma}_1^{(j')} - \gamma_1\|_1 + \|a^{(j)} - \mathbf{W}_{t,j'}^\top\hat{\tau}_1^{(j')}\|_\infty\|\hat{\phi}_1^{(j')} - \phi_1\|_1 \\ & + \|\bar{X} - \mathbf{X}_{c,j}^\top\hat{\tau}_2^{(j)}\|_\infty\|\hat{b}^{(j)} - b^{(j')}\|_1\,.\end{aligned} \tag{4.2}$$

The proof can be found in Section A of the supplementary material. Observe that each of the three terms in the upper bound of $|\Delta_n^{(j)}|$ has a *product form*, i.e., it is the product of an estimation error and an approximation error. More specifically, the first summand is the product of the estimation error of γ_1 and the error in approximating $\bar{X}$ by $\mathbf{X}_c^\top\tau_2$. The second summand is the product of the estimation error of ϕ_1 and the approximation error of a by $\mathbf{W}_t^\top\tau_1$, while the last summand is the product of the estimation error of b and the approximation error of $\bar{X}$ by $\mathbf{X}_c^\top\tau_2$. Although each individual estimation or approximation error can be significantly larger than $n^{-1/2}$ in order, the product form makes each summand asymptotically negligible. In Section 5, we provide sufficient conditions on the sparsity of the parameters and the distributions of the covariates and error terms under which $|\Delta_n^{\text{cf}}| = o_p(n^{-1/2})$ and $\sqrt{n}V_n^{\text{cf}}$ converges to a normal distribution. This establishes that $\hat{\theta}_0^{\text{cf}}$ is a $\sqrt{n}$-CAN estimator of θ_0.

4.2 Connection to Fixed-Dimensional Setting

In a fixed-dimensional setting, i.e., when the number of covariates and potential mediators is fixed, Tchetgen Tchetgen and Shpitser (2012) provided a debiased estimator of the mediation functional constructed

based on the influence function of the parameter of interest. Our proposed estimator can be viewed as a counterpart of their approach in the UHD setting. The debiased estimator proposed by Tchetgen Tchetgen and Shpitser (2012) can be written as

$$\hat{\theta}^{\mathrm{IF}} = \mathbb{E}_n\Bigg[\hat{\mu}_{10}(X) + \frac{A}{\hat{p}(A=1 \mid X)}\frac{\hat{p}(M \mid A=0,X)}{\hat{p}(M \mid A=1,X)}\Big(Y - \hat{\mu}_1(X,M)\Big) + \frac{1-A}{1-\hat{p}(A=1 \mid X)}\Big(\hat{\mu}_1(X,M) - \hat{\mu}_{10}(X)\Big)\Bigg], \tag{4.3}$$

where $\hat{\mu}_{10}(X) = \int \hat{\mu}_1(X,M)\hat{p}(M \mid A=0,X)dM$. We now show that our estimator $\hat{\theta}^{\mathrm{cf}}_{10}$, obtained via Algorithm 1, is structurally similar to $\hat{\theta}^{\mathrm{IF}}$ when the underlying data-generating process is linear. Consider the nuisance-function estimators $\hat{\mu}_1(X,M) = X^\top\hat{\beta}_1 + M^\top\hat{\gamma}_1$ and $\hat{\mu}_{10}(X) = X^\top(\hat{\beta}_1 + \hat{b})$, where $\hat{b}$ estimates $b = \mathbf{B}_0^\top\hat{\gamma}_1$. Then the estimator $\hat{\theta}^{\mathrm{IF}}$ takes the form

$$\begin{aligned}
\hat{\theta}^{\mathrm{IF}} &= \mathbb{E}_n\Bigg[X^\top\Big(\hat{\beta}_1 + \hat{b}\Big) + \frac{A}{\hat{p}(A=1 \mid X)}\frac{\hat{p}(M \mid A=0,X)}{\hat{p}(M \mid A=1,X)}\Big(Y - X^\top\hat{\beta}_1 - M^\top\hat{\gamma}_1\Big) \\
&\quad + \frac{1-A}{1-\hat{p}(A=1 \mid X)}\Big(X^\top\hat{\beta}_1 + M^\top\hat{\gamma}_1 - X^\top\Big(\hat{\beta}_1 + \hat{b}\Big)\Big)\Bigg] \\
&= \mathbb{E}_n\Bigg[X^\top\Big(\hat{\beta}_1 + \hat{b}\Big) + \omega_1 A\Big(Y - X^\top\hat{\beta}_1 - M^\top\hat{\gamma}_1\Big) + \omega_2(1-A)\Big(M^\top\hat{\gamma}_1 - X^\top\hat{b}\Big)\Bigg] \\
&= \bar{X}^\top\Big(\hat{\beta}_1 + \hat{b}\Big) + \frac{1}{n}\sum_{\{i:A_i=1\}}\omega_{1,i}\Big(Y_i - X_i^\top\hat{\beta}_1 - M_i^\top\hat{\gamma}_1\Big) \\
&\qquad + \frac{1}{n}\sum_{\{i:A_i=0\}}\omega_{2,i}\Big(M_i^\top\hat{\gamma}_1 - X_i^\top\hat{b}\Big),
\end{aligned} \tag{4.4}$$

where

$$\omega_{1,i} = \frac{1}{\hat{p}(A_i=1 \mid X_i)}\frac{\hat{p}(M_i \mid A_i=0,X_i)}{\hat{p}(M_i \mid A_i=1,X_i)}, \quad \omega_{2,i} = \frac{1}{1-\hat{p}(A_i=1 \mid X_i)}.$$

As can be seen, the estimator $\hat{\theta}^{\mathrm{IF}}$ consists of a regression-based initial estimator together with two weighted terms for bias correction. This is the same structure as our proposed estimator. For simplicity, consider the fold-specific estimator $\hat{\theta}^{1\to 2}_{10}$, suppressing the fold indices. Using $\hat{m}_c = \mathbf{M}_c^\top\tau_2$, this estimator in Algorithm 1 can be decomposed as

$$\begin{aligned}
\hat{\theta}^{1\to 2}_{10} &= \bar{X}^\top(\hat{\beta}_1 + \hat{b}) + \sum_{\{i:A_i=1\}}\hat{\tau}_{1,i}(Y_i - X_i^\top\hat{\beta}_1 - M_i^\top\hat{\gamma}_1) \\
&\qquad + \sum_{\{i:A_i=0\}}\hat{\tau}_{2,i}(M_i^\top\hat{\gamma}_1 - X_i^\top\hat{b}).
\end{aligned} \tag{4.5}$$

It is evident from Equations (4.4) and (4.5) that our estimator has the same general form as $\hat{\theta}^{\mathrm{IF}}$, with $\hat{\tau}_1$ and $\hat{\tau}_2$ playing the roles of the weights ω_1/n and ω_2/n, respectively. Consider the nuisance functions

$$\pi_1(X) = 1/p(A=0 \mid X), \qquad \pi_2(M,X) = p(M \mid A=0,X)/p(M,A=1 \mid X),$$

and recall that $\mu_1(X,M) = \mathbb{E}[Y \mid X,M,A=1]$ and $\mu_{10}(X) = \mathbb{E}[\mu_1(X,M) \mid X,A=0]$. In a recent work, Liu and Ghassami (2024) showed that, writing the estimator in Equation (4.3) as

$$\begin{aligned}
\hat{\theta}^{mr} = \mathbb{E}_n\big[&\hat{\mu}_{10}(X) + (1-A)\hat{\pi}_1(X)\{\hat{\mu}_1(M,X) - \hat{\mu}_{10}(X)\} \\
&+ A\hat{\pi}_2(M,X)\{Y - \hat{\mu}_1(M,X)\}\big],
\end{aligned}$$

the bias has the following form:

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}^{mr}\right] - \theta_0 = \mathbb{E}[&\{(1-A)\hat{\pi}_1(X) - 1\}\{\mu_{10}(X) - \hat{\mu}_{10}(X)\} \\
&+ \{A\hat{\pi}_2(M,X) - (1-A)\hat{\pi}_1(X)\}\{\mu_1(M,X) - \hat{\mu}_1(M,X)\}].
\end{aligned} \tag{4.6}$$

In the fixed-dimensional setting, by Hölder's inequality and plugging in $\hat{\mu}_1(X, M) = X^\top \hat{\beta}_1 + M^\top \hat{\gamma}_1$ and $\hat{\mu}_{10}(X) = X^\top(\hat{\beta}_1 + \hat{b})$, Equation (4.6) can be upper bounded by

$$\begin{aligned}\left|\mathbb{E}[\hat{\theta}^{mr}] - \theta_0\right| \leq\ & \mathbb{E}\left[\left\|X^\top\left\{A\hat{\pi}_2(M,X) - 1\right\}\right\|_\infty \left\|\beta_1 - \hat{\beta}_1\right\|_1\right] \\ & + \mathbb{E}\Big[\Big\{\left\|M^\top\left\{A\hat{\pi}_2(M,X) - (1-A)\hat{\pi}_1(X)\right\}\right\|_\infty \\ & \qquad + \left\|(\mathbf{B}_0 X)^\top\left\{(1-A)\hat{\pi}_1(X) - 1\right\}\right\|_\infty\Big\}\left\|\gamma_1 - \hat{\gamma}_1\right\|_1\Big] \\ & + \mathbb{E}\left[\left\|X^\top\left\{(1-A)\hat{\pi}_1(X) - 1\right\}\right\|_\infty \left\|\mathbf{B}_0^\top\hat{\gamma}_1 - \hat{b}\right\|_1\right].\end{aligned}$$

Coming back to our estimator $\hat{\theta}_{10}^{1\to 2}$, the remaining bias $\Delta_{n,1}$, as shown in Proposition 4.1, is bounded by

$$\begin{aligned}|\Delta_n| &\leq \left\|a - \mathbf{W}_t^\top\tau_1\right\|_\infty \left\|\hat{\phi}_1 - \phi_1\right\|_1 + \left\|\mathbf{B}_0\left(\bar{X} - \mathbf{X}_c^\top\tau_2\right)\right\|_\infty \left\|\hat{\gamma}_1 - \gamma_1\right\|_1 \\ &\quad + \left\|\bar{X} - \mathbf{X}_c^\top\tau_2\right\|_\infty \left\|\hat{b} - \mathbf{B}_0^\top\hat{\gamma}_1\right\|_1 \\ &\leq \left\|a - \mathbf{W}_t^\top\tau_1\right\|_\infty \left\|\hat{\beta}_1 - \beta_1\right\|_1 \\ &\quad + \left\{\left\|a - \mathbf{W}_t^\top\tau_1\right\|_\infty + \|\mathbf{B}_0\|_{1,\infty}\left\|\bar{X} - \mathbf{X}_c^\top\tau_2\right\|_\infty\right\}\left\|\hat{\gamma}_1 - \gamma_1\right\|_1 \\ &\quad + \left\|\bar{X} - \mathbf{X}_c^\top\tau_2\right\|_\infty \left\|\hat{b} - \mathbf{B}_0^\top\hat{\gamma}_1\right\|_1.\end{aligned}$$

It is apparent that the bias structures of $\hat{\theta}^{mr}$ and $\hat{\theta}_{10}^{1\to 2}$—and consequently that of $\hat{\theta}_{10}^{\text{cf}}$—are similar. Both biases are the sum of three terms, each of which has the form of a product of errors.

Relation to Balancing: We also draw a connection between our debiasing technique and balancing techniques in the causal inference literature. In the influence-function-based estimator for the ATE, the correction term contains the propensity score as a nuisance function. Imai and Ratkovic (2014) showed that the propensity score satisfies a certain balancing property and proposed a balancing estimator for the propensity score based on this observation. Zubizarreta (2015) proposed an optimization-based approach for achieving the same balance.

In the high-dimensional setting, Athey et al. (2018) illustrated that standard debiasing techniques for penalized regression-based estimators (e.g., the methods proposed in (Javanmard and Montanari, 2014; Van de Geer et al., 2014; Zhang and Zhang, 2014)) are inherently related to the balancing approach of Zubizarreta (2015). However, these approaches are not directly applicable in the presence of mediators for estimating the mediation functional. Recently, Liu and Ghassami (2024) presented counterparts of the balancing property and balancing estimators of Imai and Ratkovic (2014) for the mediation functional in a fixed-dimensional setting.

The counterparts of the balancing errors appearing in Liu and Ghassami (2024) are $\|\bar{X} - \mathbf{X}_c^\top\tau_2\|_\infty$ and $\|a - \mathbf{W}_t^\top\tau_1\|_\infty$ in our setting. In addition, the first balancing weight τ_2 is used to construct the direct approximation $\hat{m}_c = \mathbf{M}_c^\top\tau_2$ of $\mathbf{B}_0\bar{X}$, and the same weight is reused in the second bias-correction term. Hence, the two balancing optimizations in our proposed estimation strategy can be viewed as UHD adaptations of the balancing construction in Liu and Ghassami (2024), together with an additional coupling that is needed to handle the cross-arm structure of the mediation functional.

5 Theoretical Analysis

In this section, we provide a theoretical justification for our proposed methodology. We first state the assumptions used throughout our analysis and then establish that the proposed estimator of θ_{10} is $\sqrt{n}$-consistent and asymptotically normal. We subsequently extend this result to conditionally heteroskedastic outcome errors, the corresponding population-average mediation functional, and the estimators of the NDE and NIE introduced in Algorithms 2 and 3. Throughout this section, we write $d = p + q$ and allow the dimensions, sparsity levels, model parameters, and underlying distributions to vary with n, although this dependence is suppressed for notational simplicity. All constants appearing below are independent of (n, p, q) unless stated otherwise.

5.1 Assumptions

We first recall the definitions of the sub-Gaussian and sub-exponential norms used in our analysis.

Definition 5.1 (Sub-Gaussian and sub-exponential norms)**.** *For a random variable Z, its sub-Gaussian and sub-exponential norms are defined, respectively, as*

$$\|Z\|_{\psi_2} = \sup_{r\geq 1} r^{-1/2} \left(\mathbb{E}|Z|^r\right)^{1/r}, \qquad \|Z\|_{\psi_1} = \sup_{r\geq 1} r^{-1} \left(\mathbb{E}|Z|^r\right)^{1/r}.$$

For a random vector $Z \in \mathbb{R}^m$, we define $\|Z\|_{\psi_2} = \sup\{\|v^\top Z\|_{\psi_2} : \|v\|_2 = 1\}$.

We will also use the standard restricted eigenvalue condition for the high-dimensional regressions involved in our estimation procedure.

Definition 5.2 (Restricted eigenvalue and compatibility conditions)**.** *For an index set $S \subseteq \{1,\dots,m\}$ and a constant $L \geq 1$, define $\mathcal{C}(S,L) = \{\delta \in \mathbb{R}^m : \|\delta_{S^c}\|_1 \leq L\|\delta_S\|_1\}$. A positive semi-definite matrix $\Sigma \in \mathbb{R}^{m\times m}$ is said to satisfy the restricted eigenvalue condition at sparsity level r with constants $(L, \kappa_{\mathrm{RE}})$ if $\delta^\top\Sigma\delta \geq \kappa_{\mathrm{RE}}\|\delta\|_2^2$ for every $S \subseteq \{1,\dots,m\}$ with $|S| \leq r$ and every $\delta \in \mathcal{C}(S,L)$. The corresponding compatibility constant is defined by*

$$\phi^2(\Sigma; r, L) = \inf_{|S|\leq r,\ \delta\in\mathcal{C}(S,L)\ \atop \delta\neq 0} \frac{r\delta^\top\Sigma\delta}{\|\delta_S\|_1^2}.$$

We now state the two main assumptions used in our theoretical analysis.

Assumption 5.3 (Sampling, covariates, and errors)**.** *The following conditions hold.*

1. *The observations $\{(Y_i, M_i, X_i, A_i)\}_{i=1}^n$ are independent and identically distributed, and the random sample split is independent of the observations. The treatment probability satisfies $0 < \pi \leq \mathbb{P}(A = 1) \leq 1 - \pi < 1$ for some constant $\pi > 0$. Consequently, each arm-specific sample size within each data fold is proportional to n with probability tending to one.*
2. *For $a \in \{0,1\}$, let $\mu_{X,a} = \mathbb{E}[X \mid A = a]$, $\Sigma_{X,a} = \mathsf{var}(X \mid A = a)$ and $\mu_X = \mathbb{E}[X]$. The conditional distribution of $X - \mu_{X,a}$ given $A = a$ is uniformly sub-Gaussian, and $\|\mu_{X,a}\|_\infty \leq C_\mu$, $\|\mu_X\|_\infty \geq \kappa_X > 0$. Furthermore, there exist constants $0 < c_X < C_X < \infty$ such that*

$$c_X \leq \lambda_{\min}(\Sigma_{X,a}) \leq \lambda_{\max}(\Sigma_{X,a}) \leq C_X, \qquad a \in \{0,1\}.$$

   We also assume that $\mu_X^\top\Sigma_{X,a}^{-1}\mu_X \leq C_{\mu,\Sigma}$ for $a \in \{0,1\}$.
3. *Let $W = (X^\top \ \ M^\top)^\top$ and $\Omega_{W,a} = \mathbb{E}[WW^\top \mid A = a]$. Let $a_{10}^* = (\mu_X^\top \ \ (\mathbf{B}_0\mu_X)^\top)^\top$ and the target-specific population comparison vector $\eta_{10}^* = \Omega_{W,1}^{-1}a_{10}^*$. Assume $\|\eta_{10}^*\|_1 \leq C_\eta$ for some constant $C_\eta < \infty$.*
4. *Let ϵ_a denote the outcome-regression error on arm a. We assume $\mathbb{E}[\epsilon_a \mid X, M, A = a] = 0$ and $\mathsf{var}(\epsilon_a \mid X, M, A = a) = \nu_a(X, M)$ where, for some constants $0 < c_\epsilon < C_\epsilon < \infty$, $c_\epsilon \leq \nu_a(X, M) \leq C_\epsilon$ almost surely. Furthermore, the errors ϵ_a are assumed to be uniformly conditionally sub-Gaussian given $(X, M, A = a)$. The homoskedastic setup corresponds to the special case $\nu_a(X, M) = \sigma_a^2$. Let U_a denote the mediator-regression error on arm a. We assume $\mathbb{E}[U_a \mid X, A = a] = 0$ and $\mathsf{var}(U_a \mid X, A = a) = \Sigma_{U,a}$ where*

$$0 < c_U \leq \lambda_{\min}(\Sigma_{U,a}) \leq \lambda_{\max}(\Sigma_{U,a}) \leq C_U < \infty.$$

   Moreover, U_a is uniformly conditionally sub-Gaussian given $(X, A = a)$.

Assumption 5.4 (Parameters and sparsity)**.** *The following conditions hold.*

1. *There exist constants $C_\beta, C_\gamma, C_{\mathbf{B}} < \infty$ such that, for $a \in \{0,1\}$, $\|\beta_a\|_2 \leq C_\beta$, $\|\gamma_a\|_2 \leq C_\gamma$ and $\|\mathbf{B}_a\|_{1,\infty} \vee \|\mathbf{B}_a\|_{\mathrm{op}} \leq C_{\mathbf{B}}$.*
2. *There exist sparsity levels k_β, k_γ such that $\|\beta_a\|_0 \leq k_\beta$, $\|\gamma_a\|_0 \leq k_\gamma$ for $a \in \{0,1\}$. Let $k = k_\beta + k_\gamma$ denote the total sparsity. The mediator-regression matrices are row-wise sparse, i.e., $\max_{1\leq j\leq q}\|\mathbf{B}_{a,j\cdot}\|_0 \leq s$ for $a \in \{0,1\}$. This implies that the reduced-form coefficients $\rho_a = \beta_a + \mathbf{B}_a^\top\gamma_a$ satisfies $\|\rho_a\|_0 \leq k_\beta + sk_\gamma \leq (1+s)k$ and $\|\mathbf{B}_a^\top\gamma_{a'}\|_0 \leq sk_\gamma \leq sk$.*

3. *The sparsity levels and ambient dimensions satisfy* $(1+s)(1+k)\log(p+q) \ll \sqrt{n}$. *The Lasso tuning parameters are chosen at the usual high-dimensional orders* $\lambda_{n,1} \asymp \sqrt{\log(p+q)/n_t}$ *and* $\lambda_{n,\ell} \asymp \sqrt{\log p/n_a}$ *for* $\ell \in \{2,3,4\}$, *where* n_t *and* n_a *denote the relevant arm- and fold-specific sample sizes.*

Assumptions of this type are commonly used in the analysis of high-dimensional statistical models (e.g., see Bühlmann and Van De Geer (2011) and Lin et al. (2025)). In particular, the sub-Gaussian conditions impose restrictions on the tail behavior of the design vectors and regression errors, while the eigenvalue conditions prevent the arm-specific covariance matrices from becoming degenerate or excessively ill-conditioned. These assumptions facilitate the use of standard concentration inequalities for empirical means, covariance matrices, and high-dimensional regression scores. The condition $\|\eta_{10}^*\|_1 \le C_\eta$ is analogous to the bounded-ℓ_1 conditions on the coefficient vectors of Riesz representers commonly imposed in high-dimensional functional inference. Indeed, since $\Omega_{W,1}\eta_{10}^* = a_{10}^*$, the linear function $W^\top\eta_{10}^*$ satisfies $\mathbb{E}[W(W^\top\eta_{10}^*) \mid A=1] = a_{10}^*$, and therefore serves as the population balancing function associated with the target direction a_{10}^*. Conditions of this form are commonly used in the high-dimensional literature for debiasing and inference (e.g., see Condition SC(a) of Chernozhukov et al. (2022b), Assumption 6.3 of Escanciano and Pérez-Izquierdo (2023), and the conditions in Corollary D.2 of Chernozhukov et al. (2026)). For expositional clarity, the sub-Gaussian norms, covariance eigenvalues, parameter norms, and rowwise matrix norms in Assumptions 5.3 and 5.4 are uniformly bounded by constants independent of (n,p,q). It is possible to relax these conditions by allowing the corresponding bounds to grow with n. Such an extension would, however, require careful tracking of these sequences throughout the proofs, and would consequently modify the required convergence rates. Since this generalization would not contribute substantially to the main methodological ideas of the paper, we do not pursue it here.

We next discuss the sparsity-rate condition in Assumption 5.4. Conditions of the general form $r\log(d)/\sqrt{n} \to 0$, where r denotes the sparsity level and d is the ambient dimension, commonly arise in the debiased Lasso literature when establishing $\sqrt{n}$-consistency and asymptotic normality for linear contrasts of sparse high-dimensional parameters; see, for example, (Javanmard and Montanari, 2014; Van de Geer et al., 2014; Zhang and Zhang, 2014). Moreover, without imposing additional structural assumptions, the necessity of such a condition for constructing a $\sqrt{n}$-consistent estimator of a linear contrast was demonstrated by Cai and Guo (2017), who established a corresponding minimax lower bound for the estimation error. In our setting, estimation of θ_{aa} for $a \in \{0,1\}$ requires estimating a linear contrast (in the direction of $\bar{X}$ or μ_X), of the reduced-form parameter $\beta_a + \mathbf{B}_a^\top\gamma_a$. Under Assumption 5.4, $\|\beta_a + \mathbf{B}_a^\top\gamma_a\|_0 \le k_\beta + sk_\gamma \le (1+s)k$. Similarly θ_{10} involves a linear contrast (again in the direction of $\bar{X}$ or μ_X) of $\beta_1 + \mathbf{B}_0^\top\gamma_1$, and by the same argument $\|\beta_1 + \mathbf{B}_0^\top\gamma_1\|_0 \le (1+s)k$ under Assumption 5.4. Consequently, the standard sparsity-rate requirement discussed above translates in our setting to $(1+s)k\log p/\sqrt{n} \to 0$. This condition is therefore naturally required to obtain a $\sqrt{n}$-consistent and asymptotically normal estimator of θ_{aa} for each $a \in \{0,1\}$. As a consequence, this condition involving product sparsity sk cannot be relaxed without further structural assumptions. The consolidated condition $(1+s)(1+k)\log(p+q)/\sqrt{n} \to 0$ in Assumption 5.4 is a convenient sufficient condition that simultaneously controls this remainder, the treated outcome-regression remainder, and the analogous reduced-form remainders required for the NDE and NIE analyses. It is stated in a slightly stronger but simpler worst-case form to avoid introducing separate rate conditions for every high-dimensional regression.

Remark 5.5. *In the proof of Theorem 5.9, we require* $\|\mu_X\|_\infty \ge \kappa > 0$. *This is required to rule out the fact that if the contrasts are relatively small, then we would end up with* $\hat{\tau}_1, \hat{\tau}_2 \approx 0$, *resulting in a super-efficient estimator as noted in Athey et al. (2018).*

Remark 5.6. *Under Assumptions 5.3 and 5.4, the arm-specific design vector* W *is also uniformly sub-Gaussian. Indeed, on the arm* $A=a$, *we can write*

$$W = \begin{pmatrix} X \\ M \end{pmatrix} = \begin{pmatrix} X \\ \mathbf{B}_a X + U_a \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{B}_a & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} X \\ U_a \end{pmatrix}.$$

The uniform sub-Gaussianity of $X - \mu_{X,a}$ *and* U_a, *together with the uniform bound on* $\|\mathbf{B}_a\|_{\mathrm{op}}$, *implies that* $W - \mu_{W,a}$ *is uniformly sub-Gaussian conditional on* $A=a$. *Moreover, since* $\mathsf{cov}(X, U_a \mid A=a) = 0$, *the covariance matrix of* W *can be written as*

$$\Sigma_{W,a} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0} \\ \mathbf{B}_a & \mathbf{I}_q \end{pmatrix} \begin{pmatrix} \Sigma_{X,a} & \mathbf{0} \\ \mathbf{0} & \Sigma_{U,a} \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \mathbf{B}_a^\top \\ \mathbf{0} & \mathbf{I}_q \end{pmatrix}.$$

Therefore, the eigenvalues of $\Sigma_{W,a}$ are uniformly bounded above and away from zero.

Remark 5.7 (Sample restricted eigenvalue conditions)**.** *The restricted eigenvalue and compatibility conditions required for the high-dimensional regressions below need not be imposed as separate assumptions. Under Assumptions 5.3 and 5.4, the centered arm-specific design vectors X and W are uniformly sub-Gaussian, and their covariance eigenvalues are uniformly bounded away from zero and infinity. Therefore, the sample covariance/second moment matrices satisfy restricted eigenvalue conditions with probability tending to one, as soon as $n \gg s \log(ep/s)$ where p is the ambient dimension and s is the sparsity level (e.g., see (Raskutti et al., 2010; Rudelson and Zhou, 2012; Vershynin, 2018)). A formal statement and proof are provided in Section C of the supplementary material.*

5.2 Consistency and Asymptotic Normality of $\hat{\theta}_0^{\text{cf}}$

In this subsection, we present our key theoretical finding, i.e., our estimators $\hat{\theta}_{10}^{\text{cf}}, \hat{\theta}_{\text{NDE}}^{\text{cf}}, \hat{\theta}_{\text{NIE}}^{\text{cf}}$ are $\sqrt{n}$-consistent and asymptotically normal for both the sample and the population version of the parameters of interest. It is immediate from the definition of $\hat{\theta}_{10}^{\text{cf}}$ in Algorithm 1 that its construction relies on $\hat{\tau}_1$ and $\hat{\tau}_2$, obtained by solving optimization problems Equation (3.2) and (3.1) in Algorithm 1 respectively. Thus, it is imperative to establish that these optimization problems are feasible (with probability going to 1), a result we formalize in the following lemma:

Lemma 5.8. *Under Assumptions 5.3 and 5.4, there exist sufficiently large fixed constants (K_1, K_2), such that the optimization problems Equation (3.1) and (3.2) are feasible with probability going to 1. Moreover, the minimizers $\{\hat{\tau}_1^{(j)}, \hat{\tau}_2^{(j)}\}$ satisfy for $j \in \{1, 2\}$,*

$$n\left(\|\hat{\tau}_1^{(j)}\|_2^2 + \|\hat{\tau}_2^{(j)}\|_2^2\right) = O_p(1)\,.$$

The proof of this lemma is deferred to Section B.1 of the supplementary material. Lemma 5.8 shows that the optimization problems defining the debiasing weights are feasible with probability tending to one, and hence the proposed estimator is well-defined with high probability. Note that for obtaining $\hat{\theta}_{\text{NIE}}^{\text{cf}}$ via Algorithm 3, we need to perform another optimization (Equation (3.3)). However, this optimization problem is very closely related to that in Equation (3.1), and the proof for the existence of $\hat{\omega}$ is almost verbatim to that for the existence of $\hat{\tau}_2$, changing the notion of the treatment arms and data folds appropriately (recall that $\hat{\tau}_2$ involves control covariates and $\hat{\omega}$ involves treatment covariates). Therefore, we skip the proof for brevity.

We next establish the asymptotic normality of the proposed estimators. We begin with $\hat{\theta}_{10}^{\text{cf}}$ and consider inference for the sample-average mediation functional θ_{10} under homoskedastic outcome-regression errors, as the proof of this result contains the main technical ingredients of our analysis. We then extend the result to conditionally heteroskedastic outcome errors and to inference for the population-average mediation functional θ_{10}^{pop}. Finally, we present the corresponding results for $\hat{\theta}_{\text{NDE}}^{\text{cf}}$ and $\hat{\theta}_{\text{NIE}}^{\text{cf}}$. The following theorem establishes that, under homoskedasticity, $\hat{\theta}_{10}^{\text{cf}}$ is a $\sqrt{n}$-consistent and asymptotically normal estimator of θ_{10}:

Theorem 5.9. *Let $\hat{\theta}_{10}^{\text{cf}}$ be the estimator of θ_{10} obtained from Algorithm 1. Under Assumptions 5.3 and 5.4 along with homoskedasticity, i.e. $\nu_a(X, M) \equiv \sigma_a^2$ for $a \in \{0, 1\}$, we have:*

$$\frac{\sqrt{n}(\hat{\theta}_{10}^{\text{cf}} - \theta_{10})}{\sigma_n^{\text{cf}}} \xrightarrow{d} \mathcal{N}(0, 1)\,, \quad \textit{with}$$

$$(\sigma_n^{\text{cf}})^2 = \tfrac{n}{4} \textstyle\sum_{j=1}^{2} (\sigma_1^2 \|\hat{\tau}_1^{(j)}\|_2^2 + \gamma_1^\top \Sigma_{U,0} \gamma_1 \|\hat{\tau}_2^{(j)}\|_2^2)\,.$$

Remark 5.10. *When $\gamma_1 = 0$, i.e., the response is unaffected by the mediators. In that case, Theorem 5.9 implies $(\hat{\theta}_{10}^{\text{cf}} - \theta_{10})/((\sigma_1/2)\sqrt{(\|\hat{\tau}_1^{(1)}\|_2^2 + \|\hat{\tau}_1^{(2)}\|_2^2)})$ is asymptotically normal, which is Theorem 3 of Athey et al. (2018). Therefore, our result can be viewed as a generalized version of the results in (Athey et al., 2018) under cross-fitting in the presence of high-dimensional mediators along with the high-dimensional covariates.*

Theorem 5.9 establishes that $\hat{\theta}_{10}^{\text{cf}}$ is a $\sqrt{n}$-CAN estimator of the sample-average mediation functional θ_{10} under fairly general conditions. We next extend this result to the population-average mediation functional θ_{10}^{pop}. The sample-average target is often preferred in high-dimensional debiasing and approximate residual-balancing analyses because it effectively yields inference conditional on the observed covariates (e.g., see, Athey et al. (2018)). Nevertheless, the population-average functional is also a natural target in a superpopulation framework. The following corollary shows that $\hat{\theta}_{10}^{\text{cf}}$ remains a $\sqrt{n}$-CAN estimator of θ_{10}^{pop}, with an additional variance component accounting for the fluctuation of the empirical covariate average around its population mean.

Corollary 5.11. *Consider the same setup as in Theorem 5.9. Define* $\Lambda = \beta_1 + \mathbf{B}_0^\top \gamma_1$ *and set* $\sigma_X^2 = \Lambda^\top \Sigma_X \Lambda$ *where* $\Sigma_X = \mathsf{var}(X)$. *If* $\sigma_X^2 \leq C$ *for some constant* C *and* $(\sigma_n^{\text{cf}})^2/((\sigma_n^{\text{cf}})^2 + \sigma_X^2)$ *admits an asymptotically deterministic equivalent sequence, then*

$$\frac{\sqrt{n}(\hat{\theta}_{10}^{\text{cf}} - \theta_{10}^{\text{pop}})}{\sqrt{(\sigma_n^{\text{cf}})^2 + \sigma_X^2}} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

Compared with Theorem 5.9, the variance now contains the additional term $\sigma_X^2 = \Lambda^\top \Sigma_X \Lambda$, accounting for the fluctuation of $\bar{X}^\top \Lambda$ around its population counterpart $\mu_X^\top \Lambda$. The main technical difficulty remains the sample-average result established in Theorem 5.9. For the population-average target, we additionally combine this result with the central limit theorem for $\sqrt{n}(\bar{X} - \mu_X)^\top \Lambda$. Since both components depend on the realized covariates, their asymptotic joint behavior cannot be obtained by treating them as independent and directly adding their variances. We therefore use a conditional central limit theorem argument and impose the additional condition that the relative contribution of the sample-target variance to the total variance admits an asymptotically deterministic equivalent.

Our next corollary now further extends these results to the heteroskedastic outcome regression error setting of Assumption 5.3:

Corollary 5.12. *Consider the same setup as in Theorem 5.9, but instead of assuming outcome regression errors are homoskedastic, we assume there exist functions* $\nu_a(x,m) \in [c_-, c_+]$ *with* $c_- > 0$ *and* $c_+ < \infty$ *such that* $\mathsf{var}(\epsilon \mid X = x, M = m, A = a) = \nu_a(x,m)$. *Then* $\hat{\theta}_{10}^{\text{cf}}$ *satisfies:*

$$\frac{\sqrt{n}(\hat{\theta}_{10}^{\text{cf}} - \theta_{10})}{\sigma_n^{\text{cf,het}}} \xrightarrow{d} \mathcal{N}(0,1),$$

$$(\sigma_n^{\text{cf,het}})^2 = \tfrac{n}{4} \textstyle\sum_{j=1}^{2} \left[\sum_{i \in \mathcal{D}_j, A_i = 1} \left\{ (\hat{\tau}_{1,i}^{(j)})^2 \nu_1(X_i, M_i) \right\} + \gamma_1^\top \Sigma_{U,0} \gamma_1 \|\hat{\tau}_2^{(j)}\|_2^2 \right].$$

Furthermore, under the additional assumption as in Corollary 5.11 ($\sigma_X^2 \leq C$ *and* $(\sigma_n^{\text{cf,het}})^2/((\sigma_n^{\text{cf,het}})^2 + \sigma_X^2)$ *admits an asymptotically deterministic equivalent sequence) we have*

$$\frac{\sqrt{n}(\hat{\theta}_{10}^{\text{cf}} - \theta_{10}^{\text{pop}})}{\sqrt{(\sigma_n^{\text{cf,het}})^2 + \sigma_X^2}} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

The key difference between Corollary 5.12 and Theorem 5.9, together with its population extension in Corollary 5.11, lies in the variance normalization. Under conditional heteroskedasticity, the homoskedastic scale σ_n^{cf} is replaced by $\sigma_n^{\text{cf,het}}$, which explicitly incorporates the conditional outcome-error variance $\nu_1(X_i, M_i)$. In particular, setting $\nu_1(x,m) \equiv \sigma_1^2$ recovers the homoskedastic variance and, consequently, the conclusions of the preceding results. In practice, the heteroskedastic variance can be estimated using the corresponding squared regression residuals, i.e. replacing $\nu_1(X_i, M_i)$ with $(Y_i - X_i^\top \hat{\beta}_1 - M_i^\top \hat{\gamma}_1)^2$, while the control-arm mediator-error contribution can be estimated by replacing $\gamma_1^\top \Sigma_{U,0} \gamma_1$ by $(M_i^\top \hat{\gamma}_1 - X_i^\top \hat{b})^2$. Under our assumptions, this estimator is consistent and can therefore be used to construct asymptotically valid confidence intervals; see Section 6 and Section F of the supplementary material for numerical evidence.

Our next theorem presents an analogous result for $\hat{\theta}_{\text{NDE}}^{\text{cf}}$ and $\hat{\theta}_{\text{NIE}}^{\text{cf}}$. In analogy with Corollary 5.11, define $\sigma_{X,\text{NDE}}^2 = \Lambda_{\text{NDE}}^\top \Sigma_X \Lambda_{\text{NDE}}$ with $\Lambda_{\text{NDE}} = (\beta_1 - \beta_0) + \mathbf{B}_0^\top (\gamma_1 - \gamma_0)$ and $\sigma_{X,\text{NIE}}^2 = \Lambda_{\text{NIE}}^\top \Sigma_X \Lambda_{\text{NIE}}$ with $\Lambda_{\text{NIE}} = (\mathbf{B}_1 - \mathbf{B}_0)^\top \gamma_1$:

Theorem 5.13 (NDE). *Under Assumptions 5.3 and 5.4, the estimator $\hat{\theta}^{\text{cf}}_{\text{NDE}}$, as obtained from Algorithm 2 satisfies:*

$$\frac{\sqrt{n}(\hat{\theta}^{\text{cf}}_{\text{NDE}} - \theta_{\text{NDE}})}{\sigma^{\text{cf,het}}_{n,\text{NDE}}} \xrightarrow{d} \mathcal{N}(0,1)\,,$$

where, with $\omega_0(X) = (\gamma_1 - \gamma_0)^\top \Sigma_{U,0}(\gamma_1 - \gamma_0) + \mathbb{E}[\nu_0(X,M) \mid X, A=0]$,

$$(\sigma^{\text{cf,het}}_{n,\text{NDE}})^2 = \tfrac{n}{4}\sum_{j=1}^{2}\left[\sum_{i\in\mathcal{D}_j, A_i=1}(\hat{\tau}^{(j)}_{1,i})^2\nu_1(X_i,M_i) + \sum_{i\in\mathcal{D}_j, A_i=0}(\hat{\tau}^{(j)}_{2,i})^2\omega_0(X_i)\right].$$

Furthermore, if $\omega^{\text{het}}_{n,\text{NDE}} = (\sigma^{\text{cf,het}}_{n,\text{NDE}})^2/((\sigma^{\text{cf,het}}_{n,\text{NDE}})^2 + \sigma^2_{X,\text{NDE}})$ *admits an asymptotically deterministic equivalent sequence and* $\sigma^2_{X,\text{NDE}} \le C_X$ *for some* $C_X > 0$, *then*

$$\frac{\sqrt{n}\left(\hat{\theta}^{\text{cf}}_{\text{NDE}} - \theta^{\text{pop}}_{\text{NDE}}\right)}{\sqrt{(\sigma^{\text{cf,het}}_{n,\text{NDE}})^2 + \sigma^2_{X,\text{NDE}}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Theorem 5.14 (NIE). *Under Assumptions 5.3 and 5.4 and* $\|\gamma_1\|_2 \ge c_\gamma > 0$ *the estimator $\hat{\theta}^{\text{cf}}_{\text{NIE}}$, as obtained via Algorithm 3 satisfies*

$$\frac{\sqrt{n}(\hat{\theta}^{\text{cf}}_{\text{NIE}} - \theta_{\text{NIE}})}{\sigma^{\text{cf,het}}_{n,\text{NIE}}} \xrightarrow{d} \mathcal{N}(0,1)\,,$$

where, with $v^2_{U,a,1} = \gamma_1^\top \Sigma_{U,a}\gamma_1$,

$$(\sigma^{\text{cf,het}}_{n,\text{NIE}})^2 = \tfrac{n}{4}\sum_{j=1}^{2}\left[\sum_{\substack{i\in\mathcal{D}_j\\ A_i=1}}(\hat{\omega}^{(j)}_i - \hat{\tau}^{(j)}_{1,i})^2\nu_1(X_i,M_i) + v^2_{U,1,1}\|\hat{\omega}^{(j)}\|_2^2 + v^2_{U,0,1}\|\hat{\tau}^{(j)}_2\|_2^2\right].$$

Furthermore is the sequence $\omega^{\text{het}}_{n,\text{NIE}} = (\sigma^{\text{cf,het}}_{n,\text{NIE}})^2/((\sigma^{\text{cf,het}}_{n,\text{NIE}})^2 + \sigma^2_{X,\text{NIE}})$ *admits an asymptotically deterministic equivalent sequence and* $\sigma^2_{X,\text{NIE}} \le C_X < \infty$, *then*

$$\frac{\sqrt{n}(\hat{\theta}^{\text{cf}}_{\text{NIE}} - \theta^{\text{pop}}_{\text{NIE}})}{\sqrt{(\sigma^{\text{cf,het}}_{n,\text{NIE}})^2 + \sigma^2_{X,\text{NIE}}}} \xrightarrow{d} \mathcal{N}(0,1)\,.$$

Taken together, our theorems and corollaries establish the $\sqrt{n}$-CAN property of our proposed estimators for the (population and sample average) mediation functional, NDE, and NIE under conditional heteroskedasticity in the presence of UHD covariates and mediators and their interaction with the treatment assignment. These theoretical guarantees provide the basis for constructing asymptotically valid confidence intervals for the relevant causal effects.

Consistent estimation of asymptotic variance: In practice, the asymptotic variance can be estimated directly using the corresponding squared regression residuals. Specifically, define

$$\hat{\epsilon}^{(j)}_{1,i} = Y_i - X_i^\top\hat{\beta}^{(j)}_1 - M_i^\top\hat{\gamma}^{(j)}_1, \qquad \hat{r}^{(j)}_{0,i} = M_i^\top\hat{\gamma}^{(j')}_1 - X_i^\top\hat{b}^{(j)},$$

for the treated and control observations in $\mathcal{D}_j$, respectively, and set

$$(\hat{\sigma}^{\text{cf,het}}_n)^2 = \tfrac{n}{4}\sum_{j=1}^{2}\left[\sum_{\substack{i\in\mathcal{D}_j\\ A_i=1}}\left(\hat{\tau}^{(j)}_{1,i}\right)^2\left(\hat{\epsilon}^{(j)}_{1,i}\right)^2 + \sum_{\substack{i\in\mathcal{D}_j\\ A_i=0}}\left(\hat{\tau}^{(j)}_{2,i}\right)^2\left(\hat{r}^{(j)}_{0,i}\right)^2\right].$$

Under the assumptions of Corollary 5.12, this estimator is consistent for $(\sigma^{\text{cf,het}}_n)^2$. Intuitively speaking, the Lasso estimation rates ensure that the fitted residuals are asymptotically close to the corresponding true regression errors, while $n\|\hat{\tau}^{(j)}_k\|_2^2 = O_p(1)$ and the maximum-weight constraint $\|\hat{\tau}^{(j)}_k\|_\infty \le n^{-\alpha}$ ensure that no individual observation dominates the weighted sums. Consequently, the weighted squared residuals consistently estimate the conditional-variance contributions appearing in $(\sigma^{\text{cf,het}}_n)^2$, so that $(\hat{\sigma}^{\text{cf,het}}_n)^2 - (\sigma^{\text{cf,het}}_n)^2 = o_p(1)$. Analogous residual-based estimators can be constructed for the asymptotic variances of the NDE and NIE estimators, which we skip for brevity.

6 Simulation and Real Data Analysis

6.1 Simulation Designs

In this section, we present simulation studies to assess the performance of our proposed methodology. Specifically, we compare our debiased estimator ($\hat{\theta}^{\mathrm{cf}}_{10}$ in Algorithm 1) to a naive (non-debiased) estimator $\hat{\theta}^{\mathrm{plug-in}}$ (defined at the beginning of Section 3) in terms of their root mean squared error (RMSE), standard deviation (SD), confidence interval length and coverage. To understand the behavior of the proposed estimator across different regimes, we consider four data-generating processes that share the same generative mechanism (described below) but differ in terms of the sparsity, the signal strength of the nonzero coefficients, and heterogeneity.

1. We generate covariate $X \sim \mathcal{N}(\mu_X, \Sigma_X)$ with a dense but dimension-rescaled mean, $\mu_X = \frac{c}{\sqrt{p}}\mathbf{1}$ with $c = 0.5\sqrt{50}$. The covariance matrix is generated as $\Sigma_X = R\Lambda R^\top$, where R is an orthonormal matrix and $1 \le \lambda_{\min}(\Sigma_X) \le \lambda_{\max}(\Sigma_X) \le 2$.
2. We generate treatment $A \sim \mathrm{Ber}\{\mathrm{logit}(X^\top\alpha)\}$, where $\alpha \in \mathbb{R}^p$ is a sparse vector with $\|\alpha\|_0 = 3$ with nonzero elements randomly sampled from $\mathcal{U}(0, 2)$.
3. Upon generating (A, X), we generate the mediators from the model:

$$M = (1 - A)(\mathbf{B}_0 X + U) + A(\mathbf{B}_1 X + U'),$$

   where $\mathbf{B}_1, \mathbf{B}_0 \in \mathbb{R}^{q\times p}$ are row-sparse matrices, where each row of both $\mathbf{B}_0$ and $\mathbf{B}_1$ contains s (positions are randomly chosen) nonzero elements sampled from $\mathcal{U}(0.5, 1.5)$. The errors $U, U' \sim \mathcal{N}(0, \Sigma_U), \Sigma_U = R_U\Lambda_U R_U^\top$, where $R_U \in \mathbb{R}^{q\times q}$ is orthonormal and $\Lambda_U = \mathrm{diag}(\lambda_1, \ldots, \lambda_q)$ with $1 \le \lambda_{\min}(\Sigma_U) \le \lambda_{\max}(\Sigma_U) \le 2$.
4. Finally, we generate the response variable from the model:

$$Y = (1 - A)(X^\top\beta_0 + M^\top\gamma_0 + \epsilon') + A(X^\top\beta_1 + M^\top\gamma_1 + \epsilon),$$

   where $\beta_0, \beta_1 \in \mathbb{R}^p$ are sparse vectors with k_β non-zero elements. The non-zero indices are chosen randomly, and the values corresponding to those indices are sampled from $\mathcal{U}(0.5, 1.5)$. Similarly, $\gamma_0, \gamma_1 \in \mathbb{R}^q$ are also sparse vectors with k_γ non-zero elements. As before, the non-zero indices are sampled randomly, and values corresponding to those indices are sampled from $\mathcal{U}(0.5, 1.5)$, and $\epsilon, \epsilon' \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$.

With the above-mentioned data-generating process, we study four regimes:

- **Base.** The sparsity levels are fixed at $s = k_\beta = k_\gamma = 3$ across all dimensions, every nonzero coefficient is a moderate signal sampled from $\mathcal{U}(0.5, 1.5)$, and the errors are homoskedastic, $\epsilon, \epsilon' \sim \mathcal{N}(0, \sigma^2)$.
- **Increasing sparsity.** The number of nonzero coefficients grows with the ambient dimension, with $s = k_\beta = k_\gamma \in \{5, 10, 15, 20\}$ corresponding to $p + q \in \{100, 800, 1500, 2500\}$; nonzero values are again moderate, sampled from $\mathcal{U}(0.5, 1.5)$. To keep the mediator model well-scaled as its support grows, the rows of the mediator coefficient matrices $\mathbf{B}_0, \mathbf{B}_1 \in \mathbb{R}^{q\times p}$ are normalized to unit ℓ_1 norm.
- **Vanishing minimum non-zero signal.** The sparsity is now fixed at $s = k_\beta = k_\gamma = 3$, but approximately half of the nonzero coefficients of the regression coefficients are now generated from near the detection boundary, with magnitudes of order $n_{\max}^{-1/2}$ ($n_{\max} = 1250$, the maximum number of samples used in our simulation, and these coordinates are sampled from $\mathcal{U}(0.8, 1.2)/\sqrt{n_{\max}}$), while the remaining coefficients are moderate as in the base case. This regime assesses the performance of the estimator when part of the signal is only weakly identifiable.
- **Heteroskedastic noise.** We retain the sparsity ($s = k_\beta = k_\gamma = 3$) and moderate signals as in the base case, but replace the homoskedastic errors in item (4) by conditionally heteroskedastic ones whose

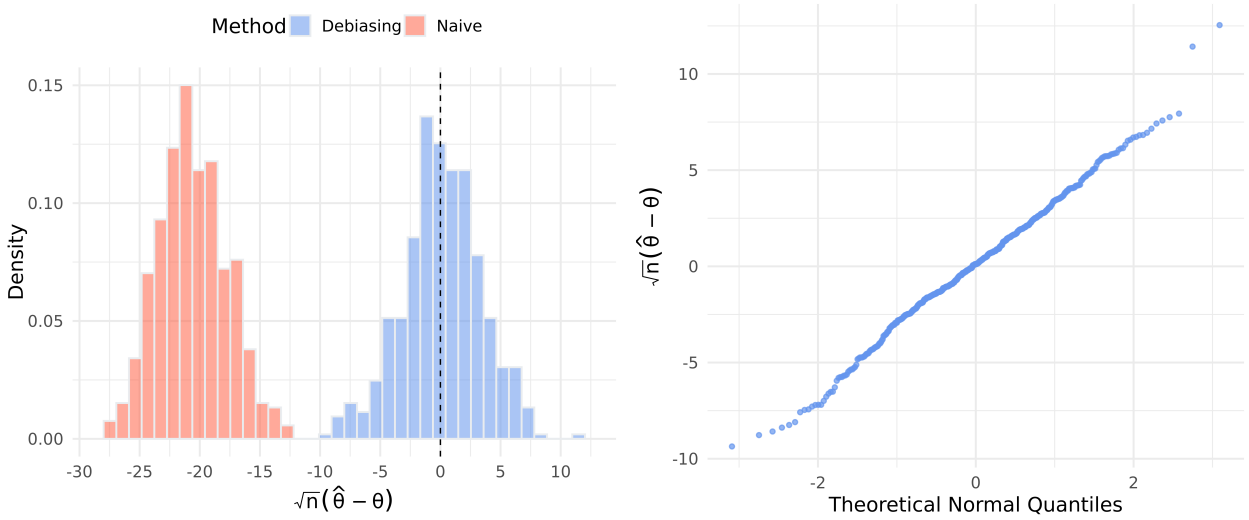


Figure 2: Histogram and QQ plot of $\sqrt{n}(\hat{\theta} - \theta)$ under $n = 750, p + q = 1500, \sigma_\epsilon = 0.1$.

variance is now a function of both the covariates and the mediators. Specifically, we generate errors as $\epsilon_i \mid (X_i, M_i) \sim \mathcal{N}\big(0,\, \sigma^2\, v(X_i, M_i)\big)$ (analogously for ϵ_i'), where

$$v(X_i, M_i) \;=\; \frac{1 + \frac{1}{2}\big[(X_i^\top a_\epsilon)^2 + (M_i^\top b_\epsilon)^2\big]}{\mathbb{E}\big[1 + \frac{1}{2}\big((X^\top a_\epsilon)^2 + (M^\top b_\epsilon)^2\big)\big]}\,, \tag{6.1}$$

with sparse $a_\epsilon \in \mathbb{R}^p$, $b_\epsilon \in \mathbb{R}^q$ satisfying $\|a_\epsilon\|_0 = \|b_\epsilon\|_0 = 3$ and nonzero entries of (a_ϵ, b_ϵ) drawn from $\mathcal{U}(0, 1)$. The normalization ensures $\mathbb{E}[v(X, M)] = 1$, so the marginal noise level matches the homoskedastic case.

In Assumption 5.3, we claim that $0 < c_\epsilon < C_\epsilon < \infty$ and $c_\epsilon \le \nu_a(X, M) \le C_\epsilon$ almost surely, whereas in Equation (6.1), $v(X_i, M_i)$ can potentially be unbounded. We include this setting as a stress test, allowing the conditional variance to be unbounded in order to assess the robustness of our estimator beyond the theoretical conditions. Our Assumption 5.3 could also be extended to accommodate such settings via an appropriate truncation that controls extreme conditional variances.

The number of samples n varies over $\{250, 500, 750, 1000, 1250\}$. For each value of n in this set, we vary $p + q$ over the set $\{100, 800, 1500, 2500\}$. Furthermore, for the homoskedastic case, the noise variance $\sigma^2 \in \{0.1, 1\}$. Recall that our parameter of interest is $\theta_{10} = \bar{X}^\top(\beta_1 + \mathbf{B}_0^\top \gamma_1)$ where $\bar{X}$ is the sample average of the covariates. We keep X and other parameters (e.g., $(\beta_a, \gamma_a, \alpha, \mathbf{B}_a)$ for $a \in \{0, 1\}$) fixed in the simulation study to keep the parameter of interest unchanged.

For brevity, the main text reports only the base regime with $\sigma^2 = 1$ (Table 1). The remaining configurations, namely the base regime with $\sigma^2 = 0.1$, the increasing-sparsity, vanishing minimal non-zero signal, and the heteroskedastic-noise regimes, together with the sensitivity analyses, can be found in Section F of the supplementary material (Tables 7– 11 and 12–14).

We use Algorithm 1 to compute our estimator. The Lasso estimates in Step 2 and Step 4 are obtained using the R package `glmnet`, where the tuning parameter is selected by cross-validation using the `lambda.1se` rule from R function `cv.glmnet` (Friedman et al. (2010)). The debiasing weights in Step 1 and Step 3 are solved by solving the corresponding optimization problem via the R package `pogs` with $K_1 = K_2 = 2.75$. This particular choice of (K_1, K_2) is motivated by our sensitivity analysis, as elaborated in Table 12 - 14 in Section F of the supplementary material. This choice remains consistent with previous studies (e.g., Javanmard and Montanari (2014)). We also record the computational cost under a representative simulation setting. Specifically, when $n = 1000$ and $p + q = 1500$, running the 500 Monte Carlo replications across 30 CPU cores takes approximately 12 minutes of wall-clock time, which amortizes to about 1.4 seconds per replication ($12 \times 60/500 \approx 1.4$).

$s=3,\ k_1=k_2=3$		$\sigma^2=1$										
Estimators		$\hat{\theta}^{\mathrm{cf}}_{10}$ (Debiasing)							$\hat{\theta}^{\mathrm{plug-in}}$			
		Monte Carlo CI				Analytical CI			Monte Carlo CI			
$p+q$	n	RMSE	SD	CI Len	Cov	SE	CI Len	Cov	RMSE	SD	CI Len	Cov
100	250	0.691	0.334	1.308	0.892	0.304	1.194	0.902	0.951	0.244	0.958	0.034
	500	0.197	0.179	0.700	0.912	0.187	0.734	0.926	0.678	0.152	0.596	0.014
	750	0.213	0.142	0.556	0.902	0.150	0.588	0.926	0.595	0.124	0.486	0.008
	1000	0.128	0.121	0.473	0.932	0.130	0.509	0.930	0.529	0.103	0.405	0.000
	1250	0.132	0.125	0.490	0.936	0.121	0.474	0.948	0.483	0.103	0.403	0.008
800	250	0.238	0.225	0.883	0.948	0.384	1.505	0.996	0.484	0.215	0.841	0.450
	500	0.169	0.146	0.574	0.906	0.204	0.801	0.982	0.432	0.056	0.221	0.000
	750	0.159	0.131	0.512	0.892	0.152	0.597	0.954	0.430	0.056	0.219	0.000
	1000	0.118	0.118	0.461	0.948	0.132	0.519	0.966	0.349	0.041	0.160	0.000
	1250	0.098	0.097	0.382	0.952	0.114	0.446	0.976	0.322	0.039	0.153	0.000
1500	250	0.445	0.254	0.995	0.682	0.375	1.471	0.908	0.245	0.173	0.679	0.817
	500	0.175	0.150	0.588	0.920	0.199	0.781	0.980	0.372	0.044	0.173	0.000
	750	0.168	0.124	0.488	0.854	0.148	0.580	0.950	0.321	0.036	0.141	0.000
	1000	0.097	0.097	0.381	0.946	0.122	0.477	0.984	0.280	0.029	0.115	0.000
	1250	0.104	0.096	0.376	0.926	0.108	0.424	0.952	0.248	0.026	0.101	0.000
2500	250	0.721	0.313	1.227	0.464	0.466	1.825	0.798	0.454	0.183	0.718	0.358
	500	0.168	0.157	0.617	0.944	0.226	0.885	0.894	0.270	0.030	0.117	0.000
	750	0.208	0.132	0.516	0.774	0.170	0.666	0.966	0.326	0.033	0.130	0.000
	1000	0.118	0.114	0.448	0.948	0.139	0.543	0.962	0.255	0.025	0.098	0.000
	1250	0.111	0.108	0.424	0.950	0.122	0.480	0.984	0.171	0.018	0.071	0.000

Table 1: Simulation results for the debiased ($\hat{\theta}^{\mathrm{cf}}_{10}$) and naive ($\hat{\theta}^{\mathrm{plug-in}}$) estimators under the base DGP with $\sigma^2=1$. Numbers are averaged over 500 replications with $K_1=K_2=2.75$.

The simulation results are summarized in Table 1 and in Section F of supplementary materials. Each table reports the root-mean-squared error (RMSE) and standard deviation (SD) of both $\hat{\theta}^{\mathrm{plug-in}}$ and $\hat{\theta}^{\mathrm{cf}}_{10}$, averaged over 500 Monte Carlo replications, along with the estimated confidence-interval length and empirical coverage under both a Monte Carlo standard deviation and our estimated analytical standard error. The results indicate that the debiased estimator $\hat{\theta}^{\mathrm{cf}}_{10}$ consistently achieves a lower RMSE than the naive estimator, with the performance gap becoming increasingly prominent as the sample size n grows. Furthermore, the standard deviations of the two estimators are comparable, suggesting that the improved RMSE of our debiased estimator is attributable to the successful bias correction. This pattern is consistent across all four regimes: in the base, minimal-signal, and heteroskedastic-noise regimes, $\hat{\theta}^{\mathrm{cf}}_{10}$ achieves close-to-nominal coverage as n increases, and the estimated analytical standard error $\hat{\sigma}^{\mathrm{cf}}_n$ closely matches the Monte Carlo standard deviation. In the increasing-sparsity regime, the same conclusions hold in moderate dimensions, while the most challenging high-dimensional, small-sample cells ($p+q=2500$, $n=250$) are, as expected, the hardest, with $\hat{\theta}^{\mathrm{cf}}_{10}$ nonetheless remaining more accurate and better calibrated than $\hat{\theta}^{\mathrm{plug-in}}$. To further illustrate the performance of the estimators, we display histograms of the naive (red) and debiased (blue) estimators in the left column of Figure 2, based on 500 Monte Carlo replications. The estimators are centered around the true parameter and scaled by $\sqrt{n}$, where n is the number of samples. The histogram clearly reveals that the naive estimator is substantially biased, while the debiased estimator is centered around zero as a result of successful bias correction. Additionally, the Q-Q plot of the debiased estimator, shown in the right column of

Figure 2, aligns closely with the standard normal distribution, providing empirical evidence for its asymptotic normality and supporting our theoretical results.

Remark 6.1 (Comparison with Guo et al. (2022))**.** *We also conducted additional numerical experiments to compare the performance of our proposed estimator with the estimator of Guo et al. (2022) in a simplified setting where both methods can be applied. A direct comparison is not possible, since the framework of Guo et al. (2022, 2023, 2024) does not accommodate treatment-covariate interactions in the mediator model or treatment–covariate/treatment–mediator interactions in the outcome model. Therefore, in the supplementary material, we consider a simplified common setting: we remove the treatment interactions from our model and include UHD covariates with a natural modification of their procedure. The details can be found in Section D.1 of the supplementary document. It is evident from the results that, especially when the sparsity level is high or the minimum active signal strength is small, our debiased estimator outperforms the modified estimator in Guo et al. (2022) in both RMSE and asymptotic coverage.*

6.2 Real Data Experiments

DNA methylation, an epigenetic modification characterized by the addition of a methyl group to the DNA molecule, is essential for regulating gene expression and is linked to various biological functions, and is often regarded as a potential mediator linking exposures to health outcomes (Fujii et al., 2022). Recent studies suggest that smoking leads to widespread DNA methylation changes, which may contribute to cancer development and progression (Tan et al., 2013; Zhang et al., 2016). Meanwhile, certain smoking-associated methylation markers have been linked to increased mortality risk among lung cancer patients (Zhang et al., 2016; Mehta et al., 2015). As an application of our developed methodology, we apply our method to the data from the lung cancer cohort of The Cancer Genome Atlas (TCGA) project to study the direct and indirect effects of smoking on the survival time of patients diagnosed with lung squamous cell carcinoma and lung adenocarcinoma. The data is publicly available at (`https://xenabrowser.net/datapages/`). The dataset includes DNA methylation data from 907 individuals, feature data for 1,299 individuals, and survival data for 1,145 individuals, among which 833 individuals were common.

- Our response variable, Y, is the logarithm of the survival time, where the survival time is measured as the number of days from diagnosis to death or the last follow-up date. We used the inter-quartile range to detect and remove the outliers in the response variable.
- We selected 20 clinically relevant features, X (e.g., age, sex, etc.; see full list in Section E of the supplementary material) for our study by excluding all those features which has more than 100 missing values. See the Supplementary Materials for the complete list of the chosen features. We then imputed the remaining missing values; for numerical features (11), we used `softImpute` (Mazumder et al., 2010) and for categorical variables (9), we used `mice` (Van Buuren and Groothuis-Oudshoorn, 2011) imputation. Furthermore, we considered up to third-order polynomial interactions for better expressiveness. These operations ultimately yielded a 42-dimensional covariate.
- The DNA methylation data comprises $350,019$ DNA methylation marker values (M), represented as continuous ratios ranging from 0 to 1, reflecting the intensity of methylation, where higher values indicate a greater degree of methylation. For more details, see (`https://www.cancer.gov/ccg/research/genome-sequencing/tcga`). We selected only those mediators with no missing values. This reduces the number of mediators to $264,797$.

After the data pre-processing, we are left with $n = 738$ units, each with 42 features (X) and $264,797$ mediators (M). The treatment indicator (A) is the smoking status of the units ($A = 1$ indicates a smoker). Among these samples, there are 42.6%(315) females and 57.3%(423) males. The patients' ages range from 33 to 90 years, with a median age of 67 years. The median survival time is 603 days (1.65 years).

To estimate the direct and indirect effects (NDE and NIE) of smoking on the survival time of lung cancer patients, we fit the following high-dimensional linear model to the pre-processed data:

$$M_i = (1 - A_i)(\mathbf{B}_0 X_i + U_i) + A_i(\mathbf{B}_1 X_i + U_i'), \qquad i = 1, \ldots, 738,$$

	Estimate	Bootstrap Mean	Quantile CI (95%)
Indirect Effect	-0.0575	-0.0506	$(-0.1536,\ -0.0395)$
Direct Effect	-0.0244	-0.0288	$(-0.1507,\ 0.0773)$
Total Effect	-0.0819	-0.0778	$(-0.2044,\ -0.0369)$

Table 2: Summary of the estimated effects, bootstrap means, and 95% quantile confidence intervals based on 300 bootstrap samples.

	Estimate	Bootstrap Mean	Quantile CI (95%)
Indirect Effect	0.2600	0.2531	(0.1279, 0.3752)
Direct Effect	-0.0223	-0.0178	(-0.2496, 0.1688)
Total Effect	0.2378	0.2353	(-0.0079, 0.4346)

Table 3: Summary of the effects in OLS (CpG sites selected from prior studies) with mean and confidence interval (CI) under 300-sample bootstrap.

$$Y_i = (1 - A_i)(X_i^\top \beta_0 + M_i^\top \gamma_0 + \epsilon_i') + A_i(X_i^\top \beta_1 + M_i^\top \gamma_1 + \epsilon_i)\,. \tag{6.2}$$

We then apply Algorithm 2 and 3 to estimate the NDE, NIE, and total treatment effect. For our real data application, one full run of the proposed algorithm (one bootstrap replicate) takes approximately 48 seconds using 10 CPU cores.

Table 2 summarizes the estimates of the NDE, NIE, and total effect of smoking on log survival time. The estimate of the total effect implies that, on average, smokers have **92.14%** ($e^{-0.0819} \times 100\%$) of the survival time of non-smokers, i.e., smoking reduces approximately **7.86%** of the survival time of lung cancer patients. Of this total effect, our results demonstrate that approximately **5.59%** $= 100 \times (1 - e^{-0.0575})\%$ corresponds to the indirect effect (NIE), i.e., the impact of smoking on survival time mediated via DNA methylation. In comparison, approximately **2.41%** $= 100 \times (1 - e^{-0.0244})\%$ corresponds to the direct effect (NDE), which can be attributed to pathways beyond the methylation mechanism considered in this analysis.

Since the confidence interval for the indirect effect estimate excludes zero, while the confidence interval for the direct effect estimate contains zero, our findings suggest that the estimated effect of smoking on survival time is primarily mediated through the DNA methylation mechanism captured in this analysis. This indicates that DNA methylation plays an important mediating role in the current dataset, although additional methylation mechanisms and other biological pathways may remain to be investigated.

We compared the results with a fixed-dimensional OLS-based approach. Note that if both X and M are low-dimensional in Equation (6.2), we can easily estimate the NDE, NIE (and consequently the total effect) by computing ordinary least squared (OLS) estimators via regressing i) Y on (X, M) (on treatment and control observations separately), and ii) M on X (on treatment and control observations separately). To make the problem low-dimensional, we selected CpG sites (cg19757631, cg05147638, cg24720672, cg08108679, cg05575921, cg24859433) which were identified as significant mediators in previous studies (Cui et al., 2021; Zhang et al., 2021a,b), as our M, and 20 clinically relevant features as our X (i.e., we do not consider their higher-order terms and interactions).

In Table 3, we present the results in a fixed-dimensional setting. The results reveal that the OLS-based approach produces conclusions about the indirect and total effects that are contrary to intuition. The estimate of the indirect and total effect imply that smokers have 128.80% ($e^{0.2531} \times 100\%$) (mediated through DNA methylation), and 126.52% ($e^{0.2353} \times 100\%$) of survival time of non-smokers, i.e., smoking increases $\approx 29\% = 100 \times (1 - e^{0.2531})\%$ (corresponds to the NIE) and $\approx 27\% = 100 \times (1 - e^{0.2353})\%$ of the survival time of lung cancer patients. Furthermore, the confidence interval for the direct effect and total effect includes zero, suggesting a lack of statistical significance. These findings further emphasize the limitations of low-dimensional mediator selection and conventional modeling approaches, reinforcing the importance of incorporating high-dimensional methodologies for accurate mediation analysis in complex biological systems.

7 Conclusion

In this paper, we introduced a novel methodology for estimating the mediation functional and, consequently, of the NDE, NIE in the presence of ultra-high-dimensional covariates and mediators. Our model is sufficiently flexible to accommodate interactions between the treatment assignment and both covariates and mediators in the conditional mean of the response, as well as interactions between the treatment assignment and covariates in the conditional expectation of the mediators. The primary contribution is to develop a novel debiasing framework to obtain a $\sqrt{n}$-CAN estimate of the mediation functional and, consequently, of the NDE, NIE in the presence of these challenges, which is potentially applicable to other problems sharing a similar structural equation model. To the best of our knowledge, this is the first study to address the challenge of estimating NDE and NIE when dealing with high-dimensional covariates, mediators, and their interactions with the treatment variable simultaneously. While this paper focuses on a linear model, extending the methodology to generalized linear models is a natural next step, which we leave as an intriguing direction for future research.

References

Albert, J. M. (2008). Mediation analysis via potential outcomes models. *Statistics in medicine*, 27(8):1282–1304.

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

Baron, R. M. and Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173.

Ben-Michael, E., Feller, A., Hirshberg, D. A., and Zubizarreta, J. R. (2021). The balancing act in causal inference. *arXiv preprint arXiv:2110.14831*.

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Bulinski, A. V. (2017). Conditional central limit theorem. *Theory of Probability & Its Applications*, 61(4):613–631.

Cai, T., Tony Cai, T., and Guo, Z. (2021). Optimal statistical inference for individualized treatment effects in high-dimensional models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(4):669–719.

Cai, T. T. and Guo, Z. (2017). Confidence intervals for high-dimensional linear regression: Minimax rates and adaptivity. *Annals of Statistics*.

Chakrabortty, A., Nandy, P., and Li, H. (2018). Inference for individual mediation effects and interventional effects in sparse high-dimensional causal graphical models. *arXiv preprint arXiv:1809.10652*.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., and Lindquist, M. A. (2018). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics*, 19(2):121–136.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022a). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022b). Debiased machine learning of global and local parameters using regularized riesz representers. *The Econometrics Journal*, 25(3):576–601.

Chernozhukov, V., Newey, W. K., Singh, R., and Syrgkanis, V. (2026). Adversarial estimation of riesz representers. *Journal of the American Statistical Association*, pages 1–12.

Cui, Y., Luo, C., Luo, L., and Yu, Z. (2021). High-dimensional mediation analysis based on additive hazards model for survival data. *Frontiers in Genetics*, 12:771932.

Escanciano, J. C. and Pérez-Izquierdo, T. (2023). Automatic locally robust gmm with machine-learning-generated regressors. *arXiv e-prints*, pages arXiv–2301.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33:1–22.

Fujii, R., Sato, S., Tsuboi, Y., Cardenas, A., and Suzuki, K. (2022). Dna methylation as a mediator of associations between the environment and chronic diseases: A scoping review on application of mediation analysis. *Epigenetics*, 17(7):759–785.

Guo, X., Li, R., Liu, J., and Zeng, M. (2022). High-dimensional mediation analysis for selecting dna methylation loci mediating childhood trauma and cortisol stress reactivity. *Journal of the American Statistical Association*, 117(539):1110–1121.

Guo, X., Li, R., Liu, J., and Zeng, M. (2023). Statistical inference for linear mediation models with high-dimensional mediators and application to studying stock reaction to covid-19 pandemic. *Journal of Econometrics*, 235(1):166–179.

Guo, X., Li, R., Liu, J., and Zeng, M. (2024). Estimations and tests for generalized mediation models with high-dimensional potential mediators. *Journal of Business & Economic Statistics*, 42(1):243–256.

Guo, Z., Wang, W., Cai, T. T., and Li, H. (2019). Optimal estimation of genetic relatedness in high-dimensional linear models. *Journal of the American Statistical Association*, 114(525):358–369.

Hirshberg, D. A. and Wager, S. (2021). Augmented minimax linear estimation. *The Annals of Statistics*, 49(6):3206–3227.

Huang, Y.-T. and Pan, W.-C. (2016). Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 72(2):402–413.

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science*.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263.

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

Jones, J., Ertefaie, A., and Strawderman, R. L. (2025). Causal mediation analysis: selection with asymptotically valid inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(3):678–700.

Lee, K. J. and Schuler, A. (2025). Rieszboost: Gradient boosting for riesz regression. *arXiv preprint arXiv:2501.04871*.

Lin, Y., Guo, Z., Sun, B., and Lin, Z. (2025). Testing high-dimensional mediation effect with arbitrary exposure–mediator coefficients. *TEST*, 34(3):530–579.

Lindquist, M. A. (2012). Functional causal mediation analysis with an application to brain connectivity. *Journal of the American Statistical Association*, 107(500):1297–1309.

Liu, C. and Ghassami, A. (2024). Bias-targeted nonparametric balancing for stable causal mediation analysis. *arXiv preprint arXiv:2404.00735*.

Loh, W. W., Moerkerke, B., Loeys, T., and Vansteelandt, S. (2022). Nonlinear mediation analysis with high-dimensional mediators whose causal structure is unknown. *Biometrics*, 78(1):46–59.

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

Mehta, A., Dobersch, S., Romero-Olmedo, A. J., and Barreto, G. (2015). Epigenetics in lung cancer diagnosis and therapy. *Cancer and Metastasis Reviews*, 34(2):229–241.

Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*.

Pearl, J. (2022). Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl*, pages 373–392. Association for Computing Machinery.

Perera, C., Zhang, H., Zheng, Y., Hou, L., Qu, A., Zheng, C., Xie, K., and Liu, L. (2022). Hima2: high-dimensional mediation analysis and its application in epigenome-wide dna methylation data. *BMC bioinformatics*, 23(1):296.

Rakshit, P. and Guo, Z. (2024). Statistical inference in high-dimensional poisson regression with applications to mediation analysis. *arXiv preprint arXiv:2410.20671*.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259.

Richiardi, L., Bellocco, R., and Zugna, D. (2013). Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 42(5):1511–1519.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Rudelson, M. and Zhou, S. (2012). Reconstruction from anisotropic random measurements. In *Conference on Learning Theory*, pages 10–1. JMLR Workshop and Conference Proceedings.

Shi, C. and Li, L. (2022). Testing mediation effects using logic of boolean matrices. *Journal of the American Statistical Association*, 117(540):2014–2027.

Tan, Q., Wang, G., Huang, J., Ding, Z., Luo, Q., Mok, T., Tao, Q., and Lu, S. (2013). Epigenomic analysis of lung adenocarcinoma reveals novel dna methylation patterns associated with smoking. *OncoTargets and therapy*, pages 1471–1479.

Tchetgen Tchetgen, E. J. and Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Annals of statistics*, 40(3):1816.

Ten Have, T. R. and Joffe, M. M. (2012). A review of causal estimation of effects in mediation analyses. *Statistical methods in medical research*, 21(1):77–107.

Van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*.

VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

VanderWeele, T. J., Vansteelandt, S., and Robins, J. M. (2014). Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology (Cambridge, Mass.)*, 25(2):300.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

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

Zhang, H., Hou, L., and Liu, L. (2021a). A review of high-dimensional mediation analyses in dna methylation studies. *Epigenome-wide association studies: methods and protocols*, pages 123–135.

Zhang, H., Zheng, Y., Hou, L., Zheng, C., and Liu, L. (2021b). Mediation analysis for survival data with high-dimensional mediators. *Bioinformatics*, 37(21):3815–3821.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., et al. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154.

Zhang, J., Wei, Z., and Chen, J. (2018). A distance-based approach for testing the mediation effect of the human microbiome. *Bioinformatics*, 34(11):1875–1883.

Zhao, Y. and Luo, X. (2016). Pathway lasso: estimate and select sparse mediation pathways with high dimensional mediators. *arXiv preprint arXiv:1603.07749*.

Zubizarreta, J. R. (2015). Stable weights that balance covariates for estimation with incomplete outcome data. *Journal of the American Statistical Association*, 110(511):910–922.

A Proof of Proposition 4.1

Proof. We first consider $\hat{\theta}_{10}^{1\to 2}$. Recall that $\theta_{10} = \bar{X}^\top(\beta_1 + \mathbf{B}_0^\top\gamma_1)$, $\phi_1^\top = (\beta_1^\top \quad \gamma_1^\top)$. We also recall the definitions of $\hat{\vartheta}_1$ and $\hat{\vartheta}_2$ in Algorithm 1. Adding and subtracting $(a^{(1)})^\top\phi_1$ in $\hat{\vartheta}_1$, we have

$$\begin{aligned}\hat{\vartheta}_1 &= \hat{\vartheta}_1 - (a^{(1)})^\top\phi_1 + (a^{(1)})^\top\phi_1 \\ &= \left(a^{(1)} - \mathbf{W}_{t,2}^\top\hat{\tau}_1^{(2)}\right)^\top\left(\hat{\phi}^{(2)} - \phi_1\right) + (a^{(1)})^\top\phi_1 + (\hat{\tau}_1^{(2)})^\top\boldsymbol{\epsilon}_t^{(2)}.\end{aligned}$$

Similarly, adding and subtracting $\bar{X}^\top b^{(2)}$ in $\hat{\vartheta}_2$ gives

$$\begin{aligned}\hat{\vartheta}_2 &= \hat{\vartheta}_2 - \bar{X}^\top b^{(2)} + \bar{X}^\top b^{(2)} \\ &= \left(\bar{X} - \mathbf{X}_{c,1}^\top\hat{\tau}_2^{(1)}\right)^\top\left(\hat{b}^{(1)} - b^{(2)}\right) + \bar{X}^\top b^{(2)} + (\hat{\tau}_2^{(1)})^\top\mathbf{U}_c^{(1)}\hat{\gamma}_1^{(2)}.\end{aligned}$$

Therefore,

$$\begin{aligned}\hat{\theta}_{10}^{1\to 2} &= \hat{\vartheta}_1 + \hat{\vartheta}_2 - \hat{m}_{c,1}^\top\hat{\gamma}_1^{(2)} \\ &= \left(a^{(1)} - \mathbf{W}_{t,2}^\top\hat{\tau}_1^{(2)}\right)^\top\left(\hat{\phi}^{(2)} - \phi_1\right) + \left(\bar{X} - \mathbf{X}_{c,1}^\top\hat{\tau}_2^{(1)}\right)^\top\left(\hat{b}^{(1)} - b^{(2)}\right) \\ &\quad + (a^{(1)})^\top\phi_1 + \bar{X}^\top b^{(2)} + (\hat{\tau}_2^{(1)})^\top\mathbf{U}_c^{(1)}\left(\hat{\gamma}_1^{(2)} - \gamma_1\right) - \hat{m}_{c,1}^\top\hat{\gamma}_1^{(2)} \\ &\quad + (\hat{\tau}_1^{(2)})^\top\boldsymbol{\epsilon}_t^{(2)} + (\hat{\tau}_2^{(1)})^\top\mathbf{U}_c^{(1)}\gamma_1. \qquad \text{(A.1)}\end{aligned}$$

Next, note that $\hat{m}_{c,1} = \mathbf{M}_{c,1}^\top\hat{\tau}_2^{(1)} = \mathbf{B}_0\mathbf{X}_{c,1}^\top\hat{\tau}_2^{(1)} + (\mathbf{U}_c^{(1)})^\top\hat{\tau}_2^{(1)}$. Moreover, $(a^{(1)})^\top\phi_1 = \bar{X}^\top\beta_1 + \hat{m}_{c,1}^\top\gamma_1$, $\bar{X}^\top b^{(2)} = \bar{X}^\top\mathbf{B}_0^\top\hat{\gamma}_1^{(2)}$. The second line in (A.1) follows that

$$\begin{aligned}&(a^{(1)})^\top\phi_1 + \bar{X}^\top b^{(2)} + (\hat{\tau}_2^{(1)})^\top\mathbf{U}_c^{(1)}\left(\hat{\gamma}_1^{(2)} - \gamma_1\right) - \hat{m}_{c,1}^\top\hat{\gamma}_1^{(2)} \\ &= \bar{X}^\top\beta_1 + \hat{m}_{c,1}^\top\gamma_1 + \bar{X}^\top\mathbf{B}_0^\top\hat{\gamma}_1^{(2)} + (\hat{\tau}_2^{(1)})^\top\mathbf{U}_c^{(1)}\left(\hat{\gamma}_1^{(2)} - \gamma_1\right) - \hat{m}_{c,1}^\top\hat{\gamma}_1^{(2)} \\ &= \bar{X}^\top\beta_1 + \bar{X}^\top\mathbf{B}_0^\top\hat{\gamma}_1^{(2)} + \left[\hat{m}_{c,1} - (\mathbf{U}_c^{(1)})^\top\hat{\tau}_2^{(1)}\right]^\top\left(\gamma_1 - \hat{\gamma}_1^{(2)}\right) \\ &= \bar{X}^\top\beta_1 + \bar{X}^\top\mathbf{B}_0^\top\hat{\gamma}_1^{(2)} + \left(\mathbf{B}_0\mathbf{X}_{c,1}^\top\hat{\tau}_2^{(1)}\right)^\top\left(\gamma_1 - \hat{\gamma}_1^{(2)}\right) \\ &= \bar{X}^\top\left(\beta_1 + \mathbf{B}_0^\top\gamma_1\right) + \left(\bar{X} - \mathbf{X}_{c,1}^\top\hat{\tau}_2^{(1)}\right)^\top\mathbf{B}_0^\top\left(\hat{\gamma}_1^{(2)} - \gamma_1\right) \\ &= \theta_{10} + \left(\bar{X} - \mathbf{X}_{c,1}^\top\hat{\tau}_2^{(1)}\right)^\top\mathbf{B}_0^\top\left(\hat{\gamma}_1^{(2)} - \gamma_1\right).\end{aligned}$$

Substituting this identity into (A.1), we obtain

$$\begin{aligned}\hat{\theta}_{10}^{1\to 2} - \theta_{10} &= \left(a^{(1)} - \mathbf{W}_{t,2}^\top\hat{\tau}_1^{(2)}\right)^\top\left(\hat{\phi}^{(2)} - \phi_1\right) \\ &\quad + \left(\bar{X} - \mathbf{X}_{c,1}^\top\hat{\tau}_2^{(1)}\right)^\top\left(\hat{b}^{(1)} - b^{(2)}\right) \\ &\quad + \left(\bar{X} - \mathbf{X}_{c,1}^\top\hat{\tau}_2^{(1)}\right)^\top\mathbf{B}_0^\top\left(\hat{\gamma}_1^{(2)} - \gamma_1\right) \\ &\quad + (\hat{\tau}_1^{(2)})^\top\boldsymbol{\epsilon}_t^{(2)} + (\hat{\tau}_2^{(1)})^\top\mathbf{U}_c^{(1)}\gamma_1 \\ &=: \Delta_n^{(1)} + V_n^{(1)}.\end{aligned}$$

By the same argument with the roles of $\mathcal{D}_1$ and $\mathcal{D}_2$ reversed, we have

$$\hat{\theta}_{10}^{2\to 1} - \theta_{10} = \left(a^{(2)} - \mathbf{W}_{t,1}^\top\hat{\tau}_1^{(1)}\right)^\top\left(\hat{\phi}^{(1)} - \phi_1\right)$$

$$
\begin{aligned}
&+ \left(\bar{X} - \mathbf{X}_{c,2}^\top \hat{\tau}_2^{(2)}\right)^\top \left(\hat{b}^{(2)} - b^{(1)}\right) \\
&+ \left(\bar{X} - \mathbf{X}_{c,2}^\top \hat{\tau}_2^{(2)}\right)^\top \mathbf{B}_0^\top \left(\hat{\gamma}_1^{(1)} - \gamma_1\right) \\
&+ (\hat{\tau}_1^{(1)})^\top \boldsymbol{\epsilon}_t^{(1)} + (\hat{\tau}_2^{(2)})^\top \mathbf{U}_c^{(2)} \gamma_1 \\
&=: \Delta_n^{(2)} + V_n^{(2)}.
\end{aligned}
$$

Letting $j' = 3 - j$, for $j \in \{1, 2\}$,

$$
\begin{aligned}
\Delta_n^{(j)} = &\left(a^{(j)} - \mathbf{W}_{t,j'}^\top \hat{\tau}_1^{(j')}\right)^\top \left(\hat{\phi}^{(j')} - \phi_1\right) \\
&+ \left(\bar{X} - \mathbf{X}_{c,j}^\top \hat{\tau}_2^{(j)}\right)^\top \left(\hat{b}^{(j)} - b^{(j')}\right) \\
&+ \left(\bar{X} - \mathbf{X}_{c,j}^\top \hat{\tau}_2^{(j)}\right)^\top \mathbf{B}_0^\top \left(\hat{\gamma}_1^{(j')} - \gamma_1\right),
\end{aligned}
$$

which gives (4.1). Since $\hat{\theta}_{10}^{\mathrm{cf}} = \frac{1}{2}\left(\hat{\theta}_{10}^{1\to 2} + \hat{\theta}_{10}^{2\to 1}\right)$, we have

$$
\begin{aligned}
\hat{\theta}_{10}^{\mathrm{cf}} - \theta_{10} &= \frac{1}{2}\left[\left(\hat{\theta}_{10}^{1\to 2} - \theta_{10}\right) + \left(\hat{\theta}_{10}^{2\to 1} - \theta_{10}\right)\right] \\
&= \frac{\Delta_n^{(1)} + \Delta_n^{(2)}}{2} + \frac{V_n^{(1)} + V_n^{(2)}}{2} \\
&= \Delta_n^{\mathrm{cf}} + V_n^{\mathrm{cf}}.
\end{aligned}
$$

Moreover, by the definitions of the fold-specific error, $(\hat{\tau}_1^{(j)})^\top \boldsymbol{\epsilon}_t^{(j)} = \sum_{\substack{i\in\mathcal{D}_j \\ A_i=1}} \hat{\tau}_{1,i}^{(j)} \epsilon_i$, $(\hat{\tau}_2^{(j)})^\top \mathbf{U}_c^{(j)} \gamma_1 = \sum_{\substack{i\in\mathcal{D}_j \\ A_i=0}} \hat{\tau}_{2,i}^{(j)} U_i^\top \gamma_1$. we obtain,

$$
V_n^{\mathrm{cf}} = \frac{1}{2}\sum_{j=1}^{2}\left[\sum_{\substack{i\in\mathcal{D}_j \\ A_i=1}} \hat{\tau}_{1,i}^{(j)} \epsilon_i + \sum_{\substack{i\in\mathcal{D}_j \\ A_i=0}} \hat{\tau}_{2,i}^{(j)} U_i^\top \gamma_1\right].
$$

The bound in (4.2) can be proved directly via applying Hölder's inequality to each of the three terms, yielding

$$
\begin{aligned}
|\Delta_n^{(j)}| \le &\left\|\mathbf{B}_0\left(\bar{X} - \mathbf{X}_{c,j}^\top \hat{\tau}_2^{(j)}\right)\right\|_\infty \left\|\hat{\gamma}_1^{(j')} - \gamma_1\right\|_1 \\
&+ \left\|a^{(j)} - \mathbf{W}_{t,j'}^\top \hat{\tau}_1^{(j')}\right\|_\infty \left\|\hat{\phi}^{(j')} - \phi_1\right\|_1 \\
&+ \left\|\bar{X} - \mathbf{X}_{c,j}^\top \hat{\tau}_2^{(j)}\right\|_\infty \left\|\hat{b}^{(j)} - b^{(j')}\right\|_1,
\end{aligned}
$$

which proves (4.2). This completes the proof. □

B Proof of Results from Section 5

B.1 Proof of Lemma 5.8

Proof. Let us consider the optimization problem Equation (3.1) on $\mathcal{D}_1$, and the proof for $\mathcal{D}_2$ is the same. To establish the feasibility, we will construct a vector τ_2^0 which satisfies the constraints in Equation (3.1). Towards that end, define:

$$
\tau_{2,i}^0 = \tfrac{1}{n_{c,1}} \mu_X^\top \Sigma_{X,0}^{-1}(X_i - \mu_{X,0}), \qquad i \in \mathcal{D}_1 \text{ with } A_i = 0\,.
$$

Here $\Sigma_{X,0} = \mathsf{var}(X \mid A = 0)$ and $\mu_{X,0} = \mathbb{E}[X \mid A = 0]$. From Assumption 5.3 we have $\mu_X^\top \Sigma_{X,0}^{-1} \mu_X < \infty$, which, along with the fact that $\lambda_{\min}(\Sigma_{X,0}) \ge c_X > 0$ implies $\mu_X^\top \Sigma_{X,0}^{-2} \mu_X \le C$ for some constant C. Furthermore, by Assumption 5.3, $\zeta_i = \mu_X^\top \Sigma_{X,0}^{-1}(X_i - \mu_{X,0})$ is a centered sub-Gaussian random variable with sub-Gaussian

constant $\mu_X^\top \Sigma_{X,0}^{-2} \mu_X$. Therefore, from the standard sub-Gaussian inequality, we have with probability going to 1:

$$\max_i |\zeta_i| \le C_1 \sqrt{\log n_{c,1}} \implies \|\tau_2^0\|_\infty \le C \frac{\sqrt{\log n_{c,1}}}{n_{c,1}} \le n_{c,1}^{-\alpha}$$

for any $\alpha \in (0,1)$ and for all large n. For the other constraint:

$$\begin{aligned}\|\bar{X} - \mathbf{X}_{c,1}^\top \tau_2^0\|_\infty &= \left\| \bar{X} - \frac{1}{n_{c,1}} \mathbf{X}_{c,1}^\top (\mathbf{X}_{c,1} - \mathbf{1}_{n_{c,1}} \mu_{X,0}^\top) \Sigma_{X,0}^{-1} \mu_X \right\|_\infty \\ &\le \|\bar{X} - \mu_X\|_\infty + \left\| \mu_X - \frac{1}{n_{c,1}} \mathbf{X}_{c,1}^\top (\mathbf{X}_{c,1} - \mathbf{1}_{n_{c,1}} \mu_{X,0}^\top) \Sigma_{X,0}^{-1} \mu_X \right\|_\infty\end{aligned}$$

The first term on the right-hand side is again $\le C\sqrt{\log p/n}$ with probability going to 1 via simple sub-Gaussian concentration. Now we establish a bound on the second term on the right-hand-side of the above equation. For notational simplicity set $v = \Sigma_{X,0}^{-1} \mu_X$ and $Z_i = X_i - \mu_{X,0}$. Then

$$\tfrac{1}{n_{c,1}} \mathbf{X}_c^\top \left(\mathbf{X}_c - \mathbf{1}_{n_{c,1}} \mu_{X,0}^\top \right) \Sigma_{X,0}^{-1} \mu_X = \tfrac{1}{n_{c,1}} \textstyle\sum_{i=1}^{n_{c,1}} X_i Z_i^\top v.$$

Furthermore, we have:

$$\mathbb{E}[X_i Z_i^\top \mid A_i = 0] = \mathbb{E}[(\mu_{X,0} + Z_i) Z_i^\top \mid A_i = 0] = \Sigma_{X,0}$$

and consequently $\mathbb{E}[X_i Z_i^\top v \mid A = 0] = \mu_X$. Therefore, defining $\tilde{Z}_i = X_i Z_i^\top v - \mu_X$, we have:

$$\tfrac{1}{n_{c,1}} \mathbf{X}_c^\top \left(\mathbf{X}_c - \mathbf{1}_{n_{c,1}} \mu_{X,0}^\top \right) \Sigma_{X,0}^{-1} \mu_X - \mu_X = \tfrac{1}{n_{c,1}} \textstyle\sum_{i=1}^{n_{c,1}} \tilde{Z}_i,$$

It is immediate that $\tilde{Z}_1, \cdots, \tilde{Z}_{n_{c,1}}$ are p-dimensional independent centered random vectors conditional on the treatment indicators. Next, we argue that $\tilde{Z}_i$ is a sub-exponential random vector. Towards that end, define the whitened covariate vector $P_i := \Sigma_{X,0}^{-1/2}(X_i - \mu_{X,0})$ for $1 \le i \le n_{c,1}$ and also define $z = \Sigma_{X,0}^{-1/2} \mu_X$ and $u_j = \Sigma_{X,0}^{1/2} e_j$ for $1 \le j \le p$. Then, we have $Z_i = X_i - \mu_{X,0} = \Sigma_{X,0}^{1/2} P_i$ which further implies $Z_i^\top v = P_i^\top z$ and $Z_{ij} = P_i^\top u_j$ for $1 \le j \le p$. As a consequence,

$$X_{ij} Z_i^\top v = \left(\mu_{X,0,j} + P_i^\top u_j \right) (P_i^\top z).$$

Moreover, since $\mathbb{E}[P_i P_i^\top \mid A_i = 0] = I_p$, we also have:

$$\mathbb{E}[(P_i^\top u_j)(P_i^\top z) \mid A_i = 0] = u_j^\top z = e_j^\top \mu_X = \mu_{X,j}.$$

Therefore,

$$\tilde{Z}_{ij} = \mu_{X,0,j} P_i^\top z + (P_i^\top u_j)(P_i^\top z) - \mathbb{E}[(P_i^\top u_j)(P_i^\top z) \mid A_i = 0]. \tag{B.1}$$

Let $K_X < \infty$ denote the uniform conditional sub-Gaussian norm bound for $X - \mu_{X,a}$ in Assumption 5.3. Since $\lambda_{\min}(\Sigma_{X,0}) \ge c_X$, the vectors P_i are uniformly conditionally sub-Gaussian with

$$\|P_i\|_{\psi_2 | A_i = 0} \le K_X / \sqrt{c_X} =: K_P.$$

Furthermore, Assumption 5.3 implies $\|u_j\|_2^2 \le e_j^\top \Sigma_{X,0} e_j \le C_X$, and $\|z\|_2^2 = \mu_X^\top \Sigma_{X,0}^{-1} \mu_X \le C_{\mu,\Sigma}$. Therefore,

$$\|P_i^\top u_j\|_{\psi_2 | A_i = 0} \le K_P \sqrt{C_X}, \qquad \|P_i^\top z\|_{\psi_2 | A_i = 0} \le K_P \sqrt{C_{\mu,\Sigma}}.$$

Since the product of two sub-Gaussian random variables is sub-exponential, we have:

$$\|(P_i^\top u_j)(P_i^\top z)\|_{\psi_1 | A_i = 0} \le C K_P^2 \sqrt{C_X C_{\mu,\Sigma}}.$$

Also, since $|\mu_{X,0,j}| \le \|\mu_{X,0}\|_\infty \le C_\mu$, we have

$$\|\mu_{X,0,j} P_i^\top z\|_{\psi_1 | A_i = 0} \le C C_\mu K_P \sqrt{C_{\mu,\Sigma}}.$$

Using the standard centering inequality $\|Y - \mathbb{E}[Y]\|_{\psi_1} \leq C\|Y\|_{\psi_1}$, we conclude that

$$\max_{1\leq j\leq p} \|\widetilde{Z}_{ij}\|_{\psi_1|A_i=0} \leq C\left\{C_\mu K_P\sqrt{C_{\mu,\Sigma}} + K_P^2\sqrt{C_X C_{\mu,\Sigma}}\right\} =: L_X.$$

As L_X does not depend on j, we have $\max_{1\leq j\leq p} \|\tilde{Z}_{ij}\|_{\psi_1} \leq L_X$, and consequently an application of Bernstein's inequality yields, for $t > 0$,

$$\mathbb{P}\left(\left|\tfrac{1}{n_{c,1}}\textstyle\sum_{i=1}^{n_{c,1}} \tilde{Z}_{ij}\right| > t\right) \leq 2\exp\left[-cn_{c,1}\min\left\{\tfrac{t^2}{L_X^2}, \tfrac{t}{L_X}\right\}\right].$$

Taking a union bound over $j = 1, \ldots, p$ yields

$$\mathbb{P}\left(\left\|\tfrac{1}{n_{c,1}}\textstyle\sum_{i=1}^{n_{c,1}} \tilde{Z}_i\right\|_\infty > t\right) \leq 2\exp\left[\log p - cn_{c,1}\min\left\{\tfrac{t^2}{L_X^2}, \tfrac{t}{L_X}\right\}\right]$$

Finally, if we choose $t = AL_X\sqrt{\log p/n_{c,1}}$ for some sufficiently large constant $A > 0$, we have:

$$\mathbb{P}\left(\left\|\tfrac{1}{n_{c,1}}\textstyle\sum_{i=1}^{n_{c,1}} \tilde{Z}_i\right\|_\infty \leq AL\sqrt{\tfrac{\log p}{n_{c,1}}}\right) \longrightarrow 1, \quad \text{as } n \uparrow \infty.$$

This immediately implies, with probability going to 1:

$$\left\|\mu_X - \tfrac{1}{n_{c,1}}\mathbf{X}_{c,1}^\top(\mathbf{X}_{c,1} - \mathbf{1}_{n_{c,1}}\mu_{X,0}^\top)\Sigma_{X,0}^{-1}\mu_X\right\|_\infty \leq AL_X\sqrt{\tfrac{\log p}{n_{c,1}}}.$$

As a consequence, τ_2^0 is a feasible solution and consequently the optimization problem in Equation (3.1) is feasible. Furthermore, from the optimality of $\hat{\tau}_2^{(1)}$, we have $n\|\hat{\tau}_2^{(1)}\|_2^2 \leq n\|\tau_2^0\|_2^2$. On the other hand, from the definition,

$$\begin{aligned} n\|\tau_2^0\|_2^2 &= \frac{n}{n_{c,1}}\frac{1}{n_{c,1}}\sum_{\substack{i\in\mathcal{D}_1\\A_i=0}} v^\top(X_i - \mu_{X,0})(X_i - \mu_{X,0})^\top v \\ &= \frac{nn_1}{n_{c,1}^2}\frac{1}{n_1}\sum_{i\in\mathcal{D}_1} v^\top(X_i - \mu_{X,0})(X_i - \mu_{X,0})^\top v\mathbb{1}_{A_i=0} = \frac{n}{n_{c,1}}v^\top\hat{\Sigma}_{X,0}^{(1)}v \end{aligned}$$

Now,

$$\begin{aligned} &\mathbb{E}\left[\frac{1}{n_1}\sum_{i\in\mathcal{D}_1}(X_i - \mu_{X,0})(X_i - \mu_{X,0})^\top\mathbb{1}_{A_i=0}\right] \\ &= \mathbb{E}\left[(X - \mu_{X,0})(X - \mu_{X,0})^\top\mathbb{1}_{A=0}\right] = \Sigma_{X,0}\mathbb{P}(A = 0) \end{aligned}$$

As $\Sigma_{X,0}$ has bounded maximum eigenvalue and $\|v\|_2^2$ is bounded,

$$\mathbb{E}\left[\tfrac{1}{n_1}\textstyle\sum_{i\in\mathcal{D}_1} v^\top(X_i - \mu_{X,0})(X_i - \mu_{X,0})^\top v\mathbb{1}_{A_i=0}\right] \leq C.$$

Therefore, an application of Markov inequality yields:

$$\tfrac{1}{n_1}\textstyle\sum_{i\in\mathcal{D}_1} v^\top(X_i - \mu_{X,0})(X_i - \mu_{X,0})^\top v\mathbb{1}_{A_i=0} = \tfrac{n}{n_{c,1}}v^\top\hat{\Sigma}_{X,0}^{(1)}v = O_p(1)$$

This, along with the fact that $n \sim n_1 \sim n_{c,1}$ (as $\mathbb{P}(A = 1)$ is bounded away from 0 and 1 and each split has half of the samples), yields the conclusion.

Next, we will establish the feasibility of the optimization problem Equation (3.2). Towards that end recall the definition of η^* and a^* from part (3) of Assumption 5.3; $a^* = (\mu_X^\top \quad (\mathbf{B}_0\mu_X)^\top)^\top$ and $\eta^* = \Omega_{W,1}^{-1}a^*$, where $\Omega_{W,1} = \mathbb{E}[WW^\top \mid A = 1]$. Define a vector τ_1^0 as:

$$\tau_{1,i}^0 = \tfrac{1}{n_{t,2}}W_i^\top\eta^*, \qquad i \in \mathcal{D}_2, \text{ with } A_i = 1.$$

Therefore, we have $\tau_1^0 = (\mathbf{W}_{t,2}\eta^*)/n_{t,2}$. Hence, we have:

$$a - \mathbf{W}_{t,2}^\top \tau_1^0 = a - \tfrac{1}{n_{t,2}} \mathbf{W}_{t,2}^\top \mathbf{W}_{t,2}\eta^* = a - \hat{\Omega}_{n,W,t}\eta^* \,.$$

As a consequence,

$$\begin{aligned}
&\left\|a - \mathbf{W}_{t,2}^\top \tau_1^0\right\|_\infty \\
&= \left\|a - a^* + a^* - a - \mathbf{W}_{t,2}^\top \tau_1^0\right\|_\infty \\
&= \left\|a - a^* + (\Omega_{W,t} - \hat{\Omega}_{n,W,t})\eta^*\right\|_\infty \\
&\le \|a - a^*\|_\infty + \|\Omega_{W,t} - \hat{\Omega}_{n,W,t}\|_{\infty,\infty}\|\eta^*\|_1 \\
&\le \|\bar{X} - \mu_X\|_\infty + \|\hat{m}_{c,1} - \mathbf{B}_0\mu_X\|_\infty + C_\eta\|\Omega_{W,t} - \hat{\Omega}_{n,W,t}\|_{\infty,\infty} \\
&\le (1 + \|\mathbf{B}_0\|_{1,\infty})\|\bar{X} - \mu_X\|_\infty + \|\hat{m}_{c,1} - \mathbf{B}_0\bar{X}\|_\infty + C_\eta\|\Omega_{W,t} - \hat{\Omega}_{n,W,t}\|_{\infty,\infty} \\
&\le C\|\bar{X} - \mu_X\|_\infty + \|\mathbf{B}_0(\mathbf{X}_{c,1}^\top \hat{\tau}_2^{(1)} - \bar{X})\|_\infty + \|(\mathbf{U}_c^{(1)})^\top \hat{\tau}_2^{(1)}\|_\infty \\
&\qquad\qquad + C_\eta\|\Omega_{W,t} - \hat{\Omega}_{n,W,t}\|_{\infty,\infty} \\
&\le C\left(\|\bar{X} - \mu_X\|_\infty + \|(\mathbf{X}_{c,1}^\top \hat{\tau}_2^{(1)} - \bar{X})\|_\infty + \|\Omega_{W,t} - \hat{\Omega}_{n,W,t}\|_{\infty,\infty}\right) + \|(\mathbf{U}_c^{(1)})^\top \hat{\tau}_2^{(1)}\|_\infty
\end{aligned}$$

In the previous inequalities, we have used the fact that $\|\eta^*\|_1 \le C_\eta$ (Assumption 5.3) and $\|\mathbf{B}_0\|_{1,\infty} \le C_{\mathbf{B}}$ (Assumption 5.4). The first and second summands on the right-hand side are $\le C\sqrt{\log p/n}$ for a large enough constant C with probability going to 1 by standard sub-Gaussian concentration for $\bar{X}$ and the definition of $\hat{\tau}_2^{(1)}$ (Equation (3.1)). For the fourth term, we use the fact that conditionally on $\mathbf{X}_c$, $U_{c,i}$'s independent sub-Gaussian vector with sub-Gaussian constants uniformly bounded by σ_U. Therefore, conditionally on $\mathbf{X}$,

$$\|(\mathbf{U}_c^{(1)})^\top \hat{\tau}_2^{(1)}\|_\infty = O_p\left(\|\hat{\tau}_2^{(1)}\|_2 \sqrt{\log q}\right) = O_p\left(\sqrt{\tfrac{\log q}{n}}\right)$$

where the last conclusion follows from the fact that $n\|\hat{\tau}_2^{(1)}\|_2^2 = O_p(1)$ as established previously. Last but not least $\|\Omega_{W,t} - \hat{\Omega}_{n,W,t}\|_{\infty,\infty} = O_p(\sqrt{\log(p+q)/n})$ which follows from standard sub-exponential Bernstein inequality, along with a union bound on the co-ordinates of $\hat{\Omega}_{n,W,t}$. Therefore, we conclude:

$$\left\|a^{(1)} - \mathbf{W}_{t,2}^\top \tau_1^0\right\|_\infty \le K_1\sqrt{\tfrac{\log(p+q)}{n_{t,2}}}$$

with probability going to 1 for large enough constant K_1. We next argue that $\|\tau_1^0\|_\infty \le n_{t,2}^{-\alpha}$. Towards that end:

$$\begin{aligned}
\max_{1\le i\le n_{t,2}} |\tau_{1,i}^0| &= \frac{1}{n_{t,2}} \max_{1\le i\le n_{t,2}} |W_i^\top \eta^*| \\
&\le \frac{|\mu_{W,1}^\top \eta^*|}{n_t} + \sqrt{\lambda_{\max}(\Sigma_{W,1})}\, \frac{1}{n_{t,2}} \max_{1\le i\le n_{t,2}} |\Sigma_{W,1}^{-1/2}(W_i - \mu_{W,1})^\top \eta^*| \\
&\le O_p\left(\frac{1}{n_{t,2}}\right) + CO_p\left(\frac{\sqrt{\log n_{t,2}}}{n_{t,2}}\right) \le n_{t,2}^{-\alpha}
\end{aligned}$$

for any $\alpha \in (0,1)$. Here we have used the fact that $|\mu_{W,1}^\top \eta^*| \le C$ which follows from the fact that $\|\mu_{W,1}\|_\infty \le C_\mu$, $\|\eta^*\|_1 \le C_\eta$, $\lambda_{\max}(\Sigma_{W,1}) \le C$ and $\Sigma_{W,1}^{-1/2}(W_i - \mu_{W,1})$ is sub-Gaussian (Assumption 5.3 and Remark 5.6). Finally, we will show that $n\|\hat{\tau}_1^{(2)}\|_2^2 = O_p(1)$ by showing that $n\|\tau_1^0\|_2^2 = O_p(1)$ and then using the fact that $n\|\hat{\tau}_1^{(2)}\|_2^2 \le n\|\tau_1^0\|_2^2$ as $\hat{\tau}_1^{(2)}$ is the optimal solution of Equation (3.2), whereas τ_1^0 is only a feasible solution. Now for τ_1^0, we have:

$$n\|\tau_1^0\|_2^2 = \tfrac{nn_2}{n_{t,2}^2}\, \tfrac{1}{n_2} \textstyle\sum_{i\in\mathcal{D}_2} (W_i^\top \eta^*)^2 \mathbb{1}_{A_i=1} \le (\eta^*)^\top \hat{\Omega}\eta^* \le \|\hat{\Omega}\|_\infty \|\eta^*\|_1^2 \,.$$

Now $\|\hat{\Omega}\|_\infty = O_p(1)$ as $\|\mathbb{E}[WW^\top \mid A = 1]\|_\infty \le \|\Sigma_{W,1}\|_\infty + \|\mu_{W,1}\|_\infty^2$. (Assumption 5.3). Furthemore, $\|\eta^*\|_1 \le C_\eta$ by Assumption 5.3. Hence, $n\|\tau_1^0\|_2^2 = O_p(1)$. This completes the proof.

□

B.2 Proof of Theorem 5.9

Proof. From Proposition 4.1, we have:

$$\frac{\sqrt{n}(\hat{\theta}_{10}^{\text{cf}} - \theta_{10})}{\sigma_n^{\text{cf}}} = \frac{\sqrt{n}(\Delta_n^{\text{cf}} + V_n^{\text{cf}})}{\sigma_n^{\text{cf}}}\,. \tag{B.2}$$

Therefore, we divide the rest of the proof into three parts:

a) We show that $\sqrt{n}\Delta_n^{\text{cf}} = o_p(1)$,

b) We show that $\sigma_n^{cf\cdot}$ is bounded away from 0 with probability going to 1.

c) We show that $\sqrt{n}V_n^{\text{cf}}/\sigma_n^{\text{cf}}$ converges to $\mathcal{N}(0,1)$.

Asymptotic negligibility of $\sqrt{n}\Delta_n^{\text{cf}}$. We show that $\sqrt{n}|\Delta_{n,1}| = o_p(1)$. The proof for $\Delta_{n,2}$ is same and hence skipped. For our proof we will use upper on $|\Delta_{n,1}|$ as established in Proposition 4.1. By standard analysis to establish the convergence rate of the Lasso estimator with respect to the ℓ_1-norm under the sparsity constraint (e.g., (Negahban et al., 2012, Corollary 2)), we have the following (we omit the notation j from the next paragraph for notational simplicity, as the argument holds for each $j \in \{1,2\}$):

$$\|\hat{\phi}_1 - \phi_1\|_1 = \mathcal{O}_p\left(k\sqrt{\tfrac{\log(p+q)}{n_t}}\right), \quad \|\hat{b} - b\|_1 = \mathcal{O}_p\left(sk_\gamma\sqrt{\tfrac{\log(p)}{n_c}} + k\sqrt{\tfrac{\log(p+q)}{n_t}}\right). \tag{B.3}$$

Note that in order to apply Corollary 2 of Negahban et al. (2012), we need the design matrices to satisfy the RE condition, and the columns should be approximately standardized. Regarding the RE condition, we need $(\mathbf{W}_t^\top \mathbf{W}_t)/n_t$ to satisfy the RE condition with high probability for establishing the bound on $\|\hat{\phi}_1 - \phi_1\|_1$ and $(\mathbf{X}_c^\top \mathbf{X}_c)/n_c$ to satisfy the RE condition with high probability to establish the bound on $\|\hat{b} - b\|_1$. For the former, note that, by Assumption 5.3, the population second moment matrix $\mathbb{E}[WW^\top \mid A = 1] = \Sigma_{W,t} + \mu_{W,t}\mu_{W,t}^\top \succeq \Sigma_{W,t}$ has minimum eigenvalue bounded away from 0 and $\mathbf{W}_t$ has i.i.d. sub-Gaussian rows (Assumption 5.3 and Remark 5.6). Therefore, by (Rudelson and Zhou, 2012, Theorem 6), we can conclude that $(\mathbf{W}_t^\top \mathbf{W}_t)/n_t$ also satisfies RE condition with probability going to 1. Similar logic also applies for $(\mathbf{X}_c^\top \mathbf{X}_c)/n_c$ and its population second moment matrix $\mathbb{E}[XX^\top \mid A = 0]$ as minimum eigenvalues bounded away from 0 and $\mathbf{X}_c$ has i.i.d. sub-Gaussian rows (Assumption 5.3). Now, as for column normalization, we need $\|\mathbf{W}_{t,*j}\|_2/\sqrt{n_t} \le C$ and $\|\mathbf{X}_{c,*j}\|_2/\sqrt{n_c} \le C$ for some constant $C > 0$ with probability going to 1. This follows from standard sub-Gaussian concentration; For $\mathbf{W}_t$, fix $1 \le j \le p+q$ and set

$$\mu_{W,t,j}^{(2)} = (\Sigma_{W,t} + \mu_{W,t}\mu_{W,t}^\top)_{jj} = \mathbb{E}[W_{t,ij}^2]\,.$$

Therefore, we have:

$$\begin{aligned}\max_{1\le j\le p+q} \frac{1}{n_t}\sum_{i\in\mathcal{D}_2: A_i=1} W_{t,ij}^2 &\le \max_{1\le j\le p+q}\left|\frac{1}{n_t}\sum_{i\in\mathcal{D}_2: A_i=1}(W_{t,ij}^2 - \mu_{W,t,j}^{(2)})\right| + \max_{1\le j\le p+q}\mu_{W,t,j}^{(2)} \\ &\le \max_{1\le j\le p+q}\mu_{W,t,j}^{(2)} + C\sqrt{\frac{\log(p+q)}{n}},\end{aligned} \tag{B.4}$$

with probability going to 1. The last bound follows from a standard sub-Gaussian concentration argument. We next verify that $\max_{1\le j\le p+q}\mu_{W,t,j}^{(2)}$ is uniformly bounded. Recall that $\mu_{W,t,j}^{(2)} = (\Sigma_{W,t})_{jj} + \mu_{W,t,j}^2$. Therefore, by Remark 5.6 and Assumptions 5.3 and 5.4, $\lambda_{\max}(\Sigma_{W,t}) \le C_W := (1+C_B)^2(C_X \vee C_U)$, where C_X, C_U, and C_B are the constants as defined in Assumptions 2 and 3. Moreover, since $\mu_{W,t} = (\mu_{X,1}^\top, (B_1\mu_{X,1})^\top)^\top$, we have

$$\|\mu_{W,t}\|_\infty \le \max\{\|\mu_{X,1}\|_\infty,\ \|B_1\|_{1,\infty}\|\mu_{X,1}\|_\infty\} \le (1 \vee C_B)C_\mu =: C_{\mu,W}.$$

As a consequence, $\max_{1\le j\le p+q}\mu_{W,t,j}^{(2)} \le C_W + C_{\mu,W}^2 := C_{W,2} < \infty$. Therefore, Equation (B.4) implies that, with probability going to one,

$$\max_{1\le j\le p+q}\tfrac{1}{n_t}\textstyle\sum_{\substack{i\in\mathcal{D}_2\\ A_i=1}} W_{t,ij}^2 \le 2C_{W,2} \implies \max_{1\le j\le p+q}\tfrac{\|\mathbf{W}_{t,*j}\|_2}{\sqrt{n_t}} \le \sqrt{2C_{W,2}}\,,$$

with probability going to one. Similarly, for the control-arm covariate design, $\mathbb{E}[X_j^2 \mid A = 0] = (\Sigma_{X,0})_{jj} + \mu_{X,0,j}^2 \le C_X + C_\mu^2 := C_{X,2}$. The same concentration argument therefore gives $\max_{1\le j\le p}\|\mathbf{X}_{c,*j}\|_2/\sqrt{n_c} \le$

$\sqrt{2C_{X,2}}$ with probability going to one. Therefore, conditions required for Corollary 2.1 of Negahban et al. (2012) are satisfied by the design matrices $\mathbf{W}_t$ and $\mathbf{X}_c$, and consequently Equation (B.3) holds. The rate for $\|\hat{b} - b\|_1$ needs some additional calculation as $b = \mathbf{B}_0^\top \hat{\gamma}_1$ may not be exactly sparse and is deferred to Lemma C.3. Now, going back to $\Delta_{n,1}$, we have, from (4.2):

$$\begin{aligned}\sqrt{n}|\Delta_{n,1}| \leq \sqrt{n}\|\mathbf{B}_0(\bar{X} - \mathbf{X}_{c,j}^\top \hat{\tau}_2^{(1)})\|_\infty \|\hat{\gamma}_1^{(2)} - \gamma_1\|_1 \\ + \sqrt{n}\|a^{(1)} - \mathbf{W}_{t,2}^\top \hat{\tau}_1^{(2)}\|_\infty \|\hat{\phi}_1^{(2)} - \phi_1\|_1 \\ + \sqrt{n}\|\bar{X} - \mathbf{X}_{c,1}^\top \hat{\tau}_2^{(1)}\|_\infty \|\hat{b}^{(1)} - b^{(2)}\|_1 := T_1 + T_2 + T_3\,.\end{aligned}$$

Now, by definition of $\hat{\tau}_2^{(1)}$ (Step 1 of Algorithm 1), and the fact that $\|\mathbf{B}_0\|_{1,\infty} \leq C_{\mathbf{B}}$ (Assumption 5.4), we have:

$$\begin{aligned}T_1 &\leq \sqrt{n}\|\mathbf{B}_0\|_{1,\infty}\|\bar{X} - \mathbf{X}_{c,j}^\top \hat{\tau}_2^{(j)}\|_\infty \|\hat{\gamma}_1^{(2)} - \gamma_1\|_1 \\ &\leq C_{\mathbf{B}} K_2 \sqrt{n}\sqrt{\frac{\log p}{n_{c,1}}}\|\hat{\gamma}_1^{(2)} - \gamma_1\|_1 \\ &= O_p\left(\sqrt{n} \times \sqrt{\frac{\log p}{n_{c,1}}} \times k\sqrt{\frac{\log(p+q)}{n_t}}\right) = o_p(1)\end{aligned}$$

The last conclusion follows from Assumption 5.4. Now, from Equation (3.2), we have $\|a^{(1)} - \mathbf{W}_{t,2}^\top \hat{\tau}_1^{(2)}\|_\infty \leq K_1\sqrt{\log(p+q)/n_{t,2}}$. Therefore, this, along with Equation (B.3) implies:

$$\begin{aligned}T_2 &= \sqrt{n}\|a^{(1)} - \mathbf{W}_{t,2}^\top \hat{\tau}_1^{(2)}\|_\infty \|\hat{\phi}_1^{(2)} - \phi_1\|_1 \\ &= O_p\left(\sqrt{n} \times \sqrt{\frac{\log(p+q)}{n}} \times k\sqrt{\frac{\log(p+q)}{n}}\right) = O_p\left(\frac{k\log(p+q)}{\sqrt{n}}\right) = o_p(1),\end{aligned}$$

where the last conclusion follows from Assumption 5.4. Similarly for T_3, we have $\|\bar{X} - \mathbf{X}_{c,1}^\top \hat{\tau}_2^{(1)}\|_\infty \leq K_2\sqrt{\log p/n_{c,1}}$ from Equation (3.1). Therefore, this and Equation (B.3) implies:

$$\begin{aligned}T_3 &= \sqrt{n}\|\bar{X} - \mathbf{X}_c^\top \tau_2\|_\infty \|\hat{b} - b\|_1 \\ &= O_p\left(\sqrt{n} \times \sqrt{\frac{\log(p)}{n}} \times \left(sk_\gamma\sqrt{\frac{\log p}{n}} + k\sqrt{\frac{\log(p+q)}{n}}\right)\right) \\ &= O_p\left(\frac{sk_\gamma \log p}{\sqrt{n}} + \frac{k\log(p+q)}{\sqrt{n}}\right) = o_p(1),\end{aligned}$$

where, as before, the last conclusion follows from Assumption 5.4. This completes the proof that $\sqrt{n}|\Delta_{n,1}| = o_p(1)$.

Lower bound on σ_n^{cf}: Our next goal is to show that σ_n^{cf} remains bounded away from zero with probability tending to one. Recall that

$$(\sigma_n^{\mathrm{cf}})^2 = \frac{n}{4}\sum_{j=1}^{2}(\sigma_1^2\|\hat{\tau}_1^{(j)}\|_2^2 + \gamma_1^\top \Sigma_{U,0}\gamma_1\|\hat{\tau}_2^{(j)}\|_2^2) \geq \frac{n\sigma_1^2\|\hat{\tau}_1^{(1)}\|_2^2}{4}\,. \tag{B.5}$$

Therefore, we next show that $n\|\hat{\tau}_1^{(1)}\|_2^2 \geq c$ for some constant $c > 0$ with probability going to 1. This, along with the fact that $\sigma_1^2 > 0$ (Assumption 5.3) will yield the conclusion. In fact, we will establish both $n\|\hat{\tau}_1^{(2)}\|_2^2$ and $n\|\hat{\tau}_2^{(1)}\|_2^2$ are bounded away from 0 with probability going to 1, which will be used in the later part of the proof. The conclusion for the balancing weights using another fold of the data (i.e., $(\hat{\tau}_1^{(1)}, \hat{\tau}_2^{(2)})$) is similar and hence skipped. We start with $n\|\hat{\tau}_1^{(2)}\|_2^2$. From the definition of $a^{(1)}$, we have:

$$\|a^{(1)}\|_\infty \geq \|\bar{X}\|_\infty \geq \|\mu_X\|_\infty - \|\bar{X} - \mu_X\|_\infty \geq \kappa - \|\bar{X} - \mu_X\|_\infty \geq \kappa/2$$

where the last inequality follows with probability approaching to 1 as $\|\bar{X} - \mu_X\|_\infty = O_p(\sqrt{\log p/n})$ from standard sub-Gaussian concentration and $\kappa > 0$. Here we have also used the fact that $\|\mu_X\|_\infty \geq \kappa$ (Assumption 5.3). Now on this event, there exists $1 \leq j \leq p+q$ such that $|a_j| \geq \kappa/2$ and $1 \leq j \leq p$ such that $|\bar{X}_j| \geq \kappa/2$. For any feasible solution τ_1 of Equation (3.2) and τ_2 of Equation (3.1) satisfy for all large n:

$$|(\mathbf{W}_{t,2}^\top \tau_1)_j| \geq |a_j^{(1)}| - |a_j^{(1)} - (\mathbf{W}_{t,2}^\top \tau_1)_j| \geq \kappa/2 - K_1\sqrt{\frac{\log(p+q)}{n_t}} \geq \kappa/4,$$
$$|(\mathbf{X}_{c,1}^\top \tau_2)_j| \geq |\bar{X}_j| - |\bar{X}_j - (\mathbf{X}_{c,1}^\top \tau_2)_j| \geq \kappa/2 - K_2\sqrt{\frac{\log p}{n_c}} \geq \kappa/4,$$

with probability tending to one, since $\log(p+q)/(n_t \wedge n_c) \to 0$. Applying Cauchy–Schwarz yields

$$\kappa_1^2 \leq (\mathbf{W}_{t,2}^\top \tau_1)_j^2 \leq \|\tau_1\|_2^2 \sum_{\substack{i \in \mathcal{D}_2 \\ A_i = 1}} W_{t,ij}^2, \qquad \kappa_2^2 \leq (\mathbf{X}_c^\top \tau_2)_j^2 \leq \|\tau_2\|_2^2 \sum_{\substack{i \in \mathcal{D}_1 \\ A_i = 0}} X_{c,ij}^2.$$

Therefore,

$$n\|\tau_1\|_2^2 \geq \frac{\kappa^2/16}{\frac{n_{t,2}}{n} \cdot \frac{1}{n_t} \sum_{\substack{i \in \mathcal{D}_2 \\ A_i = 1}} W_{t,ij}^2}, \qquad n\|\tau_2\|_2^2 \geq \frac{\kappa^2/16}{\frac{n_{c,1}}{n} \cdot \frac{1}{n_c} \sum_{\substack{i \in \mathcal{D}_1 \\ A_i = 0}} X_{c,ij}^2}.$$

For the denominator, $n_{t,2}/n \leq 1$ and $(\sum_{i:A_i=1} W_{t,ij}^2)/n_{t,2} \leq 2\mu_{W,t,j}^{(2)}$, which follows from the same calculation as in Equation (B.4). Therefore, we conclude that with probability going to 1,

$$n\|\hat{\tau}_1^{(2)}\|_2^2 \ \wedge \ n\|\hat{\tau}_2^{(1)}\|_2^2 \geq \frac{\kappa^2}{C} \quad \text{with probability going to 1.} \tag{B.6}$$

This completes the proof.

Asymptotic normality: Finally we will establish the asymptotic normality of $\sqrt{n}V_n^{\mathrm{cf}}/\sigma_n^{\mathrm{cf}}$. Recall that, from Proposition 4.1, we have:

$$\frac{\sqrt{n}V_n^{\mathrm{cf}}}{\sigma_n^{\mathrm{cf}}} = \frac{\sqrt{n}}{2\sigma_n^{\mathrm{cf}}} \left[\sum_{\substack{i \in \mathcal{D}_2 \\ A_i = 1}} \hat{\tau}_{1,i}^{(2)} \epsilon_i + \sum_{\substack{i \in \mathcal{D}_1 \\ A_i = 0}} \hat{\tau}_{2,i}^{(1)} U_i^\top \gamma_1 + \sum_{\substack{i \in \mathcal{D}_1 \\ A_i = 1}} \hat{\tau}_{1,i}^{(1)} \epsilon_i + \sum_{\substack{i \in \mathcal{D}_2 \\ A_i = 0}} \hat{\tau}_{2,i}^{(2)} U_i^\top \gamma_1 \right] := T$$

where we introduce the notation T for notational simplicity. For the rest of the analysis define $\tau_1^{(j),0}$ as *an oracle version of* $\hat{\tau}_1^{(j)}$ where $\tau_1^{(j),0}$ is obtained by solving optimization problem Equation (3.2) with replacing $a^{(j')}$ by $a^{(j'),0} \equiv a^0$ where $a^0 = (\bar{X}^\top \ (\mathbf{B}_0\bar{X})^\top)^\top$. In other words, we obtain a^0 by replacing the $\hat{m}_{c,j}$ part by its oracle counterpart $\mathbf{B}_0\bar{X}$. Note that now $\{\hat{\tau}_1^{(j),0}\}$ for $j \in \{1,2\}$ only depends on $\mathcal{H}_n := \sigma(A_{1:n}, X_{1:n}, \{M_i : A_i = 1\})$. Furthermore, by definition, $\{\hat{\tau}_2^{(j)}\}$ also is measurable function of $\sigma(A_{1:n}, X_{1:n})$, and consequently a measurable function of $\mathcal{H}_n$. Our following Lemma shows that $\hat{\tau}_1^{(j),0}$ is asymptotically close to $\hat{\tau}_1^{(j)}$:

Lemma B.1. *Under Assumption 5.3 and 5.4, we have:*

$$\sqrt{n}\|\hat{\tau}_1^{(j)} - \hat{\tau}_1^{(j),0}\|_2 = o_p(1), \quad j \in \{1,2\}.$$

Now, using this orcale $\hat{\tau}_1^{(j),0}$, we define:

$$T^0 = \frac{\sqrt{n}}{2\sigma_n^{\mathrm{cf},0}} \left[\sum_{\substack{i \in \mathcal{D}_2 \\ A_i = 1}} \hat{\tau}_{1,i}^{(2),0} \epsilon_i + \sum_{\substack{i \in \mathcal{D}_1 \\ A_i = 0}} \hat{\tau}_{2,i}^{(1)} U_i^\top \gamma_1 + \sum_{\substack{i \in \mathcal{D}_1 \\ A_i = 1}} \hat{\tau}_{1,i}^{(1),0} \epsilon_i + \sum_{\substack{i \in \mathcal{D}_2 \\ A_i = 0}} \hat{\tau}_{2,i}^{(2)} U_i^\top \gamma_1 \right]$$
$$:= \sum_{i=1}^n \zeta_{n,i}$$

where

$$\zeta_{n,i} = \frac{\sqrt{n}}{2\sigma_n^{\mathrm{cf},0}} \Big\{ \epsilon_i \left(\hat{\tau}_{1,i}^{(1),0} \mathbb{1}_{i\in\mathcal{D}_1, A_i=1} + \hat{\tau}_{1,i}^{(2),0} \mathbb{1}_{i\in\mathcal{D}_2, A_i=1} \right) + U_i^\top \gamma_1 \left(\hat{\tau}_{2,i}^{(1)} \mathbb{1}_{i\in\mathcal{D}_1, A_i=0} + \hat{\tau}_{2,i}^{(2)} \mathbb{1}_{i\in\mathcal{D}_2, A_i=0} \right) \Big\}$$

$$(\sigma_n^{\mathrm{cf},0})^2 = \frac{n}{4} \sum_{j=1}^{2} (\sigma_1^2 \|\hat{\tau}_1^{(j),0}\|_2^2 + \gamma_1^\top \Sigma_{U,0} \gamma_1 \|\hat{\tau}_2^{(j)}\|_2^2) \tag{B.7}$$

We next show that T^0 is asymptotically normal and then we will show that $T - T^0 = o_p(1)$. This would complete the proof. To establish that T^0 is asymptotically normal, first observe that conditional on $\mathcal{H}_n$ the weights $\{(\hat{\tau}_1^{(j),0}, \hat{\tau}_2^{(j)}) : j \in \{1,2\}\}$ are fixed and (U_i, ϵ_i) are independent across i. From $\mathbb{E}[\epsilon_i \mid X_i, M_i, A_i = 1] = 0$ and $\mathbb{E}[U_i \mid X_i, A_i = 0] = 0$ (Assumption 5.3), we also have $\mathbb{E}[\zeta_{n,i} \mid \mathcal{H}_n] = 0$ and

$$\begin{aligned} \mathsf{var}(T^0 \mid \mathcal{H}_n) &= \sum_i \mathsf{var}(\zeta_{n,i} \mid \mathcal{H}_n) \\ &= \frac{n}{4(\sigma_n^{\mathrm{cf},0})^2} \sum_{j=1}^{2} (\sigma_1^2 \|\hat{\tau}_1^{(j),0}\|_2^2 + \gamma_1^\top \Sigma_{U,0} \gamma_1 \|\hat{\tau}_2^{(j)}\|_2^2) = 1\,. \end{aligned}$$

Here we have used the fact that $\mathsf{var}(\epsilon_i \mid X_i, M_i, A_i = 1) = \sigma_1^2$ and $\mathsf{var}(U_i \mid X_i, A_i = 0) = \Sigma_{U,0}$ (Assumption 5.3). Now, we will verify Lyapunov's condition for the third moment, i.e., we show that $\sum_i \mathbb{E}[|\zeta_{n,i}|^3 \mid \mathcal{H}_n] \to 0$ in probability. Towards that end, we will use the fact that ϵ_i and $U_i^\top \gamma_1 / \|\gamma_1\|_2$ are sub-Gaussian (uniformly over (X, M, A)) and consequently they have finite third moment (Assumption 5.3). For notational simplicity, we will use C to denote these bounds, i.e. $\mathbb{E}[|\epsilon_i|^3 \mid \mathcal{H}_n] \le C$ and $\mathbb{E}[|U_i^\top \gamma_1|^3 \mid \mathcal{H}_n] \le C\|\gamma_1\|_2^3$. Using this we have:

$$\begin{aligned} &\sum_i \mathbb{E}[|\zeta_{n,i}|^3 \mid \mathcal{H}_n] \\ &\le C \frac{n^{3/2}}{8(\sigma_n^{\mathrm{cf},0})^3} \sum_{i=1}^{n} \Big[|\hat{\tau}_{1,i}^{(1),0}|^3 \mathbb{1}_{i\in D_1, A=1} + |\hat{\tau}_{1,i}^{(2),0}|^3 \mathbb{1}_{i\in D_2, Ai=1} \\ &\qquad\qquad + |\hat{\tau}_{2,i}^{(1)}|^3 \mathbb{1}_{i\in\mathcal{D}_1, A_i=0} + |\hat{\tau}_{2,i}^{(2)}|^3 \mathbb{1}_{i\in\mathcal{D}_2, A_i=0} \Big] \\ &\le C \frac{n^{3/2}}{8(\sigma_n^{\mathrm{cf},0})^3} \sum_{j=1}^{2} \left(\|\hat{\tau}_{1,i}^{(j),0}\|_2^2 \|\hat{\tau}_{1,i}^{(j),0}\|_\infty + \|\hat{\tau}_2^{(j)}\|_2^2 \|\hat{\tau}_{2,i}^{(j)}\|_\infty \right). \end{aligned}$$

In Lemma 5.8 we have established $n\|\hat{\tau}_{1,i}^{(j)}\|_2^2 = O_p(1)$ and $n\|\hat{\tau}_{2,i}^{(j)}\|_2^2 = O_p(1)$. This, along with Lemma B.1, implies $n\|\hat{\tau}_1^{(j),0}\|_2^2 = O_p(1)$. On the other hand, by definition of the optimization problem (Equation (3.1) and (3.2)), we have $\|\hat{\tau}_1^{(j),0}\|_\infty \vee \|\hat{\tau}_2^{(j)}\| \le Cn^{-\alpha}$ where we have used the fact that $n_{c,j} \sim n$ and $n_{t,j} \sim n$ with probability going to 1 as $\mathbb{P}(A = 1)$ is bounded away from 0 and 1, as well as each split has $n/2$ observations. (Recall that $\hat{\tau}_1^{(j),0}$ is obtained by solving Equation (3.2) with $a^{(j')}$ replaced by a^0, so by feasibility it satisfies $\|\hat{\tau}_1^{(j),0}\|_\infty \le n_{t,2}^{-\alpha} \le Cn^{-\alpha}$). Therefore, we can conclude:

$$n^{3/2} \textstyle\sum_{j=1}^{2} \left(\|\hat{\tau}_{1,i}^{(j),0}\|_2^2 \|\hat{\tau}_{1,i}^{(j),0}\|_\infty + \|\hat{\tau}_2^{(j)}\|_2^2 \|\hat{\tau}_{2,i}^{(j)}\|_\infty \right) = O_p(n^{-1-\alpha+3/2}) = o_p(1)$$

where the last conclusion follows from the fact that $\alpha > 1/2$. On the other hand, we have established in the previous part of the proof that σ_n^{cf} is bounded away from 0 with probability going to 1. We now argue that $\sigma_n^{\mathrm{cf},0}$ is also bounded away from 0 with high probability.

$$\left| (\sigma_n^{\mathrm{cf}})^2 - (\sigma_n^{\mathrm{cf},0})^2 \right| = \frac{n\sigma_1^2}{4} \sum_{j=1}^{2} \left| \|\hat{\tau}_1^{(j),0}\|_2^2 - \|\hat{\tau}_1^{(j)}\|_2^2 \right|$$

$$\leq \frac{n\sigma_1^2}{4}\sum_{j=1}^{2}\|\hat{\tau}_1^{(j),0}+\hat{\tau}_1^{(j)}\|_2\|\hat{\tau}_1^{(j),0}-\hat{\tau}_1^{(j)}\|_2$$
$$\leq \frac{\sigma_1^2}{4}\sum_{j=1}^{2}\left(\sqrt{n}(\|\hat{\tau}_1^{(j),0}\|_2+\|\hat{\tau}_1^{(j)}\|_2)\right)\left(\sqrt{n}\|\hat{\tau}_1^{(j),0}-\hat{\tau}_1^{(j)}\|_2\right)$$
$$= o_p(1)\,. \tag{B.8}$$

Here, the last conclusion follows from the fact that $\sqrt{n}\|\hat{\tau}_1^{(j),0}-\hat{\tau}_1^{(j)}\|_2 = o_p(1)$ (Lemma B.1) and $\sqrt{n}(\|\hat{\tau}_1^{(j),0}\|_2+\|\hat{\tau}_1^{(j)}\|_2) = O_p(1)$ (Lemma 5.8 and Lemma B.1). As a consequence, $\sigma_n^{\mathrm{cf},0}$ will be bounded away from 0 with probability going to 1. Hence, $\sum_i \mathbb{E}[|\zeta_{n,i}|^3 \mid \mathcal{H}_n] = o_p(1)$. Therefore, by an application of conditional central limit theorem Bulinski (2017), we conclude $T^0 \xrightarrow{d} \mathcal{N}(0,1)$.

Next we argue that $T - T^0 = o_p(1)$ which would establish that $T \xrightarrow{d} \mathcal{N}(0,1)$. Towards that end, write $T = L/\sigma_n^{\mathrm{cf}}$ and $T^0 = L^0/\sigma_n^{\mathrm{cf},0}$. Therefore, we can split the difference $T - T^0$ into two parts:

$$T - T^0 = \frac{L}{\sigma_n^{\mathrm{cf}}} - \frac{L^0}{\sigma_n^{\mathrm{cf},0}} = \frac{L-L^0}{\sigma_n^{\mathrm{cf}}} + \frac{L^0}{\sigma_n^{\mathrm{cf},0}}\left(\frac{\sigma_n^{\mathrm{cf},0}}{\sigma_n^{\mathrm{cf}}} - 1\right) := T_1 + T_2\,. \tag{B.9}$$

Our goal is to shwo that both T_1 and T_2 are $o_p(1)$. For T_1, it is enough to show that $L - L^0 = o_p(1)$ as we have already established that σ_n^{cf} is bounded away from 0. Towards that end, define $\mathcal{G}_n = \sigma(A_{1:n}, X_{1:n}, M_{1:n})$. We have:

$$\mathbb{E}[(L-L^0)^2 \mid \mathcal{G}_n]$$
$$= \frac{n}{4}\mathbb{E}\left[\left(\sum_{i\in\mathcal{D}_2, A_i=1}(\hat{\tau}_{1,i}^{(2)} - \hat{\tau}_{1,i}^{(2),0})\epsilon_i + \sum_{i\in\mathcal{D}_1, A_i=1}(\hat{\tau}_{1,i}^{(1)} - \hat{\tau}_{1,i}^{(1),0})\epsilon_i\right)^2 \mid \mathcal{G}_n\right]$$
$$= \frac{n\sigma_1^2}{4}\sum_{j=1}^{2}\|\hat{\tau}_1^{(j)} - \hat{\tau}_1^{(j),0}\|_2^2 = o_p(1)\,.$$

where the last conclusion again follows from Lemma B.1. Therefore $|L - L^0| = o_p(1)$ and consequently $T_1 = o_p(1)$. For T_2, we have already shown $|(\sigma_n^{\mathrm{cf}})^2 - (\sigma_n^{\mathrm{cf},0})^2| = o_p(1)$. This, along with the fact that σ_n^{cf} and $\sigma_n^{\mathrm{cf}.0}$ are bounded away from 0, implies $|\sigma_n^{\mathrm{cf}} - \sigma_n^{\mathrm{cf},0}| = o_p(1)$. Now, this, along with the fact that σ_n^{cf} is bounded away from 0 immplies $(\sigma_n^{\mathrm{cf}.0}/\sigma_n^{\mathrm{cf}} - 1) = o_p(1)$. As we have already established $L^0/\sigma_n^{\mathrm{cf}.0}$ is asymptotically normal (hence $O_p(1)$), we therefore have $T_2 = o_p(1)$. This completes the proof that $T \xrightarrow{d} \mathcal{N}(0,1)$ and hence concludes Theorem 5.9. □

B.3 Proof of Corollary 5.11

Proof. First, we claim, as a corollary of Theorem 5.9 that our proof essentially yields the following conditional version of the central limit theorem:

Lemma B.2 (Conditional asymptotic normality)**.** *Let $\mathcal{X}_n = \sigma(X_{1:n})$ be the σ-field generated by $\{X_1, \ldots, X_n\}$. Under the same assumptions as Theorem 5.9, we have, for every $t \in \mathbb{R}$,*

$$\mathbb{P}\left(\frac{\sqrt{n}(\hat{\theta}_{10}^{\mathrm{cf}} - \theta_{10})}{\sigma_n^{\mathrm{cf}}} \leq t \,\middle|\, \mathcal{X}_n\right) \longrightarrow \Phi(t), \qquad \textit{in probability.}$$

Here $\Phi(t)$ denotes the CDF of the standard normal distribution.

The proof of this Lemma is deferred to Section C. Now, we present the proof of Corollary 5.11 using this lemma. Towards that end, let us define $A_n := \sqrt{n}(\hat{\theta}_{10}^{\mathrm{cf}} - \theta_{10})$ and $B_n := \sqrt{n}(\theta_{10} - \theta_{10}^{\mathrm{pop}})$. Recall that, by definition, $\theta_{10} = \bar{X}^\top\Lambda$, $\theta_{10}^{\mathrm{pop}} = \mu_X^\top\Lambda$ with $\Lambda = \beta_1 + \mathbf{B}_0^\top\gamma_1$. Therefore,

$$B_n = \sqrt{n}(\bar{X} - \mu_X)^\top\Lambda = \tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}(X_i - \mu_X)^\top\Lambda,$$

and consequently B_n is measurable with respect to $\mathcal{X}_n$. By Lemma B.2, for every fixed $u \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left\{iu(A_n/\sigma_n^{\text{cf}})\right\} \mid \mathcal{X}_n\right] \xrightarrow{p} e^{-u^2/2}.$$

Since the absolute difference between the two characteristic functions is bounded by 2, convergence in probability implies convergence in L_1. Consequently,

$$\mathbb{E}\left|\mathbb{E}\left[\exp\left\{iu(A_n/\sigma_n^{\text{cf}})\right\} \mid \mathcal{X}_n\right] - e^{-u^2/2}\right| \longrightarrow 0. \tag{B.10}$$

We next establish the asymptotic joint behavior of $(A_n/\sigma_n^{\text{cf}}, B_n)$. For any fixed $u, v \in \mathbb{R}$, since B_n is measurable with respect to $\mathcal{X}_n$,

$$\mathbb{E}\left[\exp\left\{iu(A_n/\sigma_n^{\text{cf}}) + ivB_n\right\}\right] = \mathbb{E}\left[e^{ivB_n}\mathbb{E}\left[\exp\left\{iu(A_n/\sigma_n^{\text{cf}})\right\} \mid \mathcal{X}_n\right]\right].$$

As a consequence,

$$\begin{aligned}
&\left|\mathbb{E}\left[\exp\left\{iu(A_n/\sigma_n^{\text{cf}}) + ivB_n\right\}\right] - e^{-u^2/2}\mathbb{E}[e^{ivB_n}]\right| \\
&\leq \mathbb{E}\left|\mathbb{E}\left[\exp\left\{iu(A_n/\sigma_n^{\text{cf}})\right\} \mid \mathcal{X}_n\right] - e^{-u^2/2}\right| \longrightarrow 0,
\end{aligned} \tag{B.11}$$

where the last conclusion follows from Equation (B.10). For the rest of the analysis define $\omega_n = (\sigma_n^{\text{cf}})^2/(\sigma_X^2 + (\sigma_n^{\text{cf}})^2)$. By our assumption, there exists a bounded deterministic sequence $\bar{\omega}_n$ such that $\omega_n - \bar{\omega}_n \to 0$ in probability. As $\sigma_X^2 \leq C$, given any sequence, there exists a further subsequence such that on that subsequence (we use the same notation n for the sub-sequence for notational simplicity)

$$\sigma_X^2 \longrightarrow \zeta^2, \qquad \bar{\omega}_n \longrightarrow \omega,$$

for some $\zeta^2 \in [0, C]$ and $\omega \in [0, 1]$ and consequently $\omega_n \xrightarrow{p} \omega$. By the sub-Gaussian design condition, the triangular array $\left\{(X_i - \mu_X)^\top \Lambda : 1 \leq i \leq n\right\}$ satisfies the Lindeberg condition, and consequently, along this sub-sequence, by Lindeberg-Feller CLT, we have:

$$B_n \xrightarrow{d} Z_2 \sim \mathcal{N}(0, \zeta^2)\,.$$

Combining this conclusion with Equation (B.11), we have, for any fixed $u, v \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left\{iu(A_n/\sigma_n^{\text{cf}}) + ivB_n\right\}\right] \longrightarrow e^{-u^2/2}e^{-v^2\zeta^2/2}.$$

This immediately implies,

$$\left(\frac{A_n}{\sigma_n^{\text{cf}}}, B_n\right) \xrightarrow{d} (Z_1, Z_2), \tag{B.12}$$

where $Z_1 \sim \mathcal{N}(0,1), Z_2 \sim \mathcal{N}(0, \zeta^2)$ and $Z_1 \perp\!\!\!\perp Z_2$. Now define $D_n^2 := (\sigma_n^{\text{cf}})^2 + \sigma_X^2$. Then, from our definition of ω_n, we have $\omega_n = (\sigma_n^{\text{cf}})^2/D_n^2$, and consequently

$$(A_n + B_n)/D_n = \sqrt{\omega_n}\ (A_n/\sigma_n^{\text{cf}}) + (B_n/D_n). \tag{B.13}$$

Furthermore, we also have $D_n = \sigma_X/\sqrt{1-\omega_n}$. First consider the case $\zeta > 0$. Then $\omega_n \to \omega$ in probability and $\sigma_X \to \zeta > 0$. Therefore, by Slutsky's theorem, along with Equation (B.12) we have:

$$\frac{A_n + B_n}{D_n} \xrightarrow{d} \sqrt{\omega}\, Z_1 + \frac{\sqrt{1-\omega}}{\zeta} Z_2 \sim \mathcal{N}(0,1)\,.$$

Suppose next that $\zeta^2 = 0$. In this case, $\mathbb{E}[B_n^2] = \sigma_X^2 \to 0$, and consequently $B_n = o_p(1)$. Furthermore, Theorem 5.9 establishes that, for some constant $c_0 > 0$, $\mathbb{P}\left((\sigma_n^{\text{cf}})^2 \geq c_0\right) \to 1$ and consequently D_n is also bounded away from 0 with probability going to 1. Therefore, $\omega_n \to 1$ in probability and $B_n/D_n = o_p(1)$. Then it follows from Equation (B.13) that

$$\tfrac{A_n+B_n}{D_n} = \tfrac{A_n}{\sigma_n^{\text{cf}}} + o_p(1) \xrightarrow{d} \mathcal{N}(0,1),$$

where the last conclusion follows from Theorem 5.9. Hence, we have proved that given any sequence, there exists a further subsequence along which

$$\frac{A_n+B_n}{\sqrt{(\sigma_n^{\mathrm{cf}})^2+\sigma_X^2}} \xrightarrow{d} \mathcal{N}(0,1).$$

As the limiting distribution is the same for all converging subsequences, we conclude:

$$\frac{A_n+B_n}{\sqrt{(\sigma_n^{\mathrm{cf}})^2+\sigma_X^2}} \xrightarrow{d} \mathcal{N}(0,1).$$

This completes the proof recalling $A_n + B_n = \sqrt{n}(\hat{\theta}_{10}^{\mathrm{cf}} - \theta_{10}^{\mathrm{pop}})$. □

B.4 Proof of Corollary 5.12

Proof. For the first part, the proof is very closely related to the proof of Theorem 5.9, so we will highlight the key differences here. The first part of the proof of Theorem 5.9, i.e., $\sqrt{n}\Delta_n^{\mathrm{cf}} = o_p(1)$, remains identical, as the lasso rates and the conclusion of Lemma 5.8 do not depend directly on the conditional heteroskedasticity as long as errors are uniformly sub-Gaussian. The second part, i.e., lower bounding $(\sigma_n^{\mathrm{cf,het}})^2$ also remains almost identical upon observing that $\mathsf{var}(\epsilon \mid X, M, A = 1) \geq c_-$. The only difference comes in the third part. Recall the oracle balancing weights $\tau_1^{(j),0}$ as introduced in the proof of Theorem 5.9, which is obtained by replacing a with $a^0 = (\bar{X}^\top \ (\mathbf{B}_0\bar{X})^\top)^\top$ in Step 3 of Algorithm 1. Also recall the definition of T and T^0 introduced in the proof of the asymptotic normality part of Theorem 5.9 with σ_n^{cf} replaced by $\sigma_n^{\mathrm{cf,het}}$ as defined in the statement of Corollary 5.12. As before, we first establish that T^0 converges to $\mathcal{N}(0,1)$, where recall that $T^0 = \sum_{i=1}^n \zeta_{n,i,\mathrm{het}}$ with:

$$\begin{aligned}
\zeta_{n,i,\mathrm{het}} = \frac{\sqrt{n}}{2\sigma_n^{\mathrm{cf,het},0}} \Big\{ \epsilon_i \left(\hat{\tau}_{1,i}^{(1),0} \mathbb{1}_{i\in\mathcal{D}_1, A_i=1} + \hat{\tau}_{1,i}^{(2),0} \mathbb{1}_{i\in\mathcal{D}_2, A_i=1} \right) \\
+ U_i^\top \gamma_1 \left(\hat{\tau}_{2,i}^{(1)} \mathbb{1}_{i\in\mathcal{D}_1, A_i=0} + \hat{\tau}_{2,i}^{(2)} \mathbb{1}_{i\in\mathcal{D}_2, A_i=0} \right) \Big\}
\end{aligned}$$

$$(\sigma_n^{\mathrm{cf,het},0})^2 = \frac{n}{4} \sum_{j=1}^{2} \left(\sum_{i\in\mathcal{D}_j, A_i=1} \{(\hat{\tau}_{1,i}^{(j),0})^2 \nu_1(X_i, M_i)\} + \gamma_1^\top \Sigma_{U,0} \gamma_1 \|\hat{\tau}_2^{(j)}\|_2^2 \right).$$

Define $\mathcal{H}_n = \sigma(A_{1:n}, X_{1:n}, \{M_i : A_i = 1\})$. It is immediate that conditional on $\mathcal{H}_n$, $\{\zeta_{n,i,\mathrm{het}}\}$ are independent across i. Same calculation as in the proof of Theorem 5.9 yields $\mathbb{E}[\zeta_{n,i,\mathrm{het}} \mid \mathcal{H}_n] = 0$ and $\mathsf{var}(\sum_i \zeta_{n,i,\mathrm{het}} \mid \mathcal{H}_n) = 1$. We next show that the difference between $(\sigma_n^{\mathrm{cf,het}})^2$ and $(\sigma_n^{\mathrm{cf,het},0})^2$ is asymptotically negligible (similar to Equation (B.8)):

$$\begin{aligned}
&\left|(\sigma_n^{\mathrm{cf,het}})^2 - (\sigma_n^{\mathrm{cf,het},0})^2\right| \\
&= \frac{n}{4} \sum_{j=1}^{2} \left[\sum_{i\in\mathcal{D}_j, A_i=1} \left\{ \left((\hat{\tau}_{1,i}^{(j)})^2 - (\hat{\tau}_{1,i}^{(j),0})^2 \right) \nu_1(X_i, M_i) \right\} \right] \\
&\leq \frac{nc_+}{4} \sum_{j=1}^{2} \sum_{i\in\mathcal{D}_j, A_i=1} \left| (\hat{\tau}_{1,i}^{(j)})^2 - (\hat{\tau}_{1,i}^{(j),0})^2 \right| \\
&\leq \frac{nc_+}{4} \sum_{j=1}^{2} \sum_{i\in\mathcal{D}_j, A_i=1} \left| \hat{\tau}_{1,i}^{(j)} + \hat{\tau}_{1,i}^{(j),0} \right| \left| \hat{\tau}_{1,i}^{(j)} - \hat{\tau}_{1,i}^{(j),0} \right| \\
&\leq \frac{nc_+}{4} \sum_{j=1}^{2} \sqrt{\sum_{i\in\mathcal{D}_j, A_i=1} \left| \hat{\tau}_{1,i}^{(j)} + \hat{\tau}_{1,i}^{(j),0} \right|^2} \sqrt{\sum_{i\in\mathcal{D}_j, A_i=1} \left| \hat{\tau}_{1,i}^{(j)} - \hat{\tau}_{1,i}^{(j),0} \right|^2} \\
&\leq \frac{nc_+}{2\sqrt{2}} \sum_{j=1}^{2} (\|\hat{\tau}_1^{(j)}\|_2 + \|\hat{\tau}_1^{(j),0}\|_2) \|\hat{\tau}_1^{(j)} - \hat{\tau}_1^{(j),0}\|_2
\end{aligned}$$

$$= \frac{c_+}{2\sqrt{2}} \sum_{j=1}^{2} (\sqrt{n}\|\hat{\tau}_1^{(j)}\|_2 + \sqrt{n}\|\hat{\tau}_1^{(j),0}\|_2)\sqrt{n}\|\hat{\tau}_1^{(j)} - \hat{\tau}_1^{(j),0}\|_2 = o_p(1)\,. \tag{B.14}$$

The last conclusion follows from Lemma 5.8 and Lemma B.1. Therefore, $\sigma_n^{\text{cf,het},0}$ is also bounded away from 0 as probability tend to 1. As a consequence, $\sum_i \mathbb{E}[|\zeta_{n,i,\text{het}}|^3 \mid \mathcal{H}_n] \to 0$ by the exact same calculation as in the proof of Theorem 5.9; the only difference is $(\sigma_n^{\text{cf},0})^3$ is replaced by $(\sigma_n^{\text{cf,het},0})^3$ but as we have argued, $\sigma_n^{\text{cf,het},0}$ is bounded away from 0 with probability going to 1 as $\nu_1(x,m) \geq c_- > 0$. Therefore, by conditional CLT, $T^0 \xrightarrow{d} \mathcal{N}(0,1)$. Next, we need to show $T - T^0 = o_p(1)$. Following same expansion as in Equation (B.9) we need to show $L - L^0 = o_p(1)$ and $|\sigma_n^{\text{cf,het}} - \sigma_n^{\text{cf,het},0}| = o_p(1)$. The latter is already established. For the former, we have, with $\mathcal{G}_n = \sigma(A_{1:n}, X_{1:n}, M_{1:n})$:

$$\begin{aligned}
&\mathbb{E}[(L - L^0)^2 \mid \mathcal{G}_n] \\
&= \frac{n}{4}\mathbb{E}\left[\left(\sum_{i\in\mathcal{D}_2, A_i=1} (\hat{\tau}_{1,i}^{(2)} - \hat{\tau}_{1,i}^{(2),0})\epsilon_i + \sum_{i\in\mathcal{D}_1, A_i=1} (\hat{\tau}_{1,i}^{(1)} - \hat{\tau}_{1,i}^{(1),0})\epsilon_i\right)^2 \mid \mathcal{G}_n\right] \\
&\leq \frac{nc_+}{4} \sum_{j=1}^{2} \|\hat{\tau}_1^{(j)} - \hat{\tau}_1^{(j),0}\|_2^2 = o_p(1)\,.
\end{aligned}$$

where the last conclusion follows from Lemma B.1. Consequently, by conditional Markov's inequality, $L - L^0 = o_p(1)$. This completes the proof of asymptotic normality of $\hat{\theta}_{10}^{\text{cf}}$ around θ_{10}.

Next, we will establish the asymptotic normality of $\hat{\theta}_{10}^{\text{cf}}$ around θ_{10}^{pop}. The proof here is closely related to that of Corollary 5.11. We use the same notation and definition for $V_n^{\text{cf}}, V_n^{\text{cf},0}, T_n^{\text{cf}}, A_n, B_n$ and $L_n^{\text{cf},0}$ as used/introduced in the proof of Corollary 5.11. Our first goal is to prove an analogous result of Lemma B.2 with σ_n^{cf} replaced by $\sigma_n^{\text{cf,het}}$. The proof is almost exactly the same, the only thing we need to use is $(\sigma_n^{\text{cf,het},0}/\sigma_n^{\text{cf,het}}) - 1 = o_p(1)$ along with the fact that $\sigma_n^{\text{cf,het}}$ and $\sigma_n^{\text{cf,het},0}$ are bounded away from 0 with probability 1 as we have already argued in the previous part. Therefore, we also have:

$$\mathbb{P}\left(\frac{\sqrt{n}(\hat{\theta}_{10}^{\text{cf}} - \theta_{10})}{\sigma_n^{\text{cf,het}}} \leq t \mid \mathcal{X}_n\right) \longrightarrow \Phi(t), \quad \text{in probability with respect to } \mathcal{X}_n\,.$$

Then the exact same calculation as in Corollary 5.11 yields

$$\left|\mathbb{E}\left[\exp\left\{iu(A_n/\sigma_n^{\text{cf,het}}) + ivB_n\right\}\right] - e^{-u^2/2}\mathbb{E}[e^{ivB_n}]\right| \longrightarrow 0, \tag{B.15}$$

The last part is also a similar sub-sequence argument. Define $\omega_n^{\text{het}} = (\sigma_n^{\text{cf,het}})^2/((\sigma_n^{\text{cf,het}})^2 + \sigma_X^2)$, where, as in Corollary 5.11 $\sigma_X^2 = \Lambda^\top \mathsf{var}(X)\Lambda$ with $\Lambda = (\beta_1 + \mathbf{B}_0^\top\gamma_1)$. As per our assumption, there exists a deterministic sequence $\bar{\omega}_n^{\text{het}} \in [0,1]$ such that $\omega_n^{\text{het}} - \bar{\omega}_n^{\text{het}} = o_p(1)$. As $\sigma_X^2 \leq C$ (bounded), given any sequence, there exists a further subsequence such that along that subsequence $\bar{\omega}_n^{\text{het}} \to \omega \in [0,1]$ and $\sigma_X^2 \to \zeta^2 \geq 0$. Assume $\zeta > 0$. We use the same index n for this subsequence for notational simplicity. Now, by sub-Gaussianity on the design, $\{(X_i - \mu_X)^\top\Lambda : 1 \leq i \leq n\}$ satisfies the Lindeberg condition, and consequently we have along this subsequence:

$$B_n = \sqrt{n}(\bar{X} - \mu_X)^\top\Lambda \xrightarrow{d} \mathcal{N}(0, \zeta^2)\,.$$

This, along with Equation (B.15) implies:

$$\left(\tfrac{A_n}{\sigma_n^{\text{cf,het}}},\ B_n\right) \xrightarrow{d} (Z_1, Z_2), \quad Z_1 \sim \mathcal{N}(0,1), Z_2 \sim \mathcal{N}(0,\zeta^2),\ Z_1 \perp\!\!\!\perp Z_2\,.$$

Define $D_n^2 = (\sigma_n^{\text{cf,het}})^2 + \sigma_X^2$. Then $D_n = \sigma_X/\sqrt{1 - \omega_n^{\text{het}}}$. Then:

$$(A_n + B_n)/D_n = \sqrt{\omega_n^{\text{het}}}(A_n/\sigma_n^{\text{cf,het}}) + (B_n/D_n)$$

Therefore, along the subsequence, by Slutsky's theorem, we have:

$$\tfrac{A_n + B_n}{D_n} \xrightarrow{d} \sqrt{\omega}Z_1 + \tfrac{\sqrt{1-\omega}}{\zeta}Z_2 \sim \mathcal{N}(0,1)\,.$$

The case $\zeta = 0$ is handled exactly in the same way as in the proof of Corollary 5.11 and hence is skipped. Therefore, we have concluded that given any sequence, there exists a further sub-sequence such that along that sub-sequence $(A_n + B_n)/D_n \xrightarrow{d} \mathcal{N}(0,1)$. As the limiting distribution is the same along all such subsequences, we conclude that the entire sequence $(A_n + B_n)/D_n$ converges to $\mathcal{N}(0,1)$. This completes the proof. □

B.5 Proof of Theorem 5.13

Proof. Recall from the definition of NDE, we have $\hat{\theta}^{\text{cf}}_{\text{NDE}} = \hat{\theta}^{\text{cf}}_{10} - \hat{\theta}^{\text{cf}}_{00}$, where

$$\hat{\theta}^{\text{cf}}_{00} = \frac{\hat{\theta}^{(1)}_{00} + \hat{\theta}^{(2)}_{00}}{2} = \frac{1}{2}\sum_{j=1}^{2}\left(\bar{X}^\top \hat{\rho}^{(j)}_c + (\hat{\tau}^{(j)}_2)^\top (\mathbf{y}_{c,j} - \mathbf{X}_{c,j}\hat{\rho}^{(j)}_c)\right)$$

From the reduced form regression we have:

$$\begin{aligned} Y &= A(X^\top \beta_1 + M^\top \gamma_1 + \epsilon) + (1-A)(X^\top \beta_0 + M^\top \gamma_0 + \epsilon') \\ &= A(X^\top \beta_1 + X^\top \mathbf{B}_1^\top \gamma_1 + \gamma_1^\top U' + \epsilon) + (1-A)(X^\top \beta_0 + X^\top \mathbf{B}_0^\top \gamma_0 + \gamma_0^\top U + \epsilon') \end{aligned}$$

Therefore, on the control arm, $Y = X^\top \rho_c + \xi_0$, where $\rho_c = \beta_0 + \mathbf{B}_0^\top \gamma_0$ and $\xi_0 = U^\top \gamma_0 + \epsilon'$. Using this we can decompose the estimation error of $\hat{\theta}^{\text{cf}}_{00}$ as:

$$\begin{aligned} \hat{\theta}^{\text{cf}}_{00} - \theta_{00} &= \frac{1}{2}\sum_{j=1}^{2}\left(\bar{X}^\top(\hat{\rho}^{(j)}_c - \rho_c) + (\hat{\tau}^{(j)}_2)^\top(\mathbf{y}_{c,j} - \mathbf{X}_{c,j}\hat{\rho}^{(j)}_c)\right) \\ &= \frac{1}{2}\sum_{j=1}^{2}(\bar{X} - \mathbf{X}_{c,j}^\top \hat{\tau}^{(j)}_2)^\top(\hat{\rho}^{(j)}_c - \rho_c) + \frac{1}{2}\sum_{j=1}^{2}(\hat{\tau}^{(j)}_2)^\top \boldsymbol{\xi}_0 := \Delta^{\text{cf}}_{n,00} + V^{\text{cf}}_{n,00} \end{aligned}$$

This, along with the decomposition of $\hat{\theta}^{\text{cf}}_{10}$ (Equation (B.2)) yields:

$$\frac{\sqrt{n}(\hat{\theta}^{\text{cf}}_{\text{NDE}} - \theta_{\text{NDE}})}{\sigma^{\text{cf,het}}_{n,\text{NDE}}} = \frac{\sqrt{n}(\Delta^{\text{cf}}_n - \Delta^{\text{cf}}_{n,00} + V^{\text{cf}}_n - V^{\text{cf}}_{n,00})}{\sigma^{\text{cf,het}}_{n,\text{NDE}}}$$

The proof proceeds similarly to that of Theorem 5.9. We first show $\sqrt{n}(\Delta^{\text{cf}}_n - \Delta^{\text{cf}}_{n,00}) = o_p(1)$, then we show $\sigma^{\text{cf,het}}_{n,\text{NDE}}$ is bounded away from 0 with probability going to 1, and finally we establish that $\sqrt{n}(V^{\text{cf}}_n - V^{\text{cf}}_{n,00})/\sigma^{\text{cf,het}}_{n,\text{NDE}}$ converges to $\mathcal{N}(0,1)$. We have already established in the proof of Theorem 5.9 that $\sqrt{n}\Delta^{\text{cf}}_n = o_p(1)$. Here, we will first show that $\sqrt{n}\Delta^{\text{cf}}_{n,00} = o_p(1)$. Given that $\|\rho_c\|_0 = \|\beta_0 + \mathbf{B}_0^\top \gamma_0\|_0 \le k_\beta + sk_\gamma \le (1+s)k$ and the error ξ_0 satisfies $\mathbb{E}[\xi_0 \mid X, A=0] = 0$ and sub-Gaussian (uniformly over X, Assumption 5.3) and the design matrix satisfies RE condition with probability going to 1, the lasso estimator $\hat{\rho}^{(j)}_c$ satisfies by same argument in Theorem 5.9

$$\|\hat{\rho}^{(j)}_c - \rho_c\|_1 = O_p\left(sk\sqrt{\tfrac{\log p}{n}}\right), \ \|\hat{\rho}^{(j)}_c - \rho_c\|_2 = O_p\left(\sqrt{\tfrac{sk\log p}{n}}\right).$$

As a consequence:

$$\begin{aligned} \sqrt{n}|\Delta^{\text{cf}}_{n,00}| &\le \frac{1}{2}\sum_{j=1}^{2}\|\bar{X} - \mathbf{X}_{c,j}^\top \hat{\tau}^{(j)}_2\|_\infty \|\hat{\rho}^{(j)}_c - \rho_c\|_1 \\ &= O_p\left(\sqrt{n} \times \sqrt{\frac{\log p}{n}} \times sk\sqrt{\frac{\log p}{n}}\right) = O_p\left(\frac{sk\log p}{\sqrt{n}}\right) = o_p(1) \end{aligned}$$

where the last conclusion follows from Assumption 5.4. Here we have also used the fact that $\|\bar{X} - \mathbf{X}_c^\top \hat{\tau}^{(j)}_2\|_\infty \le K_2\sqrt{\log p/n}$ which is by definition of the optimization problem (3.1). This completes the proof of $\sqrt{n}(\Delta^{\text{cf}}_n - \Delta^{\text{cf}}_{n,00}) = o_p(1)$.

We next argue that $\sigma^{\text{cf,het}}_{n,\text{NDE}}$ is bounded away from 0 with probability going to 1. The proof is very similar to that of the second step of Theorem 5.9; from the definition, we have:

$$(\sigma^{\text{cf,het}}_{n,\text{NDE}})^2 = \frac{n}{4}\sum_{j=1}^{2}\left[\sum_{i\in\mathcal{D}_j, A_i=1}(\hat{\tau}^{(j)}_{1,i})^2\nu_1(X_i,M_i) + \sum_{i\in\mathcal{D}_j, A_i=0}(\hat{\tau}^{(j)}_{2,i})^2\omega_0(X_i)\right]$$
$$\geq \frac{c_-}{4}\, n\|\hat{\tau}^{(1)}_1\|_2^2$$

Here we have used the fact that $\nu_1(x,m)\geq c_-$ (Assumption 5.3). Now, we have already established in the proof of Theorem 5.9 $n\|\hat{\tau}^{(1)}_1\|_2^2$ is bounded away from 0 with probability going to 1. This immediately implies $\sigma^{\text{cf,het}}_{n,\text{NDE}}$ is bounded away from 0 with probability going to 1.

The final part is to establish the asymptotic normality. Similar to the proof of Theorem 5.9. From the definition of V^{cf}_n and $V^{\text{cf}}_{n,00}$, we have:

$$\frac{\sqrt{n}(V^{\text{cf}}_n - V^{\text{cf}}_{n,00})}{\sigma^{\text{cf,het}}_{n,\text{NDE}}} = \frac{\sqrt{n}}{2\sigma^{\text{cf,het}}_{n,\text{NDE}}}\sum_{j=1}^{2}\left\{(\hat{\tau}^{(j)}_1)^\top\boldsymbol{\epsilon}^{(j)} + (\hat{\tau}^{(j)}_2)^\top(\mathbf{U}^{(j)}(\gamma_1-\gamma_0) - \boldsymbol{\epsilon}'^{(j)})\right\}$$

Recall that definition of $\hat{\tau}^{(j),0}_1$ as introduced in the proof of Theorem 5.9. We have established in Lemma B.1 that $\sqrt{n}\|\hat{\tau}^{(j)}_1 - \hat{\tau}^{(j),0}_1\|_2 = o_p(1)$. Define T^0 as:

$$T^0 = \frac{\sqrt{n}}{2\sigma^{\text{cf,het},0}_{n,\text{NDE}}}\sum_{j=1}^{2}\left[\sum_{\substack{i\in\mathcal{D}_j\\A_i=1}}\hat{\tau}^{(j),0}_{1,i}\epsilon_i + \sum_{\substack{i\in\mathcal{D}_j\\A_i=0}}\hat{\tau}^{(j)}_{2,i}(U_i^\top(\gamma_1-\gamma_0)-\epsilon'_i)\right] := \sum_{i=1}^{n}\zeta^{\text{het}}_{n,i}$$

with

$$\zeta^{\text{het}}_{n,i} = \frac{\sqrt{n}}{2\sigma^{\text{cf,het},0}_{n,\text{NDE}}}\left\{\epsilon_i\left(\hat{\tau}^{(1),0}_{1,i}\mathbb{1}_{i\in\mathcal{D}_1} + \hat{\tau}^{(2),0}_{1,i}\mathbb{1}_{i\in\mathcal{D}_2}\right)\mathbb{1}_{A_i=1}\right.$$
$$\left.+(U_i^\top(\gamma_1-\gamma_0)-\epsilon'_i)\left(\hat{\tau}^{(1)}_{2,i}\mathbb{1}_{i\in\mathcal{D}_1} + \hat{\tau}^{(2)}_{2,i}\mathbb{1}_{i\in\mathcal{D}_2}\right)\mathbb{1}_{A_i=0}\right\}$$
$$(\sigma^{\text{cf,het},0}_{n,\text{NDE}})^2 = \frac{n}{4}\sum_{j=1}^{2}\left[\sum_{i\in\mathcal{D}_j, A_i=1}(\hat{\tau}^{(j),0}_{1,i})^2\nu_1(X_i,M_i) + \sum_{i\in\mathcal{D}_j, A_i=0}(\hat{\tau}^{(j)}_{2,i})^2\omega_0(X_i)\right].$$

Define $\mathcal{H}_n = \sigma(A_{1:n}, X_{1:n}, \{M_i : A_i = 1\})$. Then it is immediate that conditional on $\mathcal{H}_n$, ζ^{het}_i are independent across i with $\mathbb{E}[\zeta^{\text{het}}_i \mid \mathcal{H}_n] = 0$ and

$$\begin{aligned}
&\mathsf{var}(T^0 \mid \mathcal{H}_n)\\
&= \sum_i \mathsf{var}(\zeta^{\text{het}}_i \mid \mathcal{H}_n)\\
&= \frac{n}{4(\sigma^{\text{cf,het},0}_{n,\text{NDE}})^2}\sum_{j=1}^{2}\left[\sum_{i\in\mathcal{D}_j, A_i=1}(\hat{\tau}^{(j),0}_{1,i})^2\nu_1(X_i,M_i) + \sum_{i\in\mathcal{D}_j, A_i=0}(\hat{\tau}^{(j)}_{2,i})^2\omega_0(X_i)\right] = 1\,.
\end{aligned}$$

The proof that $\sum_i \mathbb{E}[|\zeta^{\text{het}}_{n,i}|^3 \mid \mathcal{H}_n] = o_p(1)$ is same as that of $\sum_i \mathbb{E}[|\zeta_{n,i}|^3 \mid \mathcal{H}_n] = o_p(1)$ in the proof of Theorem 5.9 upon using

$$\nu_1(x,m)\vee\omega_0(x)\leq C, \qquad \text{(Assumption 5.3)}$$
$$(\gamma_1-\gamma_0)^\top\Sigma_{U,0}(\gamma_1-\gamma_0)\leq\lambda_{U,0,+}\|\gamma_1-\gamma_0\|_2^2\leq 2C_\gamma^2\lambda_{U,0,+}\,. \qquad \text{(Assumption 5.3 \& 5.4)}$$

Furthermore, as we have already argued $\sigma^{\text{cf,het}}_{n,\text{NDE}}$ is bounded away from 0 with probability going to 1, the same calculation as of Equation (B.14) yields that $\sigma^{\text{cf,het},0}_{n,\text{NDE}}$ is also bounded away from 0 with probability

going to 1 (observe that $(\sigma^{\text{cf,het}}_{n,\text{NDE}})^2 - (\sigma^{\text{cf,het},0}_{n,\text{NDE}})^2$ is same as $(\sigma^{\text{cf,het}}_{n})^2 - (\sigma^{\text{cf,het}}_{n})^2$). Therefore, $T^0 \xrightarrow{d} \mathcal{N}(0,1)$. Finally defining $T = \sqrt{n}(V^{\text{cf}}_n - V^{\text{cf}}_{n,00})/\sigma^{\text{cf,het}}_{n,\text{NDE}}$, all we need to argue is $T - T^0$. This is also very similar to that of Theorem 5.9 and Corollary 5.12. To keep the notation consistent with the previous proofs, let us define $L^0 = T^0\sigma^{\text{cf,het},0}_{n,\text{NDE}}$ and $L = T\sigma^{\text{cf,het}}_{n,\text{NDE}}$. Similar decomposition as in Equation (B.9) yields:

$$T - T^0 = \frac{L}{\sigma^{\text{cf,het}}_{n,\text{NDE}}} - \frac{L^0}{\sigma^{\text{cf,het},0}_{n,\text{NDE}}} = \frac{L - L^0}{\sigma^{\text{cf,het}}_{n,\text{NDE}}} + \frac{L^0}{\sigma^{\text{cf,het},0}_{n,\text{NDE}}}\left(\frac{\sigma^{\text{cf,het},0}_{n}}{\sigma^{\text{cf,het}}_{n,\text{NDE}}} - 1\right) := T_1 + T_2\,.$$

Therefore, to establish asymptotic normality of T all we need to show is $L - L^0 = o_p(1)$ and $|\sigma^{\text{cf,het},0}_{n,\text{NDE}} - \sigma^{\text{cf,het}}_{n,\text{NDE}}| = o_p(1)$ (as this along with the fact that $\sigma^{\text{cf,het}}_{n,\text{NDE}}$ is bounded away from 0 with probability going to 1 implies $|(\sigma^{\text{cf,het},0}_{n,\text{NDE}}/\sigma^{\text{cf,het}}_{n,\text{NDE}}) - 1| = o_p(1)$. The second part can be proved immediately by the same calculation as in Equation (B.14). For the first part, define $\mathcal{G}_n = \sigma(A_{1:n}, X_{1:n}, M_{1:n})$. Then we have:

$$L - L^0 = \tfrac{\sqrt{n}}{2}\textstyle\sum_{j=1}^{2}\sum_{\substack{i\in\mathcal{D}_j\\ A_i=1}}(\hat{\tau}^{(j)}_{1,i} - \hat{\tau}^{(j),0}_{1,i})\epsilon_i\,.$$

As a consequence,

$$\mathbb{E}\left[(L - L^0)^2 \mid \mathcal{G}_n\right] \le \tfrac{nc_+}{4}\textstyle\sum_{j=1}^{2}\|\hat{\tau}^{(j)}_1 - \hat{\tau}^{(j),0}_1\|_2^2 = o_p(1)$$

where the conclusion follows from Lemma B.1. Therefore, an application of the Markov inequality yields that $L - L^0 = o_p(1)$. This completes the proof.

The proof for asymptotic normality of $\hat{\theta}^{\text{cf}}_{\text{NDE}}$ around $\theta^{\text{pop}}_{\text{NDE}}$ follows the same argument as that of Corollary 5.11; we only highlight the required changes. Define

$$A_n = \sqrt{n}\left(\hat{\theta}^{\text{cf}}_{\text{NDE}} - \theta_{\text{NDE}}\right), \qquad B_n = \sqrt{n}\left(\theta_{\text{NDE}} - \theta^{\text{pop}}_{\text{NDE}}\right).$$

Since $\theta_{\text{NDE}} = \bar{X}^\top\Lambda_{\text{NDE}}$ and $\theta^{\text{pop}}_{\text{NDE}} = \mu_X^\top\Lambda_{\text{NDE}}$, we have

$$B_n = \sqrt{n}(\bar{X} - \mu_X)^\top\Lambda_{\text{NDE}} = \tfrac{1}{\sqrt{n}}\textstyle\sum_{i=1}^{n}(X_i - \mu_X)^\top\Lambda_{\text{NDE}}.$$

Consequently, B_n is measurable with respect to $\sigma(X_{1:n}) = \mathcal{X}_n$. Similar argument as in Lemma B.2 yields, for every fixed $u \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left\{iu\tfrac{A_n}{\sigma^{\text{cf,het}}_{n,\text{NDE}}}\right\}\,\middle|\,\mathcal{X}_n\right] \xrightarrow{p} e^{-u^2/2}.$$

Since B_n is $\mathcal{X}_n$-measurable, the same conditional characteristic-function calculation as in the proof of Corollary 5.11 gives, for any $u, v \in \mathbb{R}$,

$$\left|\mathbb{E}\left[\exp\left\{iu\frac{A_n}{\sigma^{\text{cf,het}}_{n,\text{NDE}}} + ivB_n\right\}\right] - e^{-u^2/2}\mathbb{E}[e^{ivB_n}]\right| \longrightarrow 0. \tag{B.16}$$

Since $\sigma^2_{X,\text{NDE}} \le C$ and $\bar{\omega}^{\text{het}}_{n,\text{NDE}} \in [0,1]$, given any subsequence, there exists a further subsequence, which we continue to index by n, such that $\sigma^2_{X,\text{NDE}} \to \zeta^2$ and $\bar{\omega}^{\text{het}}_{n,\text{NDE}} \to \omega$ simultaneously along that subsequence for some $0 \le \zeta^2 \le C_X$ and $0 \le \omega \le 1$. By the assumed asymptotic determinism, $\omega^{\text{het}}_{n,\text{NDE}} \to \omega$ in probability along this subsequence. First assume $\zeta > 0$. By the Lindeberg condition and Equation (B.16),

$$\left(\tfrac{A_n}{\sigma^{\text{cf,het}}_{n,\text{NDE}}}, \tfrac{B_n}{\sigma_{X,\text{NDE}}}\right) \xrightarrow{d} (Z_1, Z_2), \quad Z_1, Z_2 \xrightarrow{\text{iid}} \mathcal{N}(0,1)\,.$$

Define $D^2_{n,\text{NDE}} = (\sigma^{\text{cf,het}}_{n,\text{NDE}})^2 + \sigma^2_{X,\text{NDE}}$. Then, by Slutsky's theorem,

$$\tfrac{A_n+B_n}{D_{n,\text{NDE}}} = \sqrt{\omega^{\text{het}}_{n,\text{NDE}}}\tfrac{A_n}{\sigma^{\text{cf,het}}_{n,\text{NDE}}} + \sqrt{1-\omega^{\text{het}}_{n,\text{NDE}}}\tfrac{B_n}{\sigma_{X,\text{NDE}}} \xrightarrow{d} \sqrt{\omega}\,Z_1 + \sqrt{1-\omega}\,Z_2 \sim \mathcal{N}(0,1).$$

On the other hand, if $\zeta = 0$, then $\mathbb{E}[B_n^2] = \sigma_{X,\mathrm{NDE}}^2 \to 0$ and consequently $B_n = o_p(1)$. Moreover, since $(\sigma_{n,\mathrm{NDE}}^{\mathrm{cf,het}})^2$ is bounded away from zero with probability going to one, $D_{n,\mathrm{NDE}}/\sigma_{n,\mathrm{NDE}}^{\mathrm{cf,het}} \to 1$ in probability. Thus,

$$\tfrac{A_n+B_n}{D_{n,\mathrm{NDE}}} = \tfrac{A_n}{\sigma_{n,\mathrm{NDE}}^{\mathrm{cf,het}}} + o_p(1) \xrightarrow{d} \mathcal{N}(0,1).$$

We have therefore established that given any subsequence, there exists a further subsequence along which

$$\tfrac{A_n+B_n}{\sqrt{(\sigma_{n,\mathrm{NDE}}^{\mathrm{cf,het}})^2+\sigma_{X,\mathrm{NDE}}^2}} \xrightarrow{d} \mathcal{N}(0,1).$$

As the limit is the same, the convergence holds for the entire sequence. This completes the proof. □

B.6 Proof of Theorem 5.14

Proof. On the treatment arm, the reduced form of the regression model satisfies $Y_i = X_i^\top \rho_t + \xi_{1,i}$ where $\rho_t = \beta_1 + \mathbf{B}_1^\top \gamma_1$ and $\xi_{1,i} = \gamma_1^\top U_i' + \epsilon_i$. Recall the definition of $\hat{\theta}_{11}^{\mathrm{cf}}$ (Algorithm 3):

$$\begin{aligned}
\hat{\theta}_{11}^{\mathrm{cf}} = \frac{\hat{\theta}_{11}^{(1)} + \hat{\theta}_{11}^{(2)}}{2} &= \frac{1}{2}\sum_{j=1}^{2}\left(\bar{X}^\top \hat{\rho}_t^{(j)} + (\hat{\omega}^{(j)})^\top (\mathbf{y}_{t,j} - \mathbf{X}_{t,j}\hat{\rho}_t^{(j)})\right) \\
&= \theta_{11} + \frac{1}{2}\sum_{j=1}^{2}\left((\bar{X} - \mathbf{X}_{t,j}^\top \hat{\omega}^{(j)})^\top (\hat{\rho}_t^{(j)} - \rho_t)\right) + \frac{1}{2}\sum_{j=1}^{2}(\hat{\omega}^{(j)})^\top \boldsymbol{\xi}_1^{(j)} \\
&:= \theta_{11} + \Delta_{n,11}^{\mathrm{cf}} + V_{n,11}^{\mathrm{cf}}.
\end{aligned}$$

As a consequence, we have:

$$\begin{aligned}
\frac{\sqrt{n}(\hat{\theta}_{\mathrm{NIE}}^{\mathrm{cf}} - \theta_{\mathrm{NIE}})}{\sigma_{n,\mathrm{NIE}}^{\mathrm{cf,het}}} &= \frac{\sqrt{n}((\hat{\theta}_{11}^{\mathrm{cf}} - \hat{\theta}_{10}^{\mathrm{cf}}) - (\theta_{11} - \theta_{10}))}{\sigma_{n,\mathrm{NIE}}^{\mathrm{cf,het}}} \\
&= \frac{\sqrt{n}(\Delta_{n,11}^{\mathrm{cf}} - \Delta_n^{\mathrm{cf}})}{\sigma_{n,\mathrm{NIE}}^{\mathrm{cf,het}}} + \frac{\sqrt{n}(V_{n,11}^{\mathrm{cf}} - V_n^{\mathrm{cf}})}{\sigma_{n,\mathrm{NIE}}^{\mathrm{cf,het}}}
\end{aligned}$$

where $(\Delta_n^{\mathrm{cf}}, V_n^{\mathrm{cf}})$ are same as defined in Equation (B.2). The proof structure is closely related to that of Theorem 5.9 and Theorem 5.13. First, we show that the bias term $\sqrt{n}(\Delta_{n,11}^{\mathrm{cf}} - \Delta_n^{\mathrm{cf}}) = o_p(1)$. Then we argue $\sigma_{n,\mathrm{NIE}}^{\mathrm{cf,het}}$ is bounded away from 0 with probability going to 1, and finally we show that the second term on the right-hand side of the above equation converges to $\mathcal{N}(0,1)$. For the first part, all we need to show is $\sqrt{n}\Delta_{n,11}^{\mathrm{cf}} = o_p(1)$ as we have already established in the proof of Theorem 5.9 that $\sqrt{n}\Delta_n^{\mathrm{cf}} = o_p(1)$. Towards that end, first observe that, by Assumption 5.3 and 5.4 $\xi_{1,i} = \gamma_1^\top U_i' + \epsilon_i$ is uniformly sub-Gaussian conditional on $\sigma(X_{1:n})$. Also, by Assumption 5.4, $\|\rho_t\|_0 = \|\beta_1 + \mathbf{B}_1^\top \gamma_1\|_0 \le k_\beta + sk_\gamma \le (1+s)k$. Furthermore, by Assumption 5.3 and Remark 5.7, the design matrix $\mathbf{X}_{t,j}$ satisfies RE condition with probability going to 1. Therefore, by standard Lasso concentration, we have:

$$\|\hat{\rho}_t^{(j)} - \rho_t\|_1 = O_p\left(sk\sqrt{\tfrac{\log p}{n}}\right), \quad \|\hat{\rho}_t^{(j)} - \rho_t\|_2 = O_p\left(\sqrt{\tfrac{sk\log p}{n}}\right).$$

As a consequence, we have:

$$\begin{aligned}
\sqrt{n}|\Delta_{n,11}^{\mathrm{cf}}| &\le \sqrt{n}\sum_{j=1}^{2}\left|(\bar{X} - \mathbf{X}_{t,j}^\top \hat{\omega}^{(j)})^\top (\hat{\rho}_t^{(j)} - \rho_t)\right| \\
&\le \sqrt{n}\sum_{j=1}^{2}\|\bar{X} - \mathbf{X}_{t,j}^\top \hat{\omega}^{(j)}\|_\infty \|\hat{\rho}_t^{(j)} - \rho_t\|_1 \\
&= O_p\left(\sqrt{n} \times \sqrt{\frac{\log p}{n}} \times sk\sqrt{\frac{\log p}{n}}\right) = o_p(1).
\end{aligned}$$

where, in the last conclusion we have used $(sk\log p)/\sqrt{n} = o(1)$ (Assumption 5.4). Furthermore, $\|\bar{X} - \mathbf{X}_{t,j}^\top \hat{\omega}^{(j)}\|_\infty = O_p(\sqrt{\log p/n})$ follows from the definition of the optimization problem (3.3). This establishes the first part. Now, to argue $\sigma^{\text{cf,het}}_{n,\text{NIE}}$ is bounded away from 0 with probability going to 1, it is immediate from the definition of $\sigma^{\text{cf,het}}_{n,\text{NIE}}$ that

$$(\sigma^{\text{cf,het}}_{n,\text{NIE}})^2 \geq \tfrac{n}{4} v^2_{U,0,1} \|\hat{\tau}_2^{(1)}\|_2^2 \geq \tfrac{\lambda_{\min}(\Sigma_{U,0}) c_\gamma^2}{4} n \|\hat{\tau}_2^{(1)}\|_2^2$$

We have already argued in the proof of Theorem 5.9 that $n\|\hat{\tau}_2^{(1)}\|_2^2$ bounded away from 0 with probability going to 1 (see Equation (B.6)). Hence $\sigma^{\text{cf,het}}_{n,\text{NIE}}$ is bounded away from 0 with probability going to 1. Finally, we will establish the asymptotic normality. Towards that end, define an oracle balancing weight $\tau_1^{(j),*}$ as:

$$\tau_{1,i}^{(j),*} = \tfrac{1}{n_{t,j}} W_i^\top \eta^*, \quad \eta^* = \Omega_{W,1}^{-1} a_{10}^* = \Omega_{W,1}^{-1} \begin{pmatrix} \mu_X \\ \mathbf{B}_0 \mu_X \end{pmatrix}, \quad 1 \leq i \leq n_{t,j}\,.$$

First we argue that:

$$\sqrt{n}\|\hat{\tau}_1^{(j)} - \tau_1^{(j),*}\|_2 = o_p(1)\,. \tag{B.17}$$

The proof is similar to the proof of Lemma B.1. Fix $j = 2$ (we choose $j = 2$ because it is explicitly mentioned in Algorithm 1; the proof for $j = 1$ is verbatim). Recall that the main difference between $\hat{\tau}_1^{(2)}$ is obtained by solving the optimization problem (3.2) with $a^{(1)} = (\bar{X}^\top, \hat{m}_{c,1}^\top)^\top$. Let $\hat{\eta}_n := \hat{\eta}_n(a^{(1)})$ be the rescaled dual of $\hat{\tau}_1^{(2)}$. Then from Equation (C.6) inside the proof of Lemma B.1, we have:

$$\begin{aligned}
\hat{\eta}_n &= \arg\min_\eta \left\{ \frac{1}{n_t} \sum_{i=1}^{n_t} H_{n_t b_n}\left((\mathbf{W}_t \eta)_i\right) - \eta^\top a^{(1)} + \check{\lambda}_{n_t} \|\eta\|_1 \right\} \\
&= \arg\min_\eta \ \left\{ \mathcal{L}_{n,a^{(1)}}(\eta) + \check{\lambda}_{n_t} \|\eta\|_1 \right\}\,.
\end{aligned}$$

where n_t is a shorthand for $n_{t,2}$ and $\check{\lambda}_{n_t} = K_1 \sqrt{\log(p+q)/n_t}$ for some appropriately chosen large constant K_1. Now observe that:

$$\max_{1\leq i\leq n_t} |W_{t,i}^\top \eta^*| \leq \|\eta^*\|_1 \max_i \|W_{t,i}\|_\infty = O_p(\sqrt{\log n + \log d}) = o(n^{1-\alpha})\,.$$

Therefore, all large n, with probability going to 1, $H_{n_t b_n}(W_{t,i}^\top \eta^*) = (W_{t,i}^\top \eta^*)^2/2$ for all $1 \leq i \leq n_t$. Now from optimality of $\hat{\eta}_n$, we have:

$$\mathcal{L}_{n,a^{(1)}}(\hat{\eta}_n) + \check{\lambda}_{n_t} \|\hat{\eta}_n\|_1 \leq \mathcal{L}_{n,a^{(1)}}(\eta^*) + \check{\lambda}_{n_t} \|\eta^*\|_1$$

Then, by the same calculation as used to obtain Equation (C.10) using convexity of $\mathcal{L}_{n,a^{(1)}}(\cdot)$, we conclude that $\|\hat{\eta}_n\|_1 \leq 3\|\eta^*\|_1 \leq 3C_\eta$ with probability going 1. We next claim that $\|\mathbf{W}_t(\hat{\eta}_n - \eta^*)\|_2^2/n_t = o_p(1)$. Now, as we have argued $\|\hat{\eta}_n\|_1 \leq 3C_\eta$ with probability going to 1, hence on this event,

$$\max_{1\leq i\leq n_t} |W_{t,i}^\top \hat{\eta}_n| \leq \|\hat{\eta}_n\|_1 \max_i \|W_{t,i}\|_\infty = O_p(\sqrt{\log n + \log(p+q)}) = o(n^{1-\alpha})\,.$$

Hence, we also have on this event, $H_{n_t b_n}(W_{t,i}^\top \hat{\eta}_n) = (W_{t,i}^\top \hat{\eta}_n)^2/2$. Therefore, we have, again from the optimality of $\hat{\eta}_n$:

$$\tfrac{1}{2n_t}\|\mathbf{W}_t \hat{\eta}_n\|_2^2 - \hat{\eta}_n^\top a^{(1)} + \check{\lambda}_{n_t}\|\hat{\eta}_n\|_1 \leq \tfrac{1}{2n_t}\|\mathbf{W}_t \eta^*\|_2^2 - (\eta^*)^\top a^{(1)} + \check{\lambda}_{n_t}\|\eta^*\|_1$$

Hence,

$$\begin{aligned}
&\frac{1}{2n_t}\|\mathbf{W}_t(\hat{\eta}_n - \eta^*)\|_2^2 + \check{\lambda}_{n_t}\|\hat{\eta}_n\|_1 \\
&\leq (\eta^*)^\top \hat{\Omega}_{W,t}(\eta^* - \hat{\eta}_n) + (\hat{\eta}_n - \eta^*)^\top a^{(1)} + \check{\lambda}_{n_t}\|\eta^*\|_1 \\
&= (\eta^* - \hat{\eta}_n)^\top (\hat{\Omega}_{W,t}\eta^* - a^{(1)}) + \check{\lambda}_{n_t}\|\eta^*\|_1 \\
&= (\eta^* - \hat{\eta}_n)^\top (\hat{\Omega}_{W,t}\eta^* - a_{10}^* + a_{10}^* - a^{(1)}) + \check{\lambda}_{n_t}\|\eta^*\|_1
\end{aligned}$$

$$= (\eta^* - \hat{\eta}_n)^\top ((\hat{\Omega}_{W,t} - \Omega_{W,1})\eta^* + a_{10}^* - a^{(1)}) + \check{\lambda}_{n_t}\|\eta^*\|_1 \tag{B.18}$$

Now,

$$\|(\hat{\Omega}_{W,t} - \Omega_{W,1})\eta^*\|_\infty \le \|\eta^*\|_1 \|\hat{\Omega}_{W,t} - \Omega_{W,1}\|_{\infty,\infty} \le C_\eta C\sqrt{\tfrac{\log p+q}{n}}$$

with probability going to 1, where the last inequality follows from the fact that the elements of $\hat{\Omega}_{W,t}$ are sub-exponential and consequently an application of Bernstein's inequality, along with a union bound, yields the conclusion. Furthermore, by same calculation as in Equation (C.8), we have, with probability going to 1:

$$\|a_{10}^* - a^{(1)}\|_\infty \le C'\sqrt{\tfrac{\log(p+q)}{n}}\,.$$

Hence, choosing K_1 large enough, we conclude with probability going to 1:

$$\|(\hat{\Omega}_{W,t} - \Omega_{W,1})\eta^*\|_\infty + \|a_{10}^* - a^{(1)}\|_\infty \le \tfrac{\check{\lambda}_{n_t}}{2}\,.$$

Therefore, going back to Equation (B.18), we have:

$$\tfrac{1}{2n_t}\|\mathbf{W}_t(\hat{\eta}_n - \eta^*)\|_2^2 + \check{\lambda}_{n_t}\|\hat{\eta}_n\|_1 \le \|\hat{\eta}_n - \eta^*\|_1 \tfrac{\check{\lambda}_{n_t}}{2} + \check{\lambda}_{n_t}\|\eta^*\|_1$$

Therfore, from the fact $\|\hat{\eta}_n\|_1 \le 3C_\eta$, $\|\eta^*\|_1 \le C_\eta$ and $\check{\lambda}_{n_t} = O(\sqrt{\log(p+q)/n})$, we have:

$$\tfrac{1}{2n_t}\|\mathbf{W}_t(\hat{\eta}_n - \eta^*)\|_2^2 = O_p\left(\sqrt{\tfrac{\log(p+q)}{n}}\right) = o_p(1)\,.$$

Now, going back to KKT condition of $\hat{\tau}_1^{(2)}$ (see Equation (C.4) in the proof of Lemma B.1 for more details):

$$\hat{\tau}_1^{(2)} = \mathbf{W}_{t,2}\hat{\eta}^{\mathrm{KKT}}(a^{(1)}) - \hat{\zeta}^{\mathrm{KKT}}(a^{(1)}) = \tfrac{1}{n_t}\mathbf{W}_t\hat{\eta}_n - \hat{\zeta}^{\mathrm{KKT}}(a^{(1)})$$

where we used the fact $\hat{\eta}_n = n_t\hat{\eta}^{\mathrm{KKT}}$. We next argue that $\hat{\zeta}^{\mathrm{KKT}}(a^{(1)}) = 0$. Recall that from the definition of $\hat{\zeta}^{\mathrm{KKT}}(a^{(1)})$, it is 0 if the condition $\|\hat{\tau}^{(1)}\|_\infty \le n^{-\alpha}$ is strict, i.e. $\|\hat{\tau}^{(1)}\|_\infty < n^{-\alpha}$. Towards that end, recall the rescaled dual optimization problem (Equation (C.5) in the proof of Lemma B.1):

$$(\hat{\eta}_n, \hat{\zeta}_n) = \arg\min_{\eta,\zeta}\ \left\{\tfrac{1}{2n_t}\|\mathbf{W}_t\eta - \zeta\|_2^2 - \eta^\top a + \check{\lambda}_{n_t}\|\eta\|_1 + b_n\|\zeta\|_1\right\}$$

Therefore, given $\hat{\eta}_n$, $\hat{\zeta}_n$ can be written as:

$$\hat{\zeta}_{n,i} = \mathsf{sign}(W_{t,i}^\top\hat{\eta}_n)(|W_{t,i}^\top\hat{\eta}_n| - n_t^{1-\alpha})_+, \qquad 1 \le i \le n_t\,.$$

However, we have argue previously that with probabiity going to 1, $|W_{t,i}^\top\hat{\eta}_n| \le C\sqrt{\log(p+q) + \log n} = o(n_t^{1-\alpha})$ and consequently on this even $\hat{\zeta}_{n,i} = 0$ for all $1 \le i \le n_t$. As $\zeta^{\mathrm{KKT}} = \hat{\zeta}_n/n_t$, we have $\hat{\zeta}^{\mathrm{KKT}} = 0$. Therefore,

$$\hat{\tau}_1^{(2)} = \tfrac{1}{n_t}\mathbf{W}_{t,2}\hat{\eta}_n \implies n\|\hat{\tau}_1^{(2)} - \tau_1^{(2),*}\|_2^2 = \tfrac{n}{n_t}\tfrac{1}{n_t}\|\mathbf{W}_{t,2}(\hat{\eta}_n - \eta^*)\|_2^2 = o_p(1)\,,$$

as $n \asymp n_t$. This completes the proof.

Now coming back to the proof of asymptotic normality, we can write the relevant part as:

$$\begin{aligned}
T &:= \frac{\sqrt{n}(V_{n,11}^{\mathrm{cf}} - V_n^{\mathrm{cf}})}{\sigma_{n,\mathrm{NIE}}^{\mathrm{cf,het}}} \\
&= \frac{\sqrt{n}}{2\sigma_{n,\mathrm{NIE}}^{\mathrm{cf,het}}}\sum_{j=1}^{2}\left[(\hat{\omega}^{(j)})^\top\boldsymbol{\xi}_1^{(j)} - (\hat{\tau}_1^{(j)})^\top\boldsymbol{\epsilon}^{(j)} - (\hat{\tau}_2^{(j)})^\top\mathbf{U}\gamma_1\right] \\
&= \frac{\sqrt{n}}{2\sigma_{n,\mathrm{NIE}}^{\mathrm{cf,het}}}\sum_{j=1}^{2}\left[(\hat{\omega}^{(j)})^\top\mathbf{U}'\gamma_1 + (\hat{\omega}^{(j)})^\top\boldsymbol{\epsilon}^{(j)} - (\hat{\tau}_1^{(j)})^\top\boldsymbol{\epsilon}^{(j)} - (\hat{\tau}_2^{(j)})^\top\mathbf{U}\gamma_1\right]
\end{aligned}$$

$$= \frac{\sqrt{n}}{2\sigma^{\mathrm{cf,het}}_{n,\mathrm{NIE}}} \sum_{j=1}^{2} \left[(\hat{\omega}^{(j)} - \hat{\tau}_1^{(j)})^\top \boldsymbol{\epsilon}^{(j)} + (\hat{\omega}^{(j)})^\top \mathbf{U}' \gamma_1 - (\hat{\tau}_2^{(j)})^\top \mathbf{U} \gamma_1 \right] := \frac{L}{\sigma^{\mathrm{cf,het}}_{n,\mathrm{NIE}}} .$$

Now, define an oracle version T^0 with replacing $\hat{\tau}_1^{(j)}$ by $\tau^{(j),*}$:

$$T^0 = \frac{\sqrt{n}}{2\sigma^{\mathrm{cf,het},0}_{n,\mathrm{NIE}}} \sum_{j=1}^{2} \left[(\hat{\omega}^{(j)} - \tau_1^{(j),*})^\top \boldsymbol{\epsilon}^{(j)} + (\hat{\omega}^{(j)})^\top \mathbf{U}' \gamma_1 - (\hat{\tau}_2^{(j)})^\top \mathbf{U} \gamma_1 \right] := \frac{L^0}{\sigma^{\mathrm{cf,het},0}_{n,\mathrm{NIE}}} .$$

where

$$(\sigma^{\mathrm{cf,het},0}_{n,\mathrm{NIE}})^2 = \frac{n}{4} \sum_{j=1}^{2} \left[\sum_{i \in \mathcal{D}_j, A_i = 1} \mathbb{E}\left[(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 \nu_1(X_i, M_i) \mid \mathcal{X}_n \right] \right.$$
$$\left. + v^2_{U,1,1} \|\hat{\omega}^{(j)}\|_2^2 + v^2_{U,0,1} \|\hat{\tau}_2^{(j)}\|_2^2 \right] .$$

Here $\mathcal{X}_n = \sigma(A_{1:n}, X_{1:n})$. We will also use $\mathcal{H}_n = \sigma(A_{1:n}, X_{1:n}, \{M_i : A_i = 1\})$ and $\mathcal{G}_n = \sigma(A_{1:n}, X_{1:n}, M_{1:n})$. Clearly $\mathcal{X}_n \subseteq \mathcal{H}_n \subseteq \mathcal{G}_n$. We first argue $T^0 \xrightarrow{d} \mathcal{N}(0,1)$. Towards that end, define

$$\zeta^{\mathrm{het,NIE}}_{n,i} = \frac{\sqrt{n}}{2\sigma^{\mathrm{cf,het},0}_{n,\mathrm{NIE}}} \sum_{j=1}^{2} \left[\left\{ (\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*}) \epsilon_i + \hat{\omega}_i^{(j)} (U_i')^\top \gamma_1 \right\} \mathbb{1}_{\substack{i \in \mathcal{D}_j \\ A_i = 1}} \right.$$
$$\left. - \hat{\tau}_{2,i}^{(j)} U_i^\top \gamma_1 \mathbb{1}_{\substack{i \in \mathcal{D}_j \\ A_i = 0}} \right]$$

It is immediate that conditional on $\mathcal{X}_n$, we have $\mathbb{E}[\zeta^{\mathrm{het,NIE}}_{n,i} \mid \mathcal{X}_n] = 0$. This follows from the fact that conditional on $\mathcal{X}_n$, $\hat{\omega}^{(j)}$ and $\hat{\tau}_2^{(j)}$ are fixed. Furthermore:

$$\mathbb{E}[\tau_{1,i}^{(j),*} \epsilon_i \mid \mathcal{X}_n] = \mathbb{E}\left[\mathbb{E}[\tau_{1,i}^{(j),*} \epsilon_i \mid \mathcal{H}_n] \mid \mathcal{X}_n \right] = \mathbb{E}\left[\tau_{1,i}^{(j),*} \mathbb{E}[\epsilon_i \mid \mathcal{H}_n] \mid \mathcal{X}_n \right] = 0 .$$

Here we have used the fact that $\tau_1^{(j),*}$ is measurable with respect to $\mathcal{H}_n$. Furthermore, for $i \in \mathcal{D}_j$ with $A_i = 1$

$$\begin{aligned}
&\mathsf{var}(\zeta^{\mathrm{het,NIE}}_{n,i} \mid \mathcal{X}_n) \\
&= \frac{n}{4(\sigma^{\mathrm{cf,het},0}_{n,\mathrm{NIE}})^2} \Big\{ \mathbb{E}[(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 \epsilon_i^2 \mid \mathcal{X}_n] + \mathbb{E}[(\hat{\omega}_i^{(j)})^2 ((U_i')^\top \gamma_1)^2 \mid \mathcal{X}_n] \\
&\qquad + 2\mathbb{E}[(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})(\omega_i^{(j)}) \epsilon_i (U_i')^\top \gamma_1 \mid \mathcal{X}_n] \Big\} \\
&= \frac{n}{4(\sigma^{\mathrm{cf,het},0}_{n,\mathrm{NIE}})^2} \Big\{ \mathbb{E}[(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 \nu_1(X_i, M_i) \mid \mathcal{X}_n] + (\hat{\omega}_i^{(j)})^2 \gamma_1^\top \Sigma_{U,1} \gamma_1 \Big\}
\end{aligned}$$

The expectation of product/cross term is 0 as $\mathbb{E}[\epsilon_i \mid \mathcal{H}_n] = 0$ and $\hat{\omega}_i^{(j)}$ and $(U_i')^\top \gamma_1$ is measurable with respect to $\mathcal{H}_n$. On the other hand, for $i \in \mathcal{D}_j$ with $A_i = 0$, we have:

$$\mathsf{var}(\zeta^{\mathrm{het,NIE}}_{n,i} \mid \mathcal{X}_n) = \frac{n}{4(\sigma^{\mathrm{cf,het},0}_{n,\mathrm{NIE}})^2} \mathbb{E}[(\hat{\tau}_{2,i}^{(j)} U_i^\top \gamma_1)^2 \mid \mathcal{X}_n] = \frac{n}{4(\sigma^{\mathrm{cf,het},0}_{n,\mathrm{NIE}})^2} (\hat{\tau}_{2,i}^{(j)})^2 \gamma_1^\top \Sigma_{U,0} \gamma_1 .$$

Furthermore, the terms $\{\zeta^{\mathrm{het,NIE}}_{n,i}\}$ are independent across i conditionally on $\mathcal{X}_n$. As a consequence,

$$\begin{aligned}
&\mathsf{var}\left(\sum_{i=1}^{n} \zeta^{\mathrm{het,NIE}}_{n,i} \mid \mathcal{X}_n \right) \\
&= \frac{n}{4(\sigma^{\mathrm{cf,het},0}_{n,\mathrm{NIE}})^2} \sum_{j=1}^{2} \left[\sum_{i \in \mathcal{D}_j, A_i = 1} \mathbb{E}\left[(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 \nu_1(X_i, M_i) \mid \mathcal{X}_n \right] \right.
\end{aligned}$$

$$+v_{U,1,1}^2\|\hat{\omega}^{(j)}\|_2^2 + v_{U,0,1}^2\|\hat{\tau}_2^{(j)}\|_2^2\Big]$$

$$= 1\,.$$

Therefore, all we need to show is

$$\textstyle\sum_{i=1}^n \mathbb{E}\left[\left|\zeta_{n,i}^{\text{het,NIE}}\right|^3 \mid \mathcal{X}_n\right] \xrightarrow{p} 0\,.$$

To that end,

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\left[\left|\zeta_{n,i}^{\text{het,NIE}}\right|^3 \mid \mathcal{X}_n\right] \\
&= \frac{n^{3/2}}{8(\sigma_{n,\text{NIE}}^{\text{cf,het},0})^3}\sum_i\sum_{j=1}^2\Big\{\mathbb{E}\left[|(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})\epsilon_i + \hat{\omega}_i^{(j)}(U_i')^\top\gamma_1|^3 \mid \mathcal{X}_n\right]\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=1}} \\
&\qquad\qquad + |\hat{\tau}_{2,i}^{(j)}|^3\mathbb{E}\left[|U_i^\top\gamma_1|^3 \mid \mathcal{X}_n\right]\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=0}}\Big\} \\
&\le \frac{n^{3/2}}{8(\sigma_{n,\text{NIE}}^{\text{cf,het},0})^3}\sum_i\sum_{j=1}^2\Big\{4\mathbb{E}\left[|(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})\epsilon_i|^3 \mid \mathcal{X}_n\right]\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=1}} \\
&\qquad\qquad + 4|\hat{\omega}_i^{(j)}|^3\mathbb{E}\left[|(U_i')^\top\gamma_1|^3 \mid \mathcal{X}_n\right]\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=1}} + |\hat{\tau}_{2,i}^{(j)}|^3\mathbb{E}\left[|U_i^\top\gamma_1|^3 \mid \mathcal{X}_n\right]\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=0}}\Big\}
\end{aligned}$$

Now from the fact $\|\gamma_1\|_2 \le C_\gamma$ (Assumption 5.4) and U_i are sub-Gaussian (Assumption 5.3), $\mathbb{E}[|U_i^\top\gamma|^3] \vee \mathbb{E}[|(U_i')^\top\gamma_1|^3] \le C$. Furthermore, from uniform sub-Gaussianity of ϵ (Assumption 5.3), we have; $\mathbb{E}[|\epsilon_i|^3 \mid \mathcal{H}_n] \le C$. Therefore,

$$\begin{aligned}
&\sum_{i=1}^n \mathbb{E}\left[\left|\zeta_{n,i}^{\text{het,NIE}}\right|^3 \mid \mathcal{X}_n\right] \\
&\le \frac{Cn^{3/2}}{8(\sigma_{n,\text{NIE}}^{\text{cf,het}})^3}\sum_i\sum_{j=1}^2\Big\{4\mathbb{E}\left[|\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*}|^3 \mid \mathcal{X}_n\right]\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=1}} \\
&\qquad\qquad + 4|\hat{\omega}_i^{(j)}|^3\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=1}} + |\hat{\tau}_{2,i}^{(j)}|^3\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=0}}\Big\} \\
&\le \frac{C_1n^{3/2}}{8(\sigma_{n,\text{NIE}}^{\text{cf,het}})^3}\sum_i\sum_{j=1}^2\left\{|\hat{\omega}_i^{(j)}|^3\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=1}} + \mathbb{E}\left[|\tau_{1,i}^{(j),*}|^3 \mid \mathcal{X}_n\right]\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=1}} + |\hat{\tau}_{2,i}^{(j)}|^3\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=0}}\right\} \\
&\le \frac{C_1n^{3/2}}{8(\sigma_{n,\text{NIE}}^{\text{cf,het}})^3}\sum_{j=1}^2\Big\{\|\hat{\omega}^{(j)}\|_2^2\|\hat{\omega}^{(j)}\|_\infty + \|\hat{\tau}_2^{(j)}\|_2^2\|\hat{\tau}_2^{(j)}\|_\infty \\
&\qquad\qquad + \sum_i\mathbb{E}\left[|\tau_{1,i}^{(j),*}|^3 \mid \mathcal{X}_n\right]\mathbb{1}_{\substack{i\in\mathcal{D}_j\\A_i=1}}\Big\}
\end{aligned}$$

In Lemma 5.8 we have established that $n\|\hat{\tau}_2^{(j)}\|_2^2 = O_p(1)$. By the exact same argument one can prove that $n\|\hat{\omega}^{(j)}\|_2^2 = O_p(1)$ (as they are obtained from similar optimization problem with $\mathbf{X}_{c,j}$ replaced by $\mathbf{X}_{t,j}$). Furthermore, by the constraint in Equation (3.1) (resp. (3.3)) we have $\|\hat{\tau}_2^{(j)}\|_\infty \le n_{c,j}^{-\alpha} \le Cn^{-\alpha}$ (resp. $\|\hat{\omega}^{(j)}\|_\infty \le n_{t,j}^{-\alpha} \le Cn^{-\alpha}$ where we have used the fact that $n_{c,j} \asymp n_{t,j} \asymp n$. Therefore, as $\alpha > 1/2$, we have:

$$n^{3/2}\|\hat{\omega}^{(j)}\|_2^2\|\hat{\omega}^{(j)}\|_\infty + n^{3/2}\|\hat{\tau}_2^{(j)}\|_2^2\|\hat{\tau}_2^{(j)}\|_\infty = O_p(n^{(-1-\alpha+3/2)}) = o_p(1)$$

Next we argue that $n^{3/2}\sum_j\sum_i\mathbb{E}[|\tau_{1,i}^{(j),*}|^3 \mid \mathcal{X}_n]\mathbb{1}_{i\in\mathcal{D}_j,A_i=1} = o_p(1)$. Define $Q_{n,j}$ as:

$$Q_{n,j} = \tfrac{1}{n_j}\textstyle\sum_{i\in\mathcal{D}_j}\mathbb{E}\left[|W_i^\top\eta^*|^3\mathbb{1}_{A_i=1} \mid \mathcal{X}\right]$$

Therefore,

$$\mathbb{E}[Q_{n,j}] = \mathbb{E}\left[|W^\top \eta^*|^3 \mathbb{1}_{A=1}\right] \leq C\,.$$

Here we have used the fact that $W \mid A = 1$ is sub-Gaussian and $\|\eta^*\|_2 \leq \|\eta^*\|_1 \leq C_\eta$. Therefore, $Q_{n,j} = O_p(1)$. Now,

$$\begin{aligned}
n^{3/2} \sum_j \sum_{i \in \mathcal{D}_j} \mathbb{E}\left[|\tau_{1,i}^{(j),*}|^3 \mid \mathcal{X}_n \mathbb{1}_{A_i=1}\right] &= n^{3/2} \sum_j \frac{1}{n_{t,j}^3} \sum_{i \in \mathcal{D}_j} \mathbb{E}\left[|W_i^\top \eta^*|^3 \mathbb{1}_{A_i=1}\right] \\
&= n^{3/2} \sum_j \frac{n_j}{n_{t,j}^3} \frac{1}{n_j} \sum_{i \in \mathcal{D}_j} \mathbb{E}\left[|W_i^\top \eta^*|^3 \mathbb{1}_{A_i=1}\right] \\
&= n^{3/2} \sum_j \frac{n_j}{n_{t,j}^3} Q_{n,j} = O_p(n^{-1/2}) = o_p(1)\,.
\end{aligned}$$

Here the last line follows from the fact that $Q_{n,j} = O_p(1)$, $n_j \asymp n_{t,j} \asymp n$. Now, we argue that $\sigma_{n,\text{NIE}}^{\text{cf,het},0}$ is bounded away from 0 with probability going to 1. This is very similar to the argument we used to prove $\sigma_{n,\text{NIE}}^{\text{cf,het}}$ is bounded away from 0 with probability going to 1 by observing that:

$$(\sigma_{n,\text{NIE}}^{\text{cf,het},0})^2 \geq (n/4)\|\hat{\tau}_2^{(1)}\|_2^2 v_{U,0,1}^2 \geq (\lambda_{\min}(\Sigma_{U,0}) c_\gamma^2/4)\; n\|\hat{\tau}_2^{(1)}\|_2^2$$

and the fact that $n\|\hat{\tau}_2^{(1)}\|_2^2$ is bounded away from 0 with probability going to 1. Therefore, by conditional Lindeberg-Feller CLT, we conclude that $T^0 \xrightarrow{d} \mathcal{N}(0,1)$. Finally, to conclude the proof, we show that $T - T^0 = o_p(1)$. Using the same expansion as in Equation B.9, we have:

$$T - T^0 = \frac{L}{\sigma_{n,\text{NIE}}^{\text{cf,het}}} - \frac{L^0}{\sigma_{n,\text{NIE}}^{\text{cf,het},0}} = \frac{L - L^0}{\sigma_{n,\text{NIE}}^{\text{cf,het}}} + \frac{L^0}{\sigma_{n,\text{NIE}}^{\text{cf,het},0}} \left(\frac{\sigma_{n,\text{NIE}}^{\text{cf,het},0}}{\sigma_{n,\text{NIE}}^{\text{cf,het}}} - 1\right)$$

So to conclude $T - T^0 = o_p(1)$, all we need to prove is $L - L^0 = o_p(1)$ and $|\sigma_{n,\text{NIE}}^{\text{cf,het}} - \sigma_{n,\text{NIE}}^{\text{cf,het},0}| = o_p(1)$ as these, along with the fact that $\sigma_{n,\text{NIE}}^{\text{cf,het}}$ is bounded away from 0 with probability going to 1 concludes the proof. For the former,

$$L - L^0 = \frac{\sqrt{n}}{2} \sum_{j=1}^2 (\tau_1^{(j),*} - \hat{\tau}_1^{(j)})^\top \boldsymbol{\epsilon}^{(j)}$$

Therefore,

$$\begin{aligned}
\mathbb{E}[(L - L^0)^2 \mid \mathcal{G}_n] &= \frac{n}{4} \sum_{j=1}^2 \sum_{i \in \mathcal{D}_j, A_i = 1} (\tau_{1,i}^{(j),*} - \hat{\tau}_{1,i}^{(j)})^2 \nu_1(X_i, M_i) \\
&\leq \frac{c_+}{4} \sum_j n\|\tau_1^{(j),*} - \hat{\tau}_1^{(j)}\|_2^2 = o_p(1)
\end{aligned}$$

where the last line follows from Equation (B.17). For the latter,

$$\begin{aligned}
&\left|(\sigma_{n,\text{NIE}}^{\text{cf,het}})^2 - (\sigma_{n,\text{NIE}}^{\text{cf,het},0})^2\right| \\
&= \frac{n}{4} \sum_{j=1}^2 \left| \sum_{\substack{i \in \mathcal{D}_j \\ A_i = 1}} \mathbb{E}[(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 \nu_1(X_i, M_i) \mid \mathcal{X}_n] - (\hat{\omega}_i^{(j)} - \hat{\tau}_{1,i}^{(j)})^2 \nu_1(X_i, M_i) \right| \\
&\leq \frac{n}{4} \sum_{j=1}^2 \left| \sum_{\substack{i \in \mathcal{D}_j \\ A_i = 1}} \mathbb{E}[(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 \nu_1(X_i, M_i) \mid \mathcal{X}_n] - (\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 \nu_1(X_i, M_i) \right| \\
&\quad + \frac{n}{4} \sum_{j=1}^2 \left| \sum_{\substack{i \in \mathcal{D}_j \\ A_i = 1}} (\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 \nu_1(X_i, M_i) - (\hat{\omega}_i^{(j)} - \hat{\tau}_{1,i}^{(j)})^2 \nu_1(X_i, M_i) \right|
\end{aligned}$$

$$:= S_1 + S_2 \,.$$

To bound S_2,

$$\begin{aligned} S_2 &\le \frac{nc_+}{4} \sum_{j=1}^{2} \sum_{i \in \mathcal{D}_j, A_i = 1} \left| (\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 - (\hat{\omega}_i^{(j)} - \hat{\tau}_{1,i}^{(j)})^2 \right| \\ &\le \frac{nc_+}{4} \sum_{j=1}^{2} \|2\hat{\omega}^{(j)} - \tau_1^{(j),*} - \hat{\tau}_1^{(j)}\|_2 \|\tau_1^{(j),*} - \hat{\tau}_1^{(j)}\|_2 \end{aligned}$$

We have already established that $\sqrt{n}\|\tau_1^{(j),*} - \hat{\tau}_1^{(j)}\|_2 = o_p(1)$ (Equation (B.17)). Furthermore, we know that $n\|\hat{\tau}_1^{(j)}\|_2^2 = O_p(1)$ from Lemma 5.8. By exact same argument, we also have $n\|\hat{\omega}^{(j)}\|_2^2 = O_p(1)$. Now, from the definition of $\tau_1^{(j),*}$ we have:

$$n\|\tau_1^{(j),*}\|_2^2 = \tfrac{n}{n_{t,j}^2} \textstyle\sum_{i \in \mathcal{D}_j} |W_i^\top \eta^*|^2 \mathbb{1}_{A_i = 1}$$

Now, from the sub-Gaussianity of $W \mid A = 1$, we have $\mathbb{E}[|W^\top \eta^*|^2 \mathbb{1}(A = 1)] \le C < \infty$. As a consequence,

$$\tfrac{1}{n_j} \textstyle\sum_{i \in \mathcal{D}_j} |W_i^\top \eta^*|^2 \mathbb{1}_{A_i = 1} = O_p(1) \,.$$

Therefore,

$$n\|\tau_1^{(j),*}\|_2^2 = \tfrac{n n_j}{n_{j,t}^2} \times \tfrac{1}{n_j} \textstyle\sum_{i \in \mathcal{D}_j} |W_i^\top \eta^*|^2 \mathbb{1}_{A_i = 1} = O_p(1) \,.$$

Here, the last conclusion follows from $n_{j,t} \asymp n_j \asymp n$. Therefore, $S_2 = o_p(1)$. Now we argue that $S_1 = o_p(1)$. Set

$$\upsilon_i^{(j)} = (\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 \nu_1(X_i, M_i) \mathbb{1}_{A_i = 1} - \mathbb{E}\left[(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^2 \nu_1(X_i, M_i) \mathbb{1}_{A_i = 1} \mid \mathcal{X}_n \right]$$

Then,

$$S_1 = \tfrac{n}{4} \textstyle\sum_j n_j \left| \tfrac{1}{n_j} \sum_{i \in \mathcal{D}_j} \upsilon_i^{(j)} \right| \,.$$

Now, for any $i \in \mathcal{D}_j$,

$$\begin{aligned} \mathsf{var}\left(\upsilon_i^{(j)} \mid \mathcal{X}_n \right) &\le \mathbb{E}\left[(\hat{\omega}_i^{(j)} - \tau_{1,i}^{(j),*})^4 \nu_1^2(X_i, M_i) \mathbb{1}_{A_i = 1} \mid \mathcal{X}_n \right] \\ &\le 8c_+^2 \left\{ (\hat{\omega}_i^{(j)})^4 \mathbb{1}(A_i = 1) + \mathbb{E}[(\tau_{1,i}^{(j),*})^4 \mathbb{1}(A_i = 1) \mid \mathcal{X}_n] \right\} \end{aligned}$$

Hence,

$$\begin{aligned} \mathsf{var}\left(\frac{1}{n_j} \sum_{i \in \mathcal{D}_j} \upsilon_i \mid \mathcal{X}_n \right) &= \frac{1}{n_j^2} \sum_{i \in \mathcal{D}_j} \mathsf{var}(\upsilon_i \mid \mathcal{X}_n) \\ &\le \frac{8c_+^2}{n_j^2} \|\hat{\omega}^{(j)}\|_2^2 \|\hat{\omega}^{(j)}\|_\infty^2 + \frac{8c_+^2}{n_j^2} \sum_{i \in \mathcal{D}_j} \mathbb{E}[(\tau_{1,i}^{(j),*})^4 \mathbb{1}(A_i = 1) \mid \mathcal{X}_n] \end{aligned}$$

Now, we have already argued that $\|\hat{\omega}^{(j)}\|_2^2 = O_p(n^{-1})$ and by feasibility of $\|\hat{\omega}^{(j)}\|_\infty^2 \le O_p(n^{-2\alpha})$. Consequently $\|\hat{\omega}^{(j)}\|_2^2 \|\hat{\omega}^{(j)}\|_\infty^2 / n_j^2 = O_p(n^{-3-2\alpha})$. On other other using the fact $\tau_{1,i}^{(j)} = (W_i^\top \eta^*)/n_{t,j}$ and $W \mid A = 1$ is sub-Gaussian along with $\|\eta^*\|_2 \le \|\eta^*\|_1 \le C_\eta$, we have:

$$\frac{8c_+^2}{n_j^2} \sum_{i \in \mathcal{D}_j} \mathbb{E}[(\tau_{1,i}^{(j),*})^4 \mid \mathcal{X}_n] = \frac{8c_+^2}{n_j n_{t,j}^4} \frac{1}{n_j} \sum_{i \in \mathcal{D}_j} \mathbb{E}[|W_i^\top \eta^*|^4 \mathbb{1}(A_i = 1)] = O_p(n_j^{-5}) \,.$$

Here the last conclusion follows from the fact that $n_{t,j} \asymp n_j \asymp n$. Therefore,

$$\mathsf{var}\left(\tfrac{1}{n_j} \textstyle\sum_{i \in \mathcal{D}_j} \upsilon_i \mid \mathcal{X}_n \right) = O_p(n^{-3-2\alpha}) \implies \tfrac{1}{n_j} \textstyle\sum_{i \in \mathcal{D}_j} \upsilon_i = O_p(n^{-3/2-\alpha}) \,.$$

Therefore,

$$S_1 = \frac{n}{4}\sum_j n_j \frac{1}{n_j}\sum_{i\in\mathcal{D}_j} v_i^{(j)} = O_p(n^{2-3/2-\alpha}) = O_p(n^{1/2-\alpha}) = o_P(1)\,,$$

as $\alpha > 1/2$. This completes the proof.

Finally, we establish the asymptotic normality of $\hat{\theta}^{\text{cf}}_{\text{NIE}}$ around the population parameter $\theta^{\text{pop}}_{\text{NIE}}$. The proof follows the same argument as the population extensions for θ_{10} and θ_{NDE}; therefore, we only highlight the changes specific to the NIE. To keep notational similarity, define $A_n = \sqrt{n}(\hat{\theta}^{\text{cf}}_{\text{NIE}} - \theta_{\text{NIE}})$ and $B_n = \sqrt{n}(\theta_{\text{NIE}} - \theta^{\text{pop}}_{\text{NIE}})$. Since $\theta_{\text{NIE}} = \bar{X}^\top \Lambda_{\text{NIE}}$ and $\theta^{\text{pop}}_{\text{NIE}} = \mu_X^\top \Lambda_{\text{NIE}}$, we can write

$$B_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \mu_X)^\top \Lambda_{\text{NIE}}\,.$$

Furthermore, it is immediate from the definition that B_n is measurable with respect to $\mathcal{X}_n$. A similar argument, used to prove Lemma B.2 can be used to establish that for any $u \in \mathbb{R}$,

$$\mathbb{E}\left[\exp\left\{iu\frac{A_n}{\sigma^{\text{cf,het}}_{n,\text{NIE}}}\right\} \mid \mathcal{X}_n\right] \xrightarrow{p} e^{-u^2/2}.$$

Consequently, by the same conditional characteristic-function calculation as in the proof of Corollary 5.11, we have, for any $u, v \in \mathbb{R}$,

$$\left|\mathbb{E}\left[\exp\left\{iu\frac{A_n}{\sigma^{\text{cf,het}}_{n,\text{NIE}}} + ivB_n\right\}\right] - e^{-u^2/2}\mathbb{E}[e^{ivB_n}]\right| \longrightarrow 0\,. \tag{B.19}$$

Now, as $\omega^{\text{het}}_{n,\text{NIE}}$ admits an asymptotically deterministic equivalent, there exists a deterministic sequence $\{\bar{\omega}^{\text{het}}_{n,\text{NIE}}\}_{n\in\mathbb{N}}$ with $\bar{\omega}^{\text{het}}_{n,\text{NIE}} \in [0,1]$ such that $\omega^{\text{het}}_{n,\text{NIE}} - \bar{\omega}^{\text{het}}_{n,\text{NIE}} \to 0$ in probability. Fix an arbitrary subsequence. Since $\bar{\omega}^{\text{het}}_{n,\text{NIE}} \in [0,1]$ and $\sigma^2_{X,\text{NIE}} \le C$, there exists a further subsequence, which we continue to index by n, such that

$$\bar{\omega}^{\text{het}}_{n,\text{NIE}} \to \omega \in [0,1], \qquad \sigma^2_{X,\text{NIE}} \to \zeta^2$$

for some $\zeta \ge 0$. Consequently, along this subsequence, $\omega^{\text{het}}_{n,\text{NIE}} \to \omega$ in probability. First, assume $\zeta > 0$. By Lindeberg-Feller central limit theorem (as X_i's are sub-Gaussian and $\|\Lambda_{\text{NIE}}\|_2 \le C_{\text{NIE}}$ is bounded (Assumption 5.4),

$$\frac{B_n}{\sigma_{X,\text{NIE}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Combining this with Equation (B.19) yields:

$$\left(\frac{A_n}{\sigma^{\text{cf,het}}_{n,\text{NIE}}}, \frac{B_n}{\sigma_{X,\text{NIE}}}\right) \xrightarrow{d} (Z_1, Z_2), \quad Z_1, Z_2 \overset{\text{iid}}{\sim} \mathcal{N}(0,1)\,.$$

Define $D^2_{n,\text{NIE}} = (\sigma^{\text{cf,het}}_{n,\text{NIE}})^2 + \sigma^2_{X,\text{NIE}}$. Then, by Slutsky's theorem,

$$\frac{A_n+B_n}{D_{n,\text{NIE}}} = \sqrt{\omega^{\text{het}}_{n,\text{NIE}}}\frac{A_n}{\sigma^{\text{cf,het}}_{n,\text{NIE}}} + \sqrt{1-\omega^{\text{het}}_{n,\text{NIE}}}\frac{B_n}{\sigma_{X,\text{NIE}}} \xrightarrow{d} \sqrt{\omega}\,Z_1 + \sqrt{1-\omega}\,Z_2 \sim \mathcal{N}(0,1).$$

Now, if $\zeta = 0$, then it is immediate that $\mathbb{E}[B_n^2] = \sigma^2_{X,\text{NIE}} \to 0$ and consequently $B_n = o_p(1)$. Moreover, we have already established that $(\sigma^{\text{cf,het}}_{n,\text{NIE}})^2$ is bounded away from zero with probability going to one. Consequently, $D_{n,\text{NIE}}/\sigma^{\text{cf,het}}_{n,\text{NIE}} \to 1$ in probability and $B_n/D_{n,\text{NIE}} = o_p(1)$. Hence,

$$\frac{A_n+B_n}{D_{n,\text{NIE}}} = \frac{A_n}{\sigma^{\text{cf,het}}_{n,\text{NIE}}} + o_p(1) \xrightarrow{d} \mathcal{N}(0,1)\,.$$

Thus every subsequence contains a further subsequence along which

$$\frac{A_n+B_n}{\sqrt{(\sigma^{\text{cf,het}}_{n,\text{NIE}})^2 + \sigma^2_{X,\text{NIE}}}} \xrightarrow{d} \mathcal{N}(0,1)\,,$$

and the limit is same for any subsequence. Therefore, the entire sequence converges to $\mathcal{N}(0,1)$. This completes the proof. □

C Proof of Additional Lemmas

C.1 A Lemma on Sample Restricted Eigenvalue Conditions

Lemma C.1. *Let $s_{X,n} = (1+s)k, \quad s_{W,n} = k$. Under Assumptions 5.3 and 5.4, suppose*

$$s_{X,n} \log\left(\frac{ep}{r_{X,n}}\right) = o(n), \qquad s_{W,n} \log\left(\frac{e(p+q)}{r_{W,n}}\right) = o(n).$$

Then, for any fixed cone constant $L < \infty$, there exist constants $c_{X,\mathrm{RE}}, c_{W,\mathrm{RE}} > 0$, independent of (n,p,q), such that, with probability going to one, simultaneously over both arms and both data folds,

$$\inf_{\substack{|S| \le s_{X,n} \\ \delta \neq 0,\ \|\delta_{S^c}\|_1 \le L\|\delta_S\|_1}} \frac{\delta^\top \left(\mathbf{X}_{a,j}^\top \mathbf{X}_{a,j}/n_{a,j}\right)\delta}{\|\delta\|_2^2} \ge c_{X,\mathrm{RE}},$$

and

$$\inf_{\substack{|S| \le s_{W,n} \\ \delta \neq 0,\ \|\delta_{S^c}\|_1 \le L\|\delta_S\|_1}} \frac{\delta^\top \left(\mathbf{W}_{a,j}^\top \mathbf{W}_{a,j}/n_{a,j}\right)\delta}{\|\delta\|_2^2} \ge c_{W,\mathrm{RE}}.$$

Consequently, the corresponding compatibility constants are bounded away from zero with probability tending to one.

Proof. As the proof is the same for the first and the second conclusion of the lemma, we present the proof in a unified way. Let V denote either X or W. We will use the notation $d = \dim(V)$; when $V = X$ then $d = p$ and when $V = W$ then $d = p+q$. Fix $a \in \{0,1\}$ and let $\mu_{V,a} = \mathbb{E}[V \mid A = a]$ and $Q_{V,a} = \mathbb{E}[VV^\top \mid A = a]$ be the conditional mean and second moment of V given $A = a$. Now writing $\Sigma_{V,a} = \mathsf{var}(V \mid A = a)$, we have $Q_{V,a} = \Sigma_{V,a} + \mu_{V,a}\mu_{V,a}^\top$. By Assumptions 5.3 and 5.4, along with Remark 5.6, there exist constants $0 < c_V < C_V < \infty$ (independent of (n,p,q)) such that:

$$\lambda_{\min}(Q_{V,a}) \ge c_V, \qquad \max_{1 \le \ell \le d}(Q_{V,a})_{\ell\ell} \le C_V\,.$$

Therefore, if we define $Z_{V,a} = Q_{V,a}^{-1/2} V$, then we have $\mathbb{E}[Z_{V,a} Z_{V,a}^\top \mid A = a] = \mathbf{I}_d$. Furthermore, conditional on $A = a$, $Z_{V,a}$ is sub-Gaussian random vector as

$$\|u^\top Z_{V,a}\|_{\psi_2 | A=a} \le \|u^\top Q_{V,a}^{-1/2}(V - \mu_{V,a})\|_{\psi_2|A=a} + C|u^\top Q_{V,a}^{-1/2}\mu_{V,a}| \le C\|u\|_2\,.$$

Here, we have used the fact that $\lambda_{\min}(Q_{V,a}) \ge c_V$ and $\|Q_{V,a}^{-1/2}\mu_{V,a}\|_2^2 \le 1$ which follows from the fact that $\mu_{V,a}\mu_{V,a}^\top \preceq Q_{V,a}$. Let $I_{a,j} := \{i : i \in \mathcal{D}_j,\ A_i = a\}$ denotes that set of all observations with $A_i = a$ in the j^{th} fold and let $\mathcal{I}_n = \sigma(A_{1:n})$, the sigma-field generated by the treatment indicators. It is immediate that $n_{a,j} = |I_{a,j}|$ and conditionally on $\mathcal{I}_n$, $n_{a,j}$ is fixed and the rows of $\mathbf{V}_{a,j}$ are i.i.d. with the common distribution $V \mid A = a$. Hence, we have:

$$\mathbb{E}\left[\frac{\mathbf{V}_{a,j}^\top \mathbf{V}_{a,j}}{n_{a,j}} \mid \mathcal{I}_n\right] = \mathbb{E}[VV^\top \mid A = a] = Q_{V,a}.$$

Therefore, if we write $\mathbf{V}_{a,j} = \Psi_{V,a,j} Q_{V,a}^{1/2}$, then, conditionally on $\mathcal{I}_n$, where the rows of $\Psi_{V,a,j}$ are independent along with second moment (conditionally on $A = a$) being $\mathbf{I}_d$. Therefore, an application of Theorem 15 of Rudelson and Zhou (2012) with $A = Q_{V,a}^{1/2}$, sparsity level $s_{V,n}$ implies:

$$n_{a,j} \gtrsim s_{V,n} \log\left(\tfrac{ed}{s_{V,n}}\right) \implies \tfrac{1}{n_{a,j}}\|\mathbf{V}_{a,j}\delta\|_2^2 \ge c_{V,\mathrm{RE}}\|\delta\|_2^2$$

simultaneously for all sets S with $|S| \le s_{V,n}$ and all $\delta \neq 0$ satisfying $\|\delta_{S^c}\|_1 \le L\|\delta_S\|_1$. However, to apply this theorem, we need to show $Q_{V,a}^{1/2}$ satisfies the RE condition. This immediately follows from our assumptions; since $Q_{V,a} = \Sigma_{V,a} + \mu_{V,a}\mu_{V,a}^\top \succeq \Sigma_{V,a} \succeq c_V \mathbf{I}_d$. we have, for every $\delta \in \mathbb{R}^d$,

$$\|Q_{V,a}^{1/2}\delta\|_2 \ge \sqrt{c_V}\|\delta\|_2.$$

Therefore, for every S with $|S| \le s_{V,n}$ and every δ satisfying $\|\delta_{S^c}\|_1 \le 3L\|\delta_S\|_1$, we have $\|Q_{V,a}^{1/2}\delta\|_2 \ge \sqrt{c_V}\|\delta_S\|_2$. Hence $Q_{V,a}^{1/2}$ satisfies $\mathrm{RE}(s_{V,n}, 3L)$, with $K(s_{V,n}, 3L, Q_{V,a}^{1/2}) \le c_V^{-1/2}$. Finally, for any δ in the corresponding cone, we have $\|\delta_S\|_1^2 \le |S|\|\delta_S\|_2^2 \le s_{V,n}\|\delta\|_2^2$. Therefore,

$$\frac{s_{V,n}\delta^\top(\mathbf{V}_{a,j}^\top \mathbf{V}_{a,j}/n_{a,j})\delta}{\|\delta_S\|_1^2} \ge c_{V,\mathrm{RE}},$$

so the corresponding compatibility constants are also bounded away from zero with probability going to one. □

C.2 Proof of Lemma B.1

Proof. We only prove the case for $j = 2$ (as it is exactly stated in Algorithm 1), since the case for $j = 1$ is analogous. For notational simplicity, we prove the result for $j = 2$ and suppress the fold indices throughout the proof. Define

$$a := a^{(1)} = \begin{pmatrix} \bar{X} \\ \hat{m}_c \end{pmatrix}, \qquad a^0 = \begin{pmatrix} \bar{X} \\ B_0\bar{X} \end{pmatrix}, \qquad a^* := a_{10}^* = \begin{pmatrix} \mu_X \\ B_0\mu_X \end{pmatrix},$$

where $\hat{m}_c = \hat{m}_{c,1}$. Thus, a is the actual random contrast used in Algorithm 1, a^0 is its oracle sample counterpart obtained by replacing $\hat{m}_c$ with $B_0\bar{X}$, and a^* is the corresponding population contrast. We also write $\eta^* := \eta_{10}^* = \Omega_{W,1}^{-1}a^*$. The optimization problem in Equation (3.2) can be cleanly written as:

$$\begin{aligned} \min_{\tau\in\mathbb{R}^{n_t}} \quad & \frac{1}{2}\|\tau\|_2^2 \\ \text{subject to} \quad & \left\|a - \mathbf{W}_t^\top\tau\right\|_\infty \le \check{\lambda}_n := K_1\sqrt{\frac{\log(p+q)}{n_t}}, \\ & \|\tau\|_\infty \le b_n := n_t^{-\alpha}, \end{aligned} \tag{C.1}$$

and $\hat{\tau}(a)$ is the unique optimizer of the above optimization problem (it exists and is unique as the objective is strongly convex over a convex compact set). For any $v \in \mathbb{R}^{p+q}$, let $\hat{\tau}(v)$ denote the unique optimizer of Equation (C.1) with a replaced by v. Therefore, it is immediate that $\hat{\tau}(a) = \hat{\tau}_2^{(1)}$ and $\hat{\tau}(a^0) = \hat{\tau}_2^{(1),0}$. The Lagrangian of the above optimization problem can be written as, given non-negative vectors (co-ordinatewise) $\{\eta_{i,+}\}_{i=1}^{p+q}, \{\eta_{i,-}\}_{i=1}^{p+q}$ and $\{\zeta_{i,+}\}_{i=1}^{n_t}, \{\zeta_{i,-}\}_{i=1}^{n_t}$:

$$\begin{aligned} &\mathbb{L}\left(\tau, \{\eta_{i,+}\}, \{\eta_{i,-}\}, \{\zeta_{i,+}\}, \{\zeta_{i,-}\}\right) \\ &= \frac{1}{2}\|\tau\|_2^2 + \sum_{i=1}^{p+q}\eta_{i,+}\left((a - \mathbf{W}_t^\top\tau)_i - \check{\lambda}_{n_t}\right) + \sum_{i=1}^{p+q}\eta_{i,-}\left((\mathbf{W}_t^\top\tau - a)_i - \check{\lambda}_{n_t}\right) \\ &\qquad + \sum_{i=1}^{n_t}\zeta_{i,+}\left(\tau_i - b_n\right) + \sum_{i=1}^{n_t}\zeta_{i,-}\left(-\tau_i - b_n\right) \\ &= \frac{1}{2}\|\tau\|_2^2 + \sum_{i=1}^{p+q}(\eta_{i,+} - \eta_{i,-})(a - \mathbf{W}_t^\top\tau)_i - \check{\lambda}_{n_t}\sum_{i=1}^{p+q}(\eta_{i,+} + \eta_{i,-}) \\ &\qquad + \sum_{i=1}^{n_t}(\zeta_{i,+} - \zeta_{i,-})\tau_i - b_n\sum_{i=1}^{n_t}(\zeta_{i,+} + \zeta_{i,-}) \end{aligned}$$

Define the signed multipliers $\eta := \eta_+ - \eta_-$, and $\zeta := \zeta_+ - \zeta_-$. For any fixed signed vector η, every nonnegative decomposition $\eta = \eta_+ - \eta_-$ satisfies, coordinate-wise, $\eta_{j,+} + \eta_{j,-} \ge |\eta_j|$, with equality for the canonical decomposition $\eta_{j,+} = (\eta_j)_+$ and $\eta_{j,-} = (-\eta_j)_+$. Since the dual objective contains $-\check{\lambda}_{n_t}\mathbf{1}^\top(\eta_+ + \eta_-)$, while all its remaining terms depend on (η_+, η_-) only through $\eta_+ - \eta_-$, any overlap between η_+ and η_- can be removed without changing the signed multiplier and while weakly increasing the dual objective. Hence,

without loss of generality, $\mathbf{1}^\top(\eta_+ + \eta_-) = \|\eta\|_1$. The same argument gives $\mathbf{1}^\top(\zeta_+ + \zeta_-) = \|\zeta\|_1$. Consequently, the Lagrangian may be written in terms of the signed multipliers as

$$\mathbb{L}(\tau, \{\eta_i\}, \{\zeta_i\}) = \frac{1}{2}\|\tau\|_2^2 + \eta^\top(a - \mathbf{W}_t^\top \tau) + \zeta^\top \tau - \check{\lambda}_{n_t}\|\eta\|_1 - b_n\|\zeta\|_1 \,. \tag{C.2}$$

We will also use this reduced form in our analysis. Henceforth, we will use the notation $\hat{\eta}^{\mathrm{KKT}}(a), \hat{\zeta}^{\mathrm{KKT}}(a)$ to denote the dual optimal parameters of this optimization problem. From the complementary slackness of the KKT condition, we have:

$$\hat{\eta}^{\mathrm{KKT}}_{j,+}(a)\left\{(a - \mathbf{W}_t^\top \hat{\tau}(a))_j - \check{\lambda}_{n_t}\right\} = 0,$$
$$\hat{\eta}^{\mathrm{KKT}}_{j,-}(a)\left\{(\mathbf{W}_t^\top \hat{\tau}(a) - a)_j - \check{\lambda}_{n_t}\right\} = 0.$$

Similarly, for every $i = 1, \ldots, n_t$,

$$\hat{\zeta}^{\mathrm{KKT}}_{i,+}(a)\left\{\hat{\tau}_i(a) - b_n\right\} = 0, \qquad \hat{\zeta}^{\mathrm{KKT}}_{i,-}(a)\left\{-\hat{\tau}_i(a) - b_n\right\} = 0.$$

Let us first focus on $\hat{\eta}^{\mathrm{KKT}}$. From the above condition, it is immediate that when $|(a - \mathbf{W}_t^\top \hat{\tau}(a))_j| \neq \check{\lambda}_{n_t}$, then $\hat{\eta}^{\mathrm{KKT}}_{j,+}(a) = \hat{\eta}^{\mathrm{KKT}}_{j,-}(a) = 0$ and hence $\hat{\eta}^{\mathrm{KKT}}_j(a) = 0$. On the other hand when $(a - \mathbf{W}_t^\top \hat{\tau}(a))_j = \check{\lambda}_{n_t}$, then $\hat{\eta}^{\mathrm{KKT}}_{j,-}(a) = 0$ and consequently $\eta^{\mathrm{KKT}}_j \geq 0$. Similarly, when $(a - \mathbf{W}_t^\top \hat{\tau}(a))_j = -\check{\lambda}_{n_t}$, then $\hat{\eta}^{\mathrm{KKT}}_{j,+}(a) = 0$ and consequently $\eta^{\mathrm{KKT}}_j \leq 0$. Therefore, $\hat{\eta}^{\mathrm{KKT}}(a) \in \mathcal{N}_{\mathcal{B}_\infty(\check{\lambda}_{n_t})}(a - \mathbf{W}_t^\top \hat{\tau}(a))$, where $\mathcal{N}_{\mathcal{A}}(x)$ is the normal cone of $\mathcal{A}$ at x defined as:

$$\mathcal{N}_{\mathcal{A}}(x) = \{g : \langle g, y - x\rangle \leq 0, \ \forall\ y \in \mathcal{A}\} \,. \tag{C.3}$$

By a similar logic, we also conclude that $\hat{\zeta}^{\mathrm{KKT}}(a)_i = 0$ for $|\hat{\tau}(a)_i| \neq b_n$, $\hat{\zeta}^{\mathrm{KKT}}(a)_i \geq 0$ for $\hat{\tau}(a)_i = b_n$ and $\hat{\zeta}^{\mathrm{KKT}}_i(a) \leq 0$ for $\hat{\tau}(a)_i = -b_n$. Therefore, $\hat{\zeta}^{\mathrm{KKT}}(a) \in \mathcal{N}_{\mathcal{B}_\infty(b_n)}(\hat{\tau}(a))$. The following lemma relates the difference $\hat{\tau}(a) - \hat{\tau}(a^0)$ to that of $(a - a^0)$ and the dual variable $(\hat{\eta}^{\mathrm{KKT}}(a) - \hat{\eta}^{\mathrm{KKT}}(a^0))$:

Lemma C.2. *The primal optimizer $(\hat{\tau}(a), \hat{\tau}(a^0))$ and the dual optimizer $(\hat{\eta}^{\mathrm{KKT}}(a) - \hat{\eta}^{\mathrm{KKT}}(a^0))$ satisfies:*

$$\|\hat{\tau}(a) - \hat{\tau}(a^0)\|_2^2 \leq \|a - a^0\|_\infty \|\hat{\eta}^{\mathrm{KKT}}(a) - \hat{\eta}^{\mathrm{KKT}}(a^0)\|_1 \,.$$

Proof. The key ingredient of the proof of this lemma is a simple fact about the normal cone: let $g \in \mathcal{N}_{\mathcal{A}}(x)$ and $g' \in \mathcal{N}_{\mathcal{A}}(x')$. Then $\langle g - g', x - x'\rangle \geq 0$. The proof is simple: taking $y = x'$ in the definition of $g \in N_A(x)$ and $y = x$ in the definition of $g' \in N_A(x')$ gives $\langle g, x' - x\rangle \leq 0$ and $\langle g', x - x'\rangle \leq 0$, which together imply $\langle g - g', x - x'\rangle \geq 0$. From the first KKT condition, we have, from Equation (C.2):

$$\hat{\tau}(a) - \mathbf{W}_t\hat{\eta}^{\mathrm{KKT}}(a) + \hat{\zeta}^{\mathrm{KKT}}(a) = 0 \implies \hat{\tau}(a) = \mathbf{W}_t\hat{\eta}^{\mathrm{KKT}}(a) - \hat{\zeta}^{\mathrm{KKT}}(a) \,. \tag{C.4}$$

For notational simplicity, define $\Delta\hat{\tau} = \hat{\tau}(a) - \hat{\tau}(a^0)$ (and analogously define $\Delta\hat{\eta}$ and $\Delta\hat{\zeta}$). Then, from the above equation, we have:

$$\Delta\hat{\tau} = \mathbf{W}_t\Delta\hat{\eta} - \Delta\hat{\zeta} \implies \|\Delta\hat{\tau}\|_2^2 = (\mathbf{W}_t^\top\Delta\hat{\tau})^\top\Delta\hat{\eta} - \Delta\hat{\tau}^\top\Delta\hat{\zeta} \,.$$

Now, $\hat{\zeta}(a) \in \mathcal{N}_{\mathcal{B}_\infty(b_n)}(\hat{\tau}(a))$ and $\hat{\zeta}(a^0) \in \mathcal{N}_{\mathcal{B}_\infty(b_n)}(\hat{\tau}(a^0))$. Therefore, from the property of the normal cone described above, we have $\Delta\hat{\tau}^\top\Delta\hat{\zeta} \geq 0$, and consequently $\|\Delta\hat{\tau}\|_2^2 \leq (\mathbf{W}_t^\top\Delta\hat{\tau})^\top\Delta\hat{\eta}$. Now for the next part, we rewrite:

$$\mathbf{W}_t^\top\Delta\hat{\tau} = (\mathbf{W}_t^\top\hat{\tau}(a) - a) - (\mathbf{W}_t^\top\hat{\tau}(a^0) - a^0) + (a - a^0) \,.$$

Hence,

$$(\mathbf{W}_t^\top\Delta\hat{\tau})^\top\Delta\hat{\eta} = \left((\mathbf{W}_t^\top\hat{\tau}(a) - a) - (\mathbf{W}_t^\top\hat{\tau}(a^0) - a^0)\right)^\top\Delta\hat{\eta} + (a - a^0)^\top\Delta\hat{\eta}$$

The first summand on the right-hand side is again negative from the property of the normal cone, and consequently:

$$\|\Delta\hat{\tau}\|_2^2 \leq (\mathbf{W}_t^\top\Delta\hat{\tau})^\top\Delta\hat{\eta} \leq (a - a^0)^\top\Delta\hat{\eta} \leq \|a - a^0\|_\infty\|\hat{\eta}^{\mathrm{KKT}}(a) - \hat{\eta}^{\mathrm{KKT}}(a^0)\|_1$$

where the last inequality follows from the Hölder inequality. This completes the proof. □

In the rest of the analysis, we will establish that $n\|\hat{\eta}^{\mathrm{KKT}}(a)\|_1 \vee n\|\hat{\eta}^{\mathrm{KKT}}(a^0)\|_1 = O_p(1)$. This, along with the fact that $\|a - a^0\|_\infty = o_p(1)$ establishes the claim. Towards that end, we will start with the reduced form Lagrangian in Equation (C.2). Now for any fixed (η, ζ), $\mathbb{L}(\tau, \{\eta_i\}, \{\zeta_i\})$ is minimized (with respect to τ) at $\tau^*(\eta, \zeta) = \mathbf{W}_t\eta - \zeta$. As a consequence:

$$\begin{aligned}\min_\tau \mathbb{L}(\tau, \{\eta_i\}, \{\zeta_i\}) &= \frac{1}{2}\|\mathbf{W}_t\eta - \zeta\|_2^2 + \eta^\top a - \|\mathbf{W}_t\eta - \zeta\|_2^2 - \check{\lambda}_{n_t}\|\eta\|_1 - b_n\|\zeta\|_1 \\ &= -\left\{\frac{1}{2}\|\mathbf{W}_t\eta - \zeta\|_2^2 - \eta^\top a + \check{\lambda}_{n_t}\|\eta\|_1 + b_n\|\zeta\|_1\right\}.\end{aligned}$$

Therefore, the dual optimization problem is:

$$\begin{aligned}&\max_{\eta,\zeta} \; -\left\{\frac{1}{2}\|\mathbf{W}_t\eta - \zeta\|_2^2 - \eta^\top a + \check{\lambda}_{n_t}\|\eta\|_1 + b_n\|\zeta\|_1\right\} \\ =& -\min_{\eta,\zeta} \; \left\{\frac{1}{2}\|\mathbf{W}_t\eta - \zeta\|_2^2 - \eta^\top a + \check{\lambda}_{n_t}\|\eta\|_1 + b_n\|\zeta\|_1\right\}\end{aligned}$$

Now let us re-parametrize $\eta_n = n_t\eta$ and $\zeta_n = n_t\zeta$. With this reparametrization, all we need to show is $\|\hat{\eta}(a)\|_1 \vee \|\hat{\eta}(a^0)\|_1 = O_p(1)$. Then, we can rewrite the dual problem with this new definition of (η, ζ) as:

$$-\min_{\eta,\zeta} \; \left\{\tfrac{1}{2n_t}\|\mathbf{W}_t\eta - \zeta\|_2^2 - \eta^\top a + \check{\lambda}_{n_t}\|\eta\|_1 + b_n\|\zeta\|_1\right\} \tag{C.5}$$

Now, for a fixed η, the optimization over ζ is simple, as the objective function is separable over coordinates:

$$\min_\zeta \; \left\{\tfrac{1}{2n_t}\|\mathbf{W}_t\eta - \zeta\|_2^2 + b_n\|\zeta\|_1\right\} = \min_\zeta \; \left\{\sum_{j=1}^{n_t}\left(\tfrac{1}{2n_t}((\mathbf{W}_t\eta)_i - \zeta_i)^2 + b_n|\zeta_i|\right)\right\}$$

It is well-known that for each coordinate the objective is the proximal operator of ℓ_1 norm and hence the minimizer $\zeta_i^* = \mathsf{sign}((\mathbf{W}_t\eta)_i)(|(\mathbf{W}_t\eta)_i| - n_t b_n)_+$ and the minimum value is the Huber function:

$$\min_{\zeta_i} \frac{1}{2n_t}((\mathbf{W}_t\eta)_i - \zeta_i)^2 + b_n|\zeta_i| = \begin{cases}\frac{(\mathbf{W}_t\eta)_i^2}{2n_t}, & \text{if } |(\mathbf{W}_t\eta)_i| \le n_t b_n \\ b_n|(\mathbf{W}_t\eta)_i| - \frac{n_t b_n^2}{2}, & \text{if } |(\mathbf{W}_t\eta)_i| > n_t b_n\end{cases}$$

Therefore, if we define Huber function $H_\lambda(x)$ as $x^2/2$ when $|x| \le \lambda$ and $\lambda|x| - \lambda^2/2$ for $|x| > \lambda$, then we have:

$$\begin{aligned}&\min_{\eta,\zeta} \; \left\{\frac{1}{2n_t}\|\mathbf{W}_t\eta - \zeta\|_2^2 - \eta^\top a + \check{\lambda}_{n_t}\|\eta\|_1 + b_n\|\zeta\|_1\right\} \\ &\qquad = \min_\eta \left\{\frac{1}{n_t}\sum_{i=1}^{n_t} H_{n_t b_n}\left((\mathbf{W}_t\eta)_i\right) - \eta^\top a + \check{\lambda}_{n_t}\|\eta\|_1\right\}\end{aligned}$$

Therefore, we have:

$$\begin{aligned}\hat{\eta}_n(a) &= \arg\min_\eta \left\{\frac{1}{n_t}\sum_{i=1}^{n_t} H_{n_t b_n}\left((\mathbf{W}_t\eta)_i\right) - \eta^\top a + \check{\lambda}_{n_t}\|\eta\|_1\right\} \\ &:= \arg\min_\eta \; \left\{\mathcal{L}_{n,a}(\eta) + \check{\lambda}_{n_t}\|\eta\|_1\right\}.\end{aligned} \tag{C.6}$$

Furthemore, we have $\hat{\eta}_n(a) = n_t\hat{\eta}^{\mathrm{KKT}}(a)$. A similar relation holds for a^0 as well. By Assumption 5.4, we have $\|\eta^*\|_1 \le C_\eta$. We next bound $\|a^0 - a^*\|_\infty$ and $\|a - a^*\|_\infty$:

$$\begin{aligned}\|a^0 - a^*\|_\infty &\le \max\left\{\|\bar{X} - \mu_X\|_\infty, \|\mathbf{B}_0(\bar{X} - \mu_X)\|_\infty\right\} \\ &\le (1 + \|\mathbf{B}_0\|_{1,\infty})\|\bar{X} - \mu_X\|_\infty = O_p\left(\sqrt{\frac{\log d}{n}}\right) \\ \|a - a^*\|_\infty &\le (1 + \|\mathbf{B}_0\|_{1,\infty})\|\bar{X} - \mu_X\|_\infty + \|\hat{m}_c - \mathbf{B}_0\bar{X}\|_\infty\end{aligned} \tag{C.7}$$

$$= O_p\left(\sqrt{\frac{\log d}{n}}\right). \tag{C.8}$$

Let $\hat{\Omega}_{W,1} = (\mathbf{W}_t^\top \mathbf{W}_t)/n_t$. Since the rows of $\mathbf{W}_t$ are uniformly sub-Gaussian with bounded coordinate means, we have $\|\hat{\Omega}_{W,1} - \Omega_{W,1}\|_{\max} = O_p(\sqrt{\log(p+q)/n_t})$. Consequently, simultaneously for $v \in \{a, a^0\}$,

$$\|\hat{\Omega}_{W,1}\eta^* - v\|_\infty \le \|\hat{\Omega}_{W,1} - \Omega_{W,1}\|_{\max}\|\eta^*\|_1 + \|a^* - v\|_\infty \le \frac{\check{\lambda}_{n_t}}{2} \tag{C.9}$$

with probability tending to one, after choosing the fixed constant K_1 is sufficiently large. Furthermore, from the sub-Gaussianity of W_i (Assumption 5.3 and Remark 5.6), we also have

$$\max_{1\le i\le n_t} |W_i^\top \eta^*| \le \|\eta^*\|_1 \max_{1\le i\le n_t} \|W_i\|_\infty = O_p\left(\sqrt{\log(nd)}\right).$$

Since $n_t b_n = n_t^{1-\alpha} \gg \sqrt{\log(nd)}$, it follows that, with probability going to one, $\max_{1\le i\le n_t} |W_i^\top \eta^*| < n_t b_n$ and consequently the Huber loss is quadratic at η^* and, for $v \in \{a, a^0\}$, $\nabla \mathcal{L}_{n_t,v}(\eta^*) = \hat{\Omega}_{W,1}\eta^* - v$. The same calculations also show that following the oracle balancing weight $\tau_i^\circ = (W_i^\top \eta^*)/n_t$ for $1 \le i \le n_t$, satisfies:

$$\|v - \mathbf{W}_t^\top \tau^\circ\|_\infty \le \check{\lambda}_{n_t}/2 < \check{\lambda}_{n_t}, \qquad \|\tau^\circ\|_\infty = O_p\left(\frac{\sqrt{\log(nd)}}{n}\right) = o_p(b_n),$$

Now, for $v \in \{a, a^0\}$, let $\hat{\eta}_n(v)$ denote the rescaled Huber-dual minimizer:

$$\hat{\eta}_n(v) \in \arg\min_{\eta\in\mathbb{R}^d} \left\{\mathcal{L}_{n,v}(\eta) + \check{\lambda}_{n_t}\|\eta\|_1\right\},$$

Optimality of $\hat{\eta}_n(v)$ and convexity of $\mathcal{L}_{n,v}$ imply

$$\begin{aligned}\check{\lambda}_{n_t}\|\hat{\eta}_n(v)\|_1 &\le \|\nabla \mathcal{L}_{n,v}(\eta^*)\|_\infty \|\hat{\eta}_n(v) - \eta^*\|_1 + \check{\lambda}_{n_t}\|\eta^*\|_1 \\ &\le \frac{\check{\lambda}_{n_t}}{2}\left(\|\hat{\eta}_n(v)\|_1 + \|\eta^*\|_1\right) + \check{\lambda}_{n_t}\|\eta^*\|_1.\end{aligned}$$

Therefore,

$$\frac{\check{\lambda}_{n_t}}{2}\|\hat{\eta}_n(v)\|_1 \le \frac{3\check{\lambda}_{n_t}}{2}\|\eta^*\|_1 \implies \|\hat{\eta}_n(v)\|_1 \le 3\|\eta^*\|_1 \le 3C_\eta. \tag{C.10}$$

Applying this argument to both $\hat{\eta}_n(a)$ and $\hat{\eta}_n(a^0)$, we obtain

$$\|\hat{\eta}_n(a) - \hat{\eta}_n(a^0)\|_1 \le 6C_\eta \implies \|\hat{\eta}^{\mathrm{KKT}}(a) - \hat{\eta}^{\mathrm{KKT}}(a^0)\|_1 \le \frac{6C_\eta}{n_t}.$$

Therefore, by Lemma C.2,

$$\begin{aligned}\|\hat{\tau}(a) - \hat{\tau}(a^0)\|_2^2 \le \|a - a^0\|_\infty \left\|\hat{\eta}^{\mathrm{KKT}}(a) - \hat{\eta}^{\mathrm{KKT}}(a^0)\right\|_1 &\le \frac{6C_\eta}{n_t}\|a - a^0\|_\infty \\ &= \frac{6C_\eta}{n_t}\|\hat{m}_c - B_0\bar{X}\|_\infty.\end{aligned}$$

Consequently, since $n \asymp n_t$,

$$n\|\hat{\tau}(a) - \hat{\tau}(a^0)\|_2^2 \le \frac{6C_\eta n}{n_t}\|\hat{m}_c - B_0\bar{X}\|_\infty = O_p\left(\sqrt{\frac{\log p}{n}} + \sqrt{\frac{\log q}{n}}\right) = o_p(1).$$

Hence, $\sqrt{n}\|\hat{\tau}(a) - \hat{\tau}(a^0)\|_2 = o_p(1)$, which completes the proof. □

C.3 Proof of Lemma B.2

Proof. Recall from Proposition 4.1 that we have $\hat{\theta}_{10}^{\mathrm{cf}} - \theta_{10} = \Delta_n^{\mathrm{cf}} + V_n^{\mathrm{cf}}$, where

$$V_n^{\mathrm{cf}} = \frac{1}{2}\left[\sum_{\substack{i\in\mathcal{D}_2\\A_i=1}} \hat{\tau}_{1,i}^{(2)}\epsilon_i + \sum_{\substack{i\in\mathcal{D}_1\\A_i=0}} \hat{\tau}_{2,i}^{(1)}U_i^\top\gamma_1 + \sum_{\substack{i\in\mathcal{D}_1\\A_i=1}} \hat{\tau}_{1,i}^{(1)}\epsilon_i + \sum_{\substack{i\in\mathcal{D}_2\\A_i=0}} \hat{\tau}_{2,i}^{(2)}U_i^\top\gamma_1\right].$$

For $j \in \{1,2\}$, let $\hat{\tau}_1^{(j),0}$ denote the oracle first balancing weight obtained by replacing the random mediator target $\hat{m}_{c,j}$ by $\mathbf{B}_0\bar{X}$ (same as in the proof of Theorem 5.9), and correspondingly define an oracle version of V_n^{cf} as:

$$V_n^{\mathrm{cf},0} = \tfrac{1}{2}\left[\sum_{\substack{i\in\mathcal{D}_2\\A_i=1}} \hat{\tau}_{1,i}^{(2),0}\epsilon_i + \sum_{\substack{i\in\mathcal{D}_1\\A_i=0}} \hat{\tau}_{2,i}^{(1)}U_i^\top\gamma_1 + \sum_{\substack{i\in\mathcal{D}_1\\A_i=1}} \hat{\tau}_{1,i}^{(1),0}\epsilon_i + \sum_{\substack{i\in\mathcal{D}_2\\A_i=0}} \hat{\tau}_{2,i}^{(2)}U_i^\top\gamma_1\right].$$

along with its corresponding variance scale as $(\sigma_n^{\mathrm{cf},0})^2$ defined in Equation (B.7). Furthermore, for notational simplicity, let us define $L_n^{\mathrm{cf},0} = \sqrt{n}V_n^{\mathrm{cf},0}/\sigma_n^{\mathrm{cf},0}$ and $T_n^{\mathrm{cf}} = \sqrt{n}(\hat{\theta}_{10}^{\mathrm{cf}} - \theta_{10})/\sigma_n^{\mathrm{cf}}$. We first establish that

$$T_n^{\mathrm{cf}} - L_n^{\mathrm{cf},0} = o_p(1). \tag{C.11}$$

Towards that end, define $\mathcal{G}_n = \sigma(A_{1:n}, X_{1:n}, M_{1:n})$. Conditional on $\mathcal{G}_n$, both the actual and oracle first balancing weights are fixed, whereas the treated outcome errors remain centered and independent. Therefore,

$$\mathbb{E}\left[\left(\sqrt{n}\left(V_n^{\mathrm{cf}} - V_n^{\mathrm{cf},0}\right)\right)^2 \,\middle|\, \mathcal{G}_n\right] = \tfrac{n\sigma_1^2}{4}\sum_{j=1}^2 \left\|\hat{\tau}_1^{(j)} - \hat{\tau}_1^{(j),0}\right\|_2^2 = o_p(1),$$

where the last conclusion follows from Lemma B.1. Consequently, by a conditional version of Markov's inequality, $\sqrt{n}\left(V_n^{\mathrm{cf}} - V_n^{\mathrm{cf},0}\right) = o_p(1)$. Furthermore, we have already established in Equation (B.8) that $|(\sigma_n^{\mathrm{cf}})^2 - (\sigma_n^{\mathrm{cf},0})^2| = o_p(1)$. Since σ_n^{cf} is bounded away from zero with probability going to one, the same conclusion holds for $\sigma_n^{\mathrm{cf},0}$, and using the same logic as in for the proof of lower bounding $(\sigma_n^{\mathrm{cf}})^2$ in the proof of Theorem 5.9, we conclude:

$$\frac{\sigma_n^{\mathrm{cf},0}}{\sigma_n^{\mathrm{cf}}} - 1 = o_p(1).$$

Now, going back to $L_n^{\mathrm{cf},0}$ and T_n^{cf}, we have the following decomposition:

$$T_n^{\mathrm{cf}} - L_n^{\mathrm{cf},0} = \tfrac{\sqrt{n}\,\Delta_n^{\mathrm{cf}}}{\sigma_n^{\mathrm{cf}}} + \tfrac{\sqrt{n}\left(V_n^{\mathrm{cf}} - V_n^{\mathrm{cf},0}\right)}{\sigma_n^{\mathrm{cf}}} + L_n^{\mathrm{cf},0}\left(\tfrac{\sigma_n^{\mathrm{cf},0}}{\sigma_n^{\mathrm{cf}}} - 1\right).$$

The first term is $o_p(1)$ by the bias part of the proof of Theorem 5.9; the second term is $o_p(1)$ by the preceding calculation; and the third term is $o_p(1)$ because $L_n^{\mathrm{cf},0} = O_p(1)$ and $\sigma_n^{\mathrm{cf},0}/\sigma_n^{\mathrm{cf}} - 1 = o_p(1)$. This proves (C.11). It remains to establish the conditional asymptotic normality of $L_n^{\mathrm{cf},0}$. Define $\mathcal{H}_n = \sigma\left(A_{1:n}, X_{1:n}, \{M_i : A_i = 1\}\right)$. Then $\mathcal{X}_n \subseteq \mathcal{H}_n$. Conditional on $\mathcal{H}_n$, the oracle first balancing weights $\{\hat{\tau}_1^{(j),0}\}_{j=1}^2$, the control balancing weights $\{\hat{\tau}_2^{(j)}\}_{j=1}^2$, and $\sigma_n^{\mathrm{cf},0}$ are fixed. Moreover, the treated outcome errors and the control mediator errors in $V_n^{\mathrm{cf},0}$ are conditionally centered and independent across i. Now, define the triangular-array summands

$$\xi_{n,i}^0 = \frac{\sqrt{n}}{2\sigma_n^{\mathrm{cf},0}} \begin{cases} \hat{\tau}_{1,i}^{(2),0}\epsilon_i, & i \in \mathcal{D}_2,\ A_i = 1,\\ \hat{\tau}_{2,i}^{(1)}U_i^\top\gamma_1, & i \in \mathcal{D}_1,\ A_i = 0,\\ \hat{\tau}_{1,i}^{(1),0}\epsilon_i, & i \in \mathcal{D}_1,\ A_i = 1,\\ \hat{\tau}_{2,i}^{(2)}U_i^\top\gamma_1, & i \in \mathcal{D}_2,\ A_i = 0. \end{cases}$$

Then, it is immediate that $L_n^{\mathrm{cf},0} = \sum_{i=1}^n \xi_{n,i}^0$ and we also have $\mathbb{E}[\xi_{n,i}^0 \mid \mathcal{H}_n] = 0$, along with

$$\sum_{i=1}^n \mathbb{E}\left[(\xi_{n,i}^0)^2 \,\middle|\, \mathcal{H}_n\right] = \tfrac{n}{4(\sigma_n^{\mathrm{cf},0})^2}\sum_{j=1}^2 \left\{\sigma_1^2\|\hat{\tau}_1^{(j),0}\|_2^2 + \gamma_1^\top\Sigma_{U,0}\gamma_1\|\hat{\tau}_2^{(j)}\|_2^2\right\} = 1.$$

We next verify the conditional Lyapunov condition. By the uniform conditional sub-Gaussian assumptions on ϵ_i and U_i, and the boundedness of $\|\gamma_1\|_2$, there exists a constant $C > 0$ such that

$$\sum_{i=1}^n \mathbb{E}\left[|\xi_{n,i}^0|^3 \,\middle|\, \mathcal{H}_n\right] \le (Cn^{3/2}/(\sigma_n^{\mathrm{cf},0})^3)\sum_{j=1}^2\{\|\hat{\tau}_1^{(j),0}\|_\infty\|\hat{\tau}_1^{(j),0}\|_2^2 + \|\hat{\tau}_2^{(j)}\|_\infty\|\hat{\tau}_2^{(j)}\|_2^2\}.$$

From Lemma 5.8, Lemma B.1, and the constraints in optimization problems Equation (3.1) and (3.2), we have $n\|\hat{\tau}_1^{(j),0}\|_2^2 + n\|\hat{\tau}_2^{(j)}\|_2^2 = O_p(1)$ and $\|\hat{\tau}_1^{(j),0}\|_\infty \vee \|\hat{\tau}_2^{(j)}\|_\infty = O_p(n^{-\alpha})$. Since $\sigma_n^{\mathrm{cf},0}$ is bounded away from zero with probability going to one, it follows that

$$\sum_{i=1}^n \mathbb{E}\left[|\xi_{n,i}^0|^3 \,\middle|\, \mathcal{H}_n\right] = O_p(n^{1/2-\alpha}) = o_p(1),$$

where the last conclusion follows from $\alpha > 1/2$. Therefore, by the conditional Lyapunov central limit theorem, for every $s \in \mathbb{R}$,

$$\mathbb{E}\left[e^{isL_n^{\mathrm{cf},0}} \,\middle|\, \mathcal{H}_n\right] \longrightarrow e^{-s^2/2}, \qquad \text{in probability.} \tag{C.12}$$

To go towards $\mathcal{X}_n$ from $\mathcal{H}_n$, by law of total expectation,

$$\mathbb{E}\left[e^{isL_n^{\mathrm{cf},0}} \,\middle|\, \mathcal{X}_n\right] = \mathbb{E}\left[\mathbb{E}\left(e^{isL_n^{\mathrm{cf},0}} \,\middle|\, \mathcal{H}_n\right) \,\middle|\, \mathcal{X}_n\right].$$

Now, define

$$Z_n(s) = \left|\mathbb{E}\left(e^{isL_n^{\mathrm{cf},0}} \,\middle|\, \mathcal{H}_n\right) - e^{-s^2/2}\right|.$$

By (C.12), $Z_n(s) \to 0$ in probability, and $0 \le Z_n(s) \le 2$. Consequently, $Z_n(s) \to 0$ in L_1. Hence,

$$\left|\mathbb{E}\left[e^{isL_n^{\mathrm{cf},0}} \,\middle|\, \mathcal{X}_n\right] - e^{-s^2/2}\right| \le \mathbb{E}\left[Z_n(s) \,|\, \mathcal{X}_n\right] \longrightarrow 0, \quad \text{in probability,}$$

where the last conclusion follows from Markov's inequality and $\mathbb{E}[Z_n(s)] \to 0$. Finally, define $D_n = T_n^{\mathrm{cf}} - L_n^{\mathrm{cf},0}$. We have already established that $D_n = o_p(1)$. Therefore, for any $s \in \mathbb{R}$,

$$\begin{aligned}\left|\mathbb{E}\left[e^{isT_n^{\mathrm{cf}}} \,\middle|\, \mathcal{X}_n\right] - \mathbb{E}\left[e^{isL_n^{\mathrm{cf},0}} \,\middle|\, \mathcal{X}_n\right]\right| &\le \mathbb{E}\left[\left|e^{isD_n} - 1\right| \,\middle|\, \mathcal{X}_n\right] \\ &\le \mathbb{E}\left[\min\{2, |s||D_n|\} \,|\, \mathcal{X}_n\right] \longrightarrow 0, \quad \text{in probability.}\end{aligned}$$

Combining the last two displays, we conclude that, for every $s \in \mathbb{R}$,

$$\mathbb{E}\left[e^{isT_n^{\mathrm{cf}}} \,\middle|\, \mathcal{X}_n\right] \longrightarrow e^{-s^2/2}, \qquad \text{in probability.}$$

Since the limiting characteristic function is continuous at the origin, the conditional laws of T_n^{cf} given $\mathcal{X}_n$ converge weakly in probability to the standard normal distribution. Since Φ is continuous, this implies that, for every $t \in \mathbb{R}$,

$$\mathbb{P}\left(T_n^{\mathrm{cf}} \le t \,\middle|\, \mathcal{X}_n\right) \longrightarrow \Phi(t), \qquad \text{in probability.}$$

This completes the proof. □

Lemma C.3. *Under Assumption 5.3 and 5.4, the estimate $\hat{b}$, as obtained in Step 4 of Algorithm 1 satisfies:*

$$\|\hat{b} - b\|_1 = O_p\left(sk_\gamma\sqrt{\tfrac{\log p}{n}} + k\sqrt{\tfrac{\log p + q}{n}}\right).$$

Proof. For simplicity of presentation, we will omit the fold number here. We define the response $\check{\mathbf{y}}_c = \mathbf{M}_c\hat{\gamma}_1$ (the responses used in Step 4 of Algorithm 1). We start from the basic inequality. From the definition of $\hat{b}$, we have:

$$\hat{b} = \arg\min_{\tilde{b}}\left\{\tfrac{1}{2n_c}\|\check{\mathbf{y}}_c - \mathbf{X}_c\tilde{b}\|_2^2 + \lambda\|\tilde{b}\|_1\right\}$$

Hence, we have:

$$\tfrac{1}{2n_c}\|\check{\mathbf{y}}_c - \mathbf{X}_c\hat{b}\|_2^2 + \lambda\|\hat{b}\|_1 \le \tfrac{1}{2n_c}\|\check{\mathbf{y}}_c - \mathbf{X}_c b\|_2^2 + \lambda\|b\|_1$$

Defining $\Delta = \hat{b} - b$ and using the fact that $\check{\mathbf{y}}_c = \mathbf{X}_c b + \mathbf{U}_c\hat{\gamma}_1$, we have by some simple algebra:

$$\tfrac{1}{2n_c}\|\mathbf{X}_c\Delta\|_2^2 + \lambda\|\hat{b}\|_1 \le \left\|\tfrac{1}{n_c}\mathbf{X}_c^\top\mathbf{U}_c\hat{\gamma}_1\right\|_\infty\|\Delta\|_1 + \lambda\|b\|_1. \tag{C.13}$$

Now, suppose we choose λ such that

$$\mathbb{P}\left(\left\|\tfrac{1}{n_c}\mathbf{X}_c^\top\mathbf{U}_c\hat{\gamma}_1\right\|_\infty \le \tfrac{\lambda}{2}\right) \longrightarrow 1, \quad \text{as } n \uparrow \infty. \tag{C.14}$$

We will later show that one can take $\lambda = K_1\sqrt{\log p/n_c}$ for some suitably chosen large constant K_1. Let us call the above event $\Omega_{n,1}$. Then on $\Omega_{n,1}$, we have, from Equation (C.13),

$$\tfrac{1}{2n_c}\|\mathbf{X}_c\Delta\|_2^2 \le \tfrac{\lambda}{2}\|\Delta\|_1 + \lambda(\|b\|_1 - \|b+\Delta\|_1)\,. \tag{C.15}$$

Now, let S denotes that active set of $b_0 = \mathbf{B}_0^\top\gamma_1$. Then we have $|S| \le sk_\gamma$. Furthermore,

$$\begin{aligned}\|b\|_1 - \|b+\Delta\|_1 &= \|b_S\|_1 - \|b_S+\Delta_S\|_1 + \|b_{S^c}\|_1 - \|b_{S^c}+\Delta_{S^c}\|_1\\ &\le \|\Delta_S\|_1 + \|b_{S^c}\|_1 - (\|\Delta_{S^c}\|_1 - \|b_{S^c}\|_1)\end{aligned}$$

Hence, using this in Equation (C.15), along with the fact that $(\lambda/2)\|\Delta\|_1 = (\lambda/2)\|\Delta_S\|_1 + (\lambda/2)\|\Delta_{S^c}\|_1)$ we have:

$$\tfrac{1}{2n_c}\|\mathbf{X}_c\Delta\|_2^2 + \tfrac{\lambda}{2}\|\Delta_{S^c}\|_1 \le \tfrac{3\lambda}{2}\|\Delta_S\|_1 + 2\lambda\|b_{S^c}\|_1\,. \tag{C.16}$$

As $\|\mathbf{X}_c\Delta\|_2^2 \ge 0$, it is immediate from Equation (C.16)

$$\|\Delta_{S^c}\|_1 \le 3\|\Delta_S\|_1 + 4\|b_{S^c}\|_1\,. \tag{C.17}$$

We divide the rest of the analysis into two cases, depending on whether $\|\Delta_S\|_1$ is larger than $\|b_{S^c}\|_1$ or not:
Case 1: $4\|b_{S^c}\|_1 \le 3\|\Delta_S\|_1$**.** In this case, we have from Equation (C.17), $\|\Delta_{S^c}\|_1 \le 6\|\Delta_S\|_1$. Also from Equation (C.16), we have:

$$\tfrac{1}{2n_c}\|\mathbf{X}_c\Delta\|_2^2 + \tfrac{\lambda}{2}\|\Delta_{S^c}\|_1 \le 3\lambda\|\Delta_S\|_1 \tag{C.18}$$

Furthermore, the compatibility condition holds, and define the event $\Omega_{n,2}$, on which $\mathbf{X}_c$ satisfies the RE condition. Then by Assumption 5.3 and Lemma C.1, $\mathbb{P}(\Omega_{n,2}) \to 1$ as $n \uparrow \infty$ and consequently, on $\Omega_{n,2}$, we have for some constant ϕ_0:

$$\tfrac{\|\mathbf{X}_c\Delta\|_2^2}{n_c} \ge \phi_0\|\Delta\|_2^2 \implies \|\Delta_S\|_1 \le \sqrt{sk_\gamma}\|\Delta_S\|_2 \le \sqrt{sk_\gamma}\|\Delta\|_2 \le \sqrt{sk_\gamma}\tfrac{\|\mathbf{X}_c\Delta\|_2}{\sqrt{n_c}\sqrt{\phi_0}}\,. \tag{C.19}$$

Hence, from Equation (C.18), we have:

$$\tfrac{1}{2n_c}\|\mathbf{X}_c\Delta\|_2^2 + \tfrac{\lambda}{2}\|\Delta_{S^c}\|_1 \le 3\lambda\sqrt{sk_\gamma}\tfrac{\|\mathbf{X}_c\Delta\|_2}{\sqrt{n_c}\sqrt{\phi_0}} \le \tfrac{1}{4n_c}\|\mathbf{X}_c\Delta\|_2^2 + \tfrac{9}{\phi_0}sk_\gamma\lambda^2 \tag{C.20}$$

Hence, changing sides, yields:

$$\tfrac{1}{n_c}\|\mathbf{X}_c\Delta\|_2^2 \le \tfrac{36}{\phi_0}sk_\gamma\lambda^2 \implies \tfrac{1}{\sqrt{n_c}}\|\mathbf{X}_c\Delta\|_2 \le \tfrac{6}{\sqrt{\phi_0}}\lambda\sqrt{sk_\gamma}$$

This, along with Equation (C.19) implies $\|\Delta_S\|_1 \le (6\lambda sk_\gamma)/\phi_0$ and consequently, using the fact under this case 1 $\|\Delta_{S^c}\|_1 \le 6\|\Delta_S\|_1$, we conclude that:

$$\|\Delta_1\|_1 = \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1 \le 7\|\Delta_S\|_1 \le \tfrac{42}{\phi_0}\lambda sk_\gamma = O_p\left(sk_\gamma\sqrt{\tfrac{\log p}{n}}\right)\,. \tag{C.21}$$

Here in the last line we have used the fact $\lambda = K_1\sqrt{\log p/n_c}$, a fact that we will prove at the end, and $n_t \sim n_c \sim n$.
Case 2: $4\|b_{S^c}\|_1 > 3\|\Delta_S\|_1$**:** In this case, we have from Equation (C.16),

$$\tfrac{1}{2n_c}\|\mathbf{X}_c\Delta\|_2^2 + \tfrac{\lambda}{2}\|\Delta_{S^c}\|_1 \le 4\lambda\|b_{S^c}\|_1 \implies \|\Delta_{S^c}\|_1 \le 8\|b_{S^c}\|_1 \implies \|\Delta\|_1 \le \tfrac{28}{3}\|b_{S^c}\|_1\,. \tag{C.22}$$

Now, $\|b_{S^c}\|_1 = \|(b-b_0)_{S^c}\|_1 \le \|b-b_0\|_1$, where in the first equality, we have used the fact that $(b_0)_{S^c} \equiv 0$ by definition of S. On the other have, we have:

$$\|b-b_0\|_1 = \|\mathbf{B}_0^\top(\hat\gamma_1 - \gamma_1)\|_1 \le \|\mathbf{B}_0\|_{1,\infty}\|\hat\gamma_1 - \gamma_1\|_1 = O_p\left(k\sqrt{\tfrac{\log(p+q)}{n}}\right)$$

where the last conclusion follows from the fact that $\|\mathbf{B}_0\|_{1,\infty} \le C_{\mathbf{B}} < \infty$ and $\|\hat\gamma_1 - \gamma_1\|_1 \le \|\hat\phi_1 - \phi_1\|_1 = O_p(k\sqrt{\log(p+q)/n})$. This completes the proof. Now, all we need to show is that $\lambda = K_1\sqrt{\log p/n_c}$ is a valid choice for Equation (C.14) to hold. As $\hat\gamma_1$ is obtained from separate data (treatment data), and U_i's

are sub-Gaussian conditional on X, it is immediate from the sub-Gaussian concentration inequality that for a large enough constant C,

$$\mathbb{P}\left(\frac{1}{n_c\|\hat{\gamma}_1\|_2}\|\mathbf{X}_c^\top \mathbf{U}_c\hat{\gamma}_1\|_\infty > C\sqrt{\frac{\log p}{n_c}} \mid \hat{\gamma}_1\right) \to 0, \quad \text{as } n \uparrow \infty\,.$$

Consequently $\|\mathbf{X}_c^\top \mathbf{U}_c\hat{\gamma}_1\|_\infty/n_c \le C\|\hat{\gamma}_1\|_2\sqrt{\log p/n_c}$ with probability going to 1. Now $\|\hat{\gamma}_1\|_2 \le \|\gamma_1\|_2 + \|\hat{\gamma}_1 - \gamma_1\|_2 \le 1+C_\gamma$ with probability going to 1 as $\|\hat{\gamma}_1-\gamma_1\|_2 \le \|\hat{\phi}_1-\phi_1\|_2 = o_p(1)$ which follows from the consistency of the lasso estimator. Therefore, $\|\mathbf{X}_c^\top \mathbf{U}_c\hat{\gamma}_1\|_\infty/n_c \le C(1+C_\gamma)\sqrt{\log p/n_c}$. Taking $K_1 = 2C(1+C_\gamma)$ would conclude the proof. □

C.4 Calculation of Effects

By taking expectations of the outcome and mediator regressions conditional on (X, M, A) and (X, A), respectively, we obtain

$$\begin{aligned}
\mathbb{E}[Y(1, M(1))] &= \alpha_1 + \mathbb{E}[X]^\top\beta_1 + \mathbb{E}[\mathbb{E}[M \mid X, A=1]]^\top\gamma_1 \\
&= (\alpha_1 + \delta_1^\top\gamma_1) + \mathbb{E}[X]^\top(\beta_1 + \mathbf{B}_1^\top\gamma_1), \\
\mathbb{E}[Y(0, M(0))] &= \alpha_0 + \mathbb{E}[X]^\top\beta_0 + \mathbb{E}[\mathbb{E}[M \mid X, A=0]]^\top\gamma_0 \\
&= (\alpha_0 + \delta_0^\top\gamma_0) + \mathbb{E}[X]^\top(\beta_0 + \mathbf{B}_0^\top\gamma_0), \\
\mathbb{E}[Y(1, M(0))] &= \alpha_1 + \mathbb{E}[X]^\top\beta_1 + \mathbb{E}[\mathbb{E}[M \mid X, A=0]]^\top\gamma_1 \\
&= (\alpha_1 + \delta_0^\top\gamma_1) + \mathbb{E}[X]^\top(\beta_1 + \mathbf{B}_0^\top\gamma_1), \\
\mathbb{E}[Y(0, M(1))] &= \alpha_0 + \mathbb{E}[X]^\top\beta_0 + \mathbb{E}[\mathbb{E}[M \mid X, A=1]]^\top\gamma_0 \\
&= (\alpha_0 + \delta_1^\top\gamma_0) + \mathbb{E}[X]^\top(\beta_0 + \mathbf{B}_1^\top\gamma_0).
\end{aligned}$$

Therefore, the natural indirect effect simplifies as

$$\begin{aligned}
\text{NIE} &= \mathbb{E}[Y(1, M(1))] - \mathbb{E}[Y(1, M(0))] \\
&= (\alpha_1 + \delta_1^\top\gamma_1) + \mathbb{E}[X]^\top(\beta_1 + \mathbf{B}_1^\top\gamma_1) - (\alpha_1 + \delta_0^\top\gamma_1) - \mathbb{E}[X]^\top(\beta_1 + \mathbf{B}_0^\top\gamma_1) \\
&= (\delta_1 - \delta_0)^\top\gamma_1 + \mathbb{E}[X]^\top(\beta_1 - \beta_1 + \mathbf{B}_1^\top\gamma_1 - \mathbf{B}_0^\top\gamma_1) \\
&= (\delta_1 - \delta_0)^\top\gamma_1 + \mathbb{E}[X]^\top(\mathbf{B}_1 - \mathbf{B}_0)^\top\gamma_1.
\end{aligned}$$

Similarly, the natural direct effect simplifies as

$$\begin{aligned}
\text{NDE} &= \mathbb{E}[Y(1, M(0))] - \mathbb{E}[Y(0, M(0))] \\
&= (\alpha_1 + \delta_0^\top\gamma_1) + \mathbb{E}[X]^\top(\beta_1 + \mathbf{B}_0^\top\gamma_1) - (\alpha_0 + \delta_0^\top\gamma_0) - \mathbb{E}[X]^\top(\beta_0 + \mathbf{B}_0^\top\gamma_0) \\
&= \alpha_1 - \alpha_0 + \delta_0^\top(\gamma_1 - \gamma_0) + \mathbb{E}[X]^\top(\beta_1 + \mathbf{B}_0^\top\gamma_1 - \beta_0 - \mathbf{B}_0^\top\gamma_0) \\
&= \alpha_1 - \alpha_0 + \delta_0^\top(\gamma_1 - \gamma_0) + \mathbb{E}[X]^\top(\beta_1 - \beta_0) + \mathbb{E}[X]^\top\mathbf{B}_0^\top(\gamma_1 - \gamma_0).
\end{aligned}$$

D Connection with the Existing Literatures

D.1 Comparison with Guo et al. (2022, 2023, 2024) and the role of treatment interactions

In this section, we elaborate on the relationship between our method and existing high-dimensional mediation inference methods, with particular emphasis on the trilogy of (Guo et al., 2022, 2023, 2024). These works make significant contributions to high-dimensional mediation analysis, and our inferential objective is aligned with theirs at a high level: namely, to conduct valid statistical inference for mediated treatment effect, e.g., the natural indirect effect (NIE) and the natural direct effect (NDE) in high-dimensional settings. However, the framework studied in our paper extends this line of work primarily in three directions:

Algorithm A1 Debiased estimation of the natural indirect effect in the absence of interactions

1: **Step 1.** Regress Y_i on $(A_i, \mathbf{X}_i)$ with ℓ_1 penalty on β:

$$\hat{\phi}_1 = (\hat{\alpha}_1^{(1)}, \hat{\beta}^{(1)}) = \arg\min_{\alpha_1,\beta} \sum_i \left(Y_i - A_i\alpha_1 - \mathbf{X}_i^\top \beta\right)^2 + \lambda_1 \|\beta\|_1.$$

Let $\mathbf{W}_i^{(1)} = (A_i, \mathbf{X}_i^\top)^\top$ and $a = (1, \mathbf{0}^\top)^\top$. The vector a extracts the treatment coefficient in the reduced outcome regression. Compute τ_a via

$$\tau_a = \arg\min_{\tilde{\tau}} \|\tilde{\tau}\|_2^2 \text{ s.t. } \left\| a - (\mathbf{W}^{(1)})^\top \tilde{\tau} \right\|_\infty \le K_a \sqrt{\frac{\log(1+p)}{n}},\ \|\tilde{\tau}\|_\infty \le n^{-2/3}.$$

Set

$$\hat{\mathrm{TE}} = a^\top \hat{\phi}_1 + \sum_i \tau_{a,i} \left(Y_i - A_i \hat{\alpha}_1^{(1)} - \mathbf{X}_i^\top \hat{\beta}^{(1)}\right).$$

2: **Step 2.** Regress Y_i on $(A_i, \mathbf{X}_i, \mathbf{M}_i)$ with ℓ_1 penalty on (β, γ):

$$\hat{\phi}_2 = (\hat{\alpha}_1^{(2)}, \hat{\beta}^{(2)}, \hat{\gamma}) = \arg\min_{\alpha_1,\beta,\gamma} \sum_i \left(Y_i - A_i\alpha_1 - \mathbf{X}_i^\top \beta - \mathbf{M}_i^\top \gamma\right)^2 + \lambda_2(\|\beta\|_1 + \|\gamma\|_1).$$

Let $\mathbf{W}_i^{(2)} = (A_i, \mathbf{X}_i^\top, \mathbf{M}_i^\top)^\top$ and $b = (1, \mathbf{0}^\top, \mathbf{0}^\top)^\top$. The vector b extracts the treatment coefficient after adjusting for both covariates and mediators. Compute τ_b analogously and set

$$\hat{\mathrm{NDE}} = b^\top \hat{\phi}_2 + \sum_i \tau_{b,i} \left(Y_i - A_i \hat{\alpha}_1^{(2)} - \mathbf{X}_i^\top \hat{\beta}^{(2)} - \mathbf{M}_i^\top \hat{\gamma}\right).$$

3: **Step 3.** Compute the natural indirect effect as the difference between the total effect and the natural direct effect:

$$\hat{\mathrm{NIE}} = \hat{\mathrm{TE}} - \hat{\mathrm{NDE}}.$$

4: **Return** $\hat{\mathrm{NIE}}$.

1. We allow ultra-high-dimensional (UHD) covariates X in addition to UHD mediators M.
2. More importantly, we allow treatment interactions in both stages of the data-generating mechanism; in particular, we allow A–X and A–M interactions in the generation of the response, and A–X interactions in the generation of the mediators.
3. Our assumptions required for high-dimensional inference are weaker than those of (Guo et al., 2022, 2023, 2024)

It is the combination of UHD covariates and treatment interactions that fundamentally changes the inferential structure. Note that there exist several works in the causal inference literature discussing the challenges added by interaction effects (e.g., the textbook VanderWeele (2015)), yet such investigations are, to the best of our knowledge, scarce in the high-dimensional mediation analysis literature. To clarify the source of the difficulty, it is useful to compare the basic mediation functional $\theta_0 := \mathbb{E}[Y^{(1,M^{(0)})}]$ under a no-interaction model and under our interaction-rich model:

1. The role of the simultaneous presence of interaction and UHD X. To isolate the role of interaction, suppose we naturally extend (Guo et al., 2022, 2023, 2024) by allowing UHD covariates X, but without treatment interactions:

$$\begin{aligned} Y &= \alpha_0 + A\alpha_1 + X^\top \beta + M^\top \gamma + \varepsilon, \\ M &= \delta_0 + A\delta_1 + BX + U. \end{aligned} \tag{D.1}$$

Under this model, the sample average mediation functional is

$$\theta_0 = \mathbb{E}[Y^{(1,M^{(0)})}] = \alpha_0 + \alpha_1 + \delta_0^\top \gamma + \bar{X}^\top (\beta + B^\top \gamma).$$

Although this expression contains the term $B^\top\gamma$, in the no-interaction model, this term can be absorbed into a single reduced-form outcome regression. Indeed, substituting the mediator equation into the outcome

Algorithm A2 Guo et al. (2022) estimator for the natural indirect effect

1: **Step 1.** Estimate

$$(\hat{\gamma}, \hat{\alpha}_1, \hat{\beta}) = \arg\min_{\gamma,\alpha_1,\beta} \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - A_i\alpha_1 - X_i^\top\beta - M_i^\top\gamma\right)^2 + \sum_{j=1}^{q} p_\lambda(|\gamma_j|),$$

where $p_\lambda(\cdot)$ denotes the SCAD penalty. The tuning parameter λ is selected by HBIC.

2: **Step 2.** Define the selected mediator set $\hat{\mathcal{A}} = \{j : \hat{\gamma}_j \neq 0\}$ and construct the restricted mediator matrix $M_{\hat{\mathcal{A}}}$.

3: **Step 3.** Refit the outcome model on the selected mediators:

$$(\hat{\gamma}^{\mathrm{rf}}_{\hat{\mathcal{A}}}, \hat{\alpha}^{\mathrm{rf}}_1, \hat{\beta}^{\mathrm{rf}}) = \arg\min_{\gamma_{\hat{\mathcal{A}}},\alpha_1,\beta} \sum_{i=1}^{n} \left(Y_i - A_i\alpha_1 - X_i^\top\beta - M_{i,\hat{\mathcal{A}}}^\top\gamma_{\hat{\mathcal{A}}}\right)^2.$$

Set $\widehat{\mathrm{NDE}} = \hat{\alpha}^{\mathrm{rf}}_1$.

4: **Step 4.** Using the reduced-form representation

$$Y = \alpha_0 + \delta_0^\top\gamma + A(\alpha_1 + \delta_1^\top\gamma) + X^\top(\beta + \mathbf{B}^\top\gamma) + U^\top\gamma + \epsilon,$$

estimate the total effect by regressing Y_i on (A_i, X_i):

$$(\hat{\theta}_A, \hat{\theta}_X) = \arg\min_{\theta_A,\theta_X} \sum_{i=1}^{n} \left(Y_i - A_i\theta_A - X_i^\top\theta_X\right)^2.$$

Set $\widehat{\mathrm{TE}} = \hat{\theta}_A$.

5: **Step 5.** Compute

$$\widehat{\mathrm{NIE}} = \widehat{\mathrm{TE}} - \widehat{\mathrm{NDE}} = \hat{\theta}_A - \hat{\alpha}^{\mathrm{rf}}_1.$$

6: **Return** $\widehat{\mathrm{NIE}}$.

equation gives

$$Y = (\alpha_0 + \delta_0^\top\gamma) + A(\alpha_1 + \delta_1^\top\gamma) + X^\top(\beta + B^\top\gamma) + U^\top\gamma + \varepsilon.$$

Thus, the high-dimensional component $\bar{X}^\top(\beta + B^\top\gamma)$ is now a standard linear functional of a single reduced-form regression coefficient $\beta + B^\top\gamma$ of X. Consequently, even if X is UHD, the problem remains similar to standard high-dimensional linear functional inference: one may first estimate this regression coefficient using penalized regression of Y on (A, X) and then debias it in the direction $\bar{X}$ using the techniques of (Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014).

The situation changes fundamentally once treatment interactions are introduced. Under our model,

$$\begin{aligned} Y &= (1-A)(\alpha_0 + X^\top\beta_0 + M^\top\gamma_0 + \varepsilon') + A(\alpha_1 + X^\top\beta_1 + M^\top\gamma_1 + \varepsilon), \\ M &= (1-A)(\delta_0 + B_0X + U) + A(\delta_1 + B_1X + U'), \end{aligned}$$

the same mediation functional becomes

$$\theta_0 = \mathbb{E}[Y^{(1,M^{(0)})}] = \alpha_1 + \delta_0^\top\gamma_1 + \bar{X}^\top(\beta_1 + B_0^\top\gamma_1),$$

Suppressing intercept terms, which are not the main difficulty, the challenging part is $\bar{X}^\top B_0^\top\gamma_1$. Here B_0 is identified from the mediator model in the control arm, whereas γ_1 and β_1 are identified from the outcome model in the treated arm. Therefore, $\bar{X}^\top B_0^\top\gamma_1$ is not a fixed linear functional of a single high-dimensional regression coefficient. It is a cross-arm functional involving a UHD matrix B_0 (whose rows and columns can be much larger than the sample size) and a UHD vector γ_1. This is precisely the structural feature that prevents a direct reduction to a standard one-step debiased Lasso problem. To see this point from another

Algorithm A3 Guo et al. (2022) estimator for the natural indirect effect with modification for high-dimensional covariates

1: **Step 1.** Estimate

$$(\hat{\gamma}, \hat{\alpha}_1, \hat{\beta}) = \arg\min_{\gamma,\alpha_1,\beta} \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - A_i\alpha_1 - X_i^\top\beta - M_i^\top\gamma\right)^2 + \sum_{j=1}^{q} p_\lambda(|\gamma_j|) + \sum_{k=1}^{p} p_\lambda(|\beta_k|),$$

where the tuning parameter λ is selected by HBIC.

2: **Step 2.** Define $\hat{\mathcal{A}}_M = \{j : \hat{\gamma}_j \neq 0\}$ and $\hat{\mathcal{A}}_X = \{k : \hat{\beta}_k \neq 0\}$. Construct $M_{\hat{\mathcal{A}}_M}$ and $X_{\hat{\mathcal{A}}_X}$.

3: **Step 3.** Refit the outcome model on the selected mediators and covariates:

$$(\hat{\gamma}^{\mathrm{rf}}_{\hat{\mathcal{A}}_M}, \hat{\alpha}^{\mathrm{rf}}_1, \hat{\beta}^{\mathrm{rf}}_{\hat{\mathcal{A}}_X}) = \arg\min_{\gamma_{\hat{\mathcal{A}}_M},\alpha_1,\beta_{\hat{\mathcal{A}}_X}} \sum_{i=1}^{n} \left(Y_i - A_i\alpha_1 - X_{i,\hat{\mathcal{A}}_X}^\top\beta_{\hat{\mathcal{A}}_X} - M_{i,\hat{\mathcal{A}}_M}^\top\gamma_{\hat{\mathcal{A}}_M}\right)^2.$$

Set $\widehat{\mathrm{NDE}} = \hat{\alpha}^{\mathrm{rf}}_1$.

4: **Step 4.** Using the reduced-form representation

$$Y = \alpha_0 + \delta_0^\top\gamma + A(\alpha_1 + \delta_1^\top\gamma) + X^\top(\beta + \mathbf{B}^\top\gamma) + U^\top\gamma + \epsilon,$$

estimate the total effect by regressing Y_i on $(A_i, X_{i,\hat{\mathcal{A}}_X})$:

$$(\hat{\theta}_A, \hat{\theta}_X) = \arg\min_{\theta_A,\theta_X} \sum_{i=1}^{n} \left(Y_i - A_i\theta_A - X_{i,\hat{\mathcal{A}}_X}^\top\theta_X\right)^2.$$

Set $\widehat{\mathrm{TE}} = \hat{\theta}_A$.

5: **Step 5.** Compute

$$\widehat{\mathrm{NIE}} = \widehat{\mathrm{TE}} - \widehat{\mathrm{NDE}} = \hat{\theta}_A - \hat{\alpha}^{\mathrm{rf}}_1.$$

6: **Return** $\widehat{\mathrm{NIE}}$.

angle, if the A–X interaction in the mediator model were absent, then $B_0 = B_1$. In that case, the high-dimensional coefficient $\beta_1 + B_0^\top\gamma_1$ would coincide with the coefficient of X in the treated-arm reduced-form regression. However, when $B_0 \neq B_1$, the coefficient obtained from the treated-arm reduced-form regression is $\beta_1 + B_1^\top\gamma_1$, whereas the counterfactual mediation functional θ_0 requires $\beta_1 + B_0^\top\gamma_1$. This cross-arm structure is absent in the no-interaction model and is also absent in standard single-model high-dimensional linear functional inference.

2. Comparison of assumptions. Another important difference between our approach and that of Guo et al. is the set of conditions under which high-dimensional inference is performed. In (Guo et al., 2022, 2023, 2024), the authors do not use a debiasing step. They do not penalize the low-dimensional parameter of interest and penalize only the high-dimensional nuisance component. To illustrate the idea, let us consider the simple high-dimensional linear model

$$Y_i = X_i^\top\beta_0 + \varepsilon_i = X_{i1}\beta_{0,1} + X_{i,-1}^\top\beta_{0,-1} + \varepsilon_i,$$

where the goal is to infer on the first coordinate $\beta_{0,1}$. An analogue of the partial-penalization approach of Guo et al. (2023) would estimate $\beta_0 = (\beta_{0,1}, \beta_{0,-1})$ by

$$(\hat{\beta}_1, \hat{\beta}_{-1}) = \arg\min_{b_1,b_{-1}} \left\{ \frac{1}{2n}\sum_{i=1}^{n} \left(Y_i - X_{i1}b_1 - X_{i,-1}^\top b_{-1}\right)^2 + \sum_{j=2}^{p} \rho_{\lambda_n}(|b_j|) \right\},$$

where ρ_{λ_n} is typically a nonconvex penalty such as SCAD or MCP. Importantly, the first coordinate is not penalized. Under their assumptions, Guo et al. (2023) then show that the resulting estimator of the

low-dimensional target parameter (here $\beta_{0,1}$) is $\sqrt{n}$-consistent and asymptotically normal, i.e., asymptotic normality is achieved without an additional debiasing step.

However, the achievement of asymptotic normality without debiasing comes at the cost of stronger assumptions. In particular, their Condition (A1) includes a support-recovery-oriented irrepresentability type of condition, namely an $L_{2,\infty}$-type bound such as $\|M_{\mathcal{A}^c}^\top(X, M_{\mathcal{A}})\|_{2,\infty} = O_p(n)$ (where $\mathcal{A}$ is the active set and in our linear regression example above it roughly translates to $\|X_{-1,\mathcal{A}^c}^\top(X_1, X_{-1,\mathcal{A}})\|_{2,\infty} = O_p(n)$), and, more importantly, their Condition (A2) imposes a β-min condition on the nonzero entries of the high-dimensional regression vector, i.e.,

$$d_n = \frac{1}{2}\min_{j\in A}|\alpha_{0j}^*| = \frac{1}{2}\min_{j\in\mathcal{A}}|\beta_{0,j}| \gg \lambda_n \gg \max\left\{\sqrt{s/n}, \sqrt{\log p/n}\right\}. \tag{D.2}$$

Hence, under their condition, all active coefficients must be sufficiently separated from zero. This is precisely the type of condition that debiased-Lasso methods are designed to avoid. For example, Zhang and Zhang (2014) emphasize that the main distinction between their low-dimensional projection(LDP)/debiasing approach and variable-selection-based approaches is the latter's reliance on a uniform signal strength condition and to quote: "However, a main complaint about the variable selection approach is the practicality of the uniform signal strength condition, and the crucial difference between the two approaches is precisely in the case where the condition fails to hold. Without the condition, the LDP approach is still able to select correctly all coefficients above a threshold of order $C\sqrt{(2/n)\log p}$ and all zero coefficients" (kindly see the Introduction section of Zhang and Zhang (2014) for more details). In other words, debiased-Lasso inference of Zhang and Zhang (2014) (or Van de Geer et al. (2014); Javanmard and Montanari (2014)) does not require a β-min condition on the nonzero nuisance coefficients; weak but nonzero coefficients are also allowed, provided the usual sparsity and other regularity conditions hold.

Condition (A2) (Equation (D.2)) of Guo et al. (2023) also has further implications. This condition requires $d_n \gg \sqrt{s/n}$. Now, in the standard debiased Lasso approach, we require $s\log p/\sqrt{n} \to 0$, i.e., we can allow sparsity up to order $\sqrt{n}/\log p$. Therefore, if we set $s \sim \sqrt{n}/(\log p)^2$, then the minimal signal strength condition of Guo et al. (2023) requires $d_n \gg \sqrt{s/n} \gg n^{-1/4}(\log p)^{-1}$, which is fairly strong condition on the minimal active signal strength and rules out relatively weak nonzero coefficients. Furthermore, the asymptotic normality result, Theorem 1 and Corollary 1 of Guo et al. (2023), requires $s = o(n^{1/3})$, whereas, as mentioned before, debiased-Lasso analyses typically allow sparsity of order $(s\log p)/\sqrt{n} \to 0$. Since our approach is also motivated by the debiased-Lasso idea, we *do not impose a minimal signal strength condition on the regression coefficients*, and our analysis also *allows the relevant sparsity levels* (s_β, s_γ, s_k) *to be of order* $o(\sqrt{n})$*, up to the logarithmic factors*. In this sense, our inferential theory requires weaker signal-strength and sparsity conditions than (Guo et al., 2022, 2023, 2024). This distinction is also reflected in our simulations: when the sparsity level is high (Table 4) or the minimum active coefficient is small (Table 5), our method outperforms theirs.

In summary, when treatment interactions are absent, the mediation functional reduces to a standard linear functional of a high-dimensional reduced-form regression coefficient, making conventional Lasso-plus-debiasing arguments more directly applicable. By contrast, in the interaction-rich model considered in our work, even the basic mediation functional $\mathbb{E}[Y^{(1,M^{(0)})}]$ involves the composite cross-arm term $\bar{X}^\top B_0^\top \gamma_1$, which is the primary source of the additional technical difficulty and motivates the proposed multi-step debiasing procedure. Furthermore, our assumptions on sparsity are weaker than (Guo et al., 2022, 2023, 2024), and we do not require any assumption on the minimal signal strength. Thus, although our inferential objective is aligned with (Guo et al., 2022, 2023, 2024), the structural model studied in this paper is fundamentally more challenging. The simultaneous presence of UHD covariates, UHD mediators, and treatment interactions changes the structure of the mediation functional in a way that is not covered by settings where either UHD covariates or interaction effects are absent. This distinction necessitates both the multi-step debiasing strategy developed in our paper and a separate theoretical analysis.

Comparison via numerical experiments: We next provide an additional numerical comparison with Guo et al. (2022). The purpose of this comparison is to evaluate the inferential behavior of the proposed debiased estimator relative to an existing inference-oriented method under a common data-generating framework where both methods can be implemented. Since existing high-dimensional mediation methods, such as Guo et al. (2022, 2023, 2024), do not directly accommodate the full interaction-rich UHD covariate setting, we conduct

this comparison under the reduced no-interaction setting in Equation (D.1). This ensures that both the proposed method and the method of Guo et al. (2022) can be applied.

We compare the specialization of our proposed debiased estimator to the no-interaction data-generating process in Equation (D.1), summarized in Algorithm A1, with a version of the estimator proposed by Guo et al. (2022). Under this reduced model, the target admits a reduced-form representation; hence, this special-case implementation does not require estimation of the full mediator-on-covariate coefficient matrix at all. Since the original procedure in Guo et al. (2022) does not account for UHD covariates, we consider a natural modification of their method that incorporates penalization of both the mediator and covariate coefficients. For completeness, Algorithm A2 records the baseline estimator following Guo et al. (2022), while Algorithm A3 presents the modified version used for high-dimensional covariates. We next describe our simulation setup. The generative model for (X, A, Y, M) remains the same; the two different regimes we consider differ in terms of the strength of the nonzero coefficients, as specified below.

1. **Covariates.** For each simulation run, we generate $X_i \overset{iid}{\sim} \mathcal{N}(\mu\mathbf{1}, \Sigma_X)$, where $\mu = 0.5$ and $\Sigma_X = R_X \Lambda_X R_X^\top$ with R_X orthonormal and $\Lambda_X = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, where $\lambda_j \sim \mathcal{U}(0.1, 0.5)$ (generated once and kept fixed throughout the Monte Carlo iterations).
2. **Treatment assignment.** The binary treatment A is generated from
$$\mathbb{P}(A = 1 \mid X) = \mathrm{logit}^{-1}(X^\top \alpha_A),$$
where $\|\alpha_A\|_0 = s$ is with nonzero entries sampled from $\mathcal{U}(0, 2)$.
3. **Mediators.** Given (A, X), we generate mediators from $M = A\delta_1 + BX + U$, where $\|\delta_1\|_0 = s$ and B is row-sparse with $\|B_{i,*}\|_0 = s_B$, with non-zero entries sampled from $\mathcal{U}(0.05, 0.2)$. The noise $U \sim \mathcal{N}(0, \Sigma_U)$, where $\Sigma_U = R_U \Lambda_U R_U^\top$ with R_U orthonormal and $\lambda_j \sim \mathcal{U}(0.1, 0.5)$.
4. **Outcome.** The outcomes are generated from the model $Y = \alpha_1 A + X^\top \beta + M^\top \gamma + \epsilon$, where $\alpha_1 = 0.15$; $\|\beta\|_0 = \|\gamma_0\|_0 = s$ and $\epsilon \sim \mathcal{N}(0, 1)$.

We set $p = q \in \{50, 400, 750, 1250\}$ and $n \in \{250, 500, 750, 1000, 1250\}$. Across all regimes the sparsity levels grow with the ambient dimension according to

$$(s, s_B) = \begin{cases} (5, 5), & p = 50, \\ (10, 10), & p = 400 \text{ or } 750, \\ (20, 20), & p = 1250. \end{cases}$$

The two regimes differ only in how the nonzero entries of the mediation and outcome coefficients $(\delta_1, \beta, \gamma)$ are generated:

- **Increasing sparsity with moderate minimum active signal strength:** Every nonzero coefficient is a moderate signal sampled from $\mathcal{U}(0.05, 0.2)$.
- **Increasing sparsity with vanishing minimum active signal strength:** Half of the nonzero coefficients (chosen at random) are moderate signals sampled from $\mathcal{U}(0.05, 0.2)$, while the remaining half are weak signals of order $n^{-1/2}$, sampled from $\mathcal{U}(0.8, 1.2)/\sqrt{n_{\max}}$ with $n_{\max} = 1250$.

We report RMSE, empirical standard deviation, and confidence interval length, and empirical coverage to compare the performance. The results are summarized in Tables 4 and 5, with the corresponding visualizations provided in Figure 3. Across both regimes, the comparison reveals that the modified procedure based on Guo et al. (2022) can perform reasonably in some lower-dimensional or larger-sample settings, but it becomes unstable in the ultra-high-dimensional regimes, especially when the sample size is small. In contrast, the proposed debiased estimator demonstrates more stable inferential performance across the regimes considered, with empirical coverage generally closer to the nominal level relative to the competing estimator.

increasing sparsity		$\sigma^2 = 1$													
Estimators		Debiasing (ours)							(Guo et al., 2022)						
		Monte Carlo CI				Analytical CI			Monte Carlo CI				Analytical CI		
$p+q$	n	RMSE	SD	CI Len	Cov	SE	CI Len	Cov	RMSE	SD	CI Len	Cov	SE	CI Len	Cov
100	250	0.101	0.095	0.372	0.936	0.099	0.387	0.870	0.065	0.065	0.253	0.966	0.026	0.102	0.568
	500	0.087	0.079	0.311	0.916	0.084	0.330	0.936	0.050	0.049	0.193	0.974	0.017	0.067	0.534
	750	0.078	0.074	0.289	0.940	0.077	0.303	0.948	0.044	0.044	0.171	0.958	0.018	0.069	0.626
	1000	0.069	0.065	0.253	0.928	0.071	0.279	0.952	0.043	0.041	0.159	0.954	0.017	0.068	0.626
	1250	0.066	0.062	0.241	0.928	0.052	0.204	0.948	0.049	0.036	0.141	0.854	0.018	0.072	0.504
800	250	0.126	0.115	0.451	0.926	0.099	0.390	0.880	0.171	0.097	0.378	0.732	0.063	0.247	0.452
	500	0.096	0.091	0.357	0.948	0.083	0.325	0.914	0.216	0.105	0.413	0.610	0.048	0.187	0.098
	750	0.079	0.076	0.300	0.950	0.081	0.316	0.906	0.268	0.134	0.525	0.688	0.042	0.165	0.068
	1000	0.074	0.075	0.292	0.952	0.070	0.276	0.936	0.333	0.163	0.640	0.622	0.039	0.154	0.056
	1250	0.071	0.070	0.276	0.946	0.061	0.241	0.962	0.426	0.239	0.936	0.718	0.037	0.146	0.072
1500	250	0.141	0.128	0.503	0.916	0.099	0.387	0.838	0.161	0.102	0.398	0.796	0.064	0.250	0.562
	500	0.132	0.099	0.388	0.876	0.084	0.330	0.792	0.196	0.108	0.425	0.714	0.050	0.196	0.252
	750	0.085	0.083	0.324	0.936	0.076	0.300	0.922	0.259	0.122	0.478	0.576	0.044	0.174	0.058
	1000	0.079	0.076	0.297	0.936	0.072	0.281	0.920	0.311	0.137	0.539	0.510	0.041	0.161	0.052
	1250	0.070	0.069	0.272	0.942	0.061	0.239	0.922	0.398	0.199	0.779	0.660	0.038	0.151	0.042
2500	250	0.186	0.146	0.574	0.870	0.102	0.399	0.666	3.169	1.177	4.612	0.312	0.258	1.012	0.002
	500	0.162	0.128	0.502	0.894	0.089	0.349	0.702	3.950	0.890	3.487	0.034	0.190	0.743	0.002
	750	0.131	0.097	0.382	0.830	0.081	0.317	0.734	3.911	1.130	4.431	0.128	0.165	0.648	0.000
	1000	0.097	0.080	0.313	0.898	0.072	0.283	0.860	3.956	1.189	4.661	0.144	0.142	0.556	0.000
	1250	0.087	0.084	0.330	0.942	0.065	0.254	0.934	3.860	1.385	5.427	0.214	0.125	0.488	0.004

Table 4: No-interaction NIE comparison, increasing sparsity with moderate minimum active signal strength DGP, $\sigma^2 = 1$: our debiased estimator vs. Guo et al. (2022). $s = 5, 10, 15, 20$ for $p + q = 100, 800, 1500, 2500$. Averaged over 500 replications, $K_1 = K_2 = 2.75$.

minimal signal		$\sigma^2 = 1$													
Estimators		Debiasing (ours)							(Guo et al., 2022)						
		Monte Carlo CI				Analytical CI			Monte Carlo CI				Analytical CI		
$p+q$	n	RMSE	SD	CI Len	Cov	SE	CI Len	Cov	RMSE	SD	CI Len	Cov	SE	CI Len	Cov
100	250	0.087	0.076	0.298	0.916	0.098	0.385	0.972	0.078	0.078	0.306	0.946	0.022	0.087	0.400
	500	0.084	0.077	0.301	0.922	0.085	0.331	0.942	0.067	0.067	0.263	0.958	0.015	0.058	0.320
	750	0.074	0.069	0.269	0.934	0.077	0.304	0.958	0.063	0.063	0.246	0.948	0.014	0.055	0.356
	1000	0.068	0.062	0.245	0.922	0.071	0.279	0.954	0.054	0.053	0.206	0.958	0.015	0.060	0.472
	1250	0.065	0.061	0.237	0.944	0.052	0.204	0.974	0.062	0.051	0.202	0.926	0.017	0.068	0.362
800	250	0.113	0.111	0.436	0.956	0.100	0.392	0.922	0.167	0.091	0.355	0.724	0.053	0.207	0.332
	500	0.087	0.086	0.339	0.944	0.082	0.321	0.930	0.219	0.102	0.399	0.576	0.042	0.166	0.046
	750	0.075	0.075	0.295	0.958	0.080	0.312	0.970	0.305	0.160	0.628	0.736	0.038	0.150	0.014
	1000	0.074	0.074	0.291	0.958	0.070	0.273	0.938	0.372	0.172	0.674	0.586	0.036	0.143	0.004
	1250	0.067	0.066	0.259	0.948	0.061	0.238	0.922	0.470	0.219	0.858	0.602	0.035	0.139	0.010
1500	250	0.133	0.123	0.484	0.936	0.100	0.393	0.872	0.148	0.102	0.399	0.842	0.056	0.219	0.606
	500	0.126	0.092	0.362	0.848	0.084	0.329	0.802	0.226	0.140	0.548	0.804	0.046	0.182	0.254
	750	0.085	0.079	0.311	0.938	0.076	0.298	0.924	0.338	0.177	0.694	0.704	0.044	0.173	0.048
	1000	0.081	0.074	0.292	0.916	0.071	0.279	0.910	0.406	0.189	0.741	0.594	0.041	0.161	0.014
	1250	0.068	0.067	0.264	0.950	0.061	0.237	0.926	0.512	0.217	0.849	0.528	0.040	0.156	0.010
2500	250	0.157	0.139	0.544	0.928	0.104	0.410	0.816	2.056	0.722	2.829	0.284	0.173	0.680	0.000
	500	0.125	0.116	0.455	0.934	0.089	0.347	0.838	2.548	0.484	1.897	0.012	0.130	0.510	0.000
	750	0.101	0.090	0.351	0.914	0.080	0.313	0.878	2.593	0.497	1.947	0.034	0.114	0.446	0.000
	1000	0.079	0.077	0.303	0.954	0.071	0.280	0.880	2.589	0.526	2.061	0.042	0.098	0.385	0.000
	1250	0.084	0.084	0.328	0.950	0.064	0.250	0.920	2.528	0.652	2.557	0.086	0.087	0.340	0.000

Table 5: No-interaction NIE comparison, increasing sparsity with vanishing minimum active signal strength DGP, $\sigma^2 = 1$: our debiased estimator vs. Guo et al. (2022). Sparsity fixed at $s = 3$. Averaged over 500 replications, $K_1 = K_2 = 2.75$.

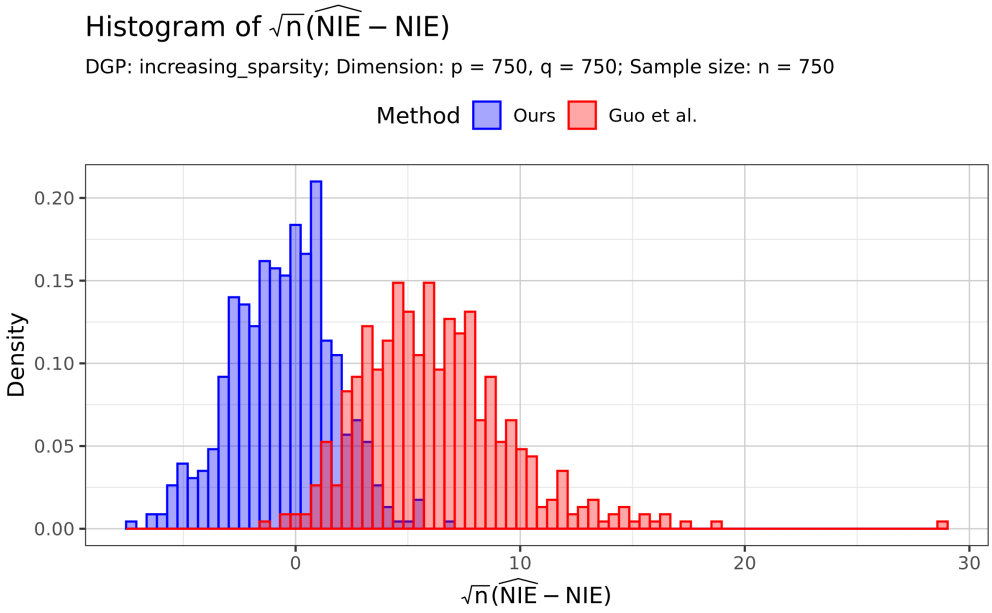


Figure 3: Comparison with Guo et al. (2022)

D.2 Relation to Rakshit and Guo (2024)

We next elaborate on the similarity and distinction between our work and Rakshit and Guo (2024). The Poisson-link mediation model studied by Rakshit and Guo (2024) is more general than the linear model in terms of the outcome link function. Thus, our contribution should not be interpreted as a generalization in the sense of moving from a linear model to a nonlinear exponential-link model. Rather, our work extends the high-dimensional mediation literature along a different modeling dimension: we allow ultra-high-dimensional covariates X and ultra-high-dimensional mediators M, along with A–X and A–M interactions in the outcome model for Y, as well as A–X interactions in the mediator model. Rakshit and Guo (2024) makes an important contribution by developing inference for high-dimensional sparse Poisson regression, with an application to mediation analysis for count outcomes and, in particular, testing treatment–mediator interactions. In their mediation formulation, the outcome model can be written as

$$g\big(\mathbb{E}[Y \mid A, M]\big) = \alpha_0 + \alpha_1 A + M^\top \beta_0 + A M^\top \beta_1 = \alpha_0 + \alpha_1 A + (\beta_0^\top, \beta_1^\top) \begin{pmatrix} M \\ AM \end{pmatrix},$$

where g is the Poisson GLM link function. Thus, the treatment–mediator interaction is encoded by the coefficient β_1. However, they do not include UHD covariates, and consequently their model does not allow for A–X interactions in either the mediator-generation or outcome-generation model. This is precisely why the structural problem considered in our paper is fundamentally different from that of Rakshit and Guo (2024). In particular, our work extends this line of work by allowing UHD covariates and UHD mediators simultaneously, together with their interactions with the treatment variable, although we focus on a linear link rather than a Poisson link. We have also elaborated in Section D.1 that the simultaneous presence of UHD mediators and covariates, together with their interactions with the treatment, makes estimation of the mediation functional substantially more involved, since the target depends on cross-arm nuisance parameters involving both a UHD matrix and a UHD vector.

D.3 Relation to other methods on high-dimensional mediation analysis

We also discuss several other recent works in the high-dimensional mediation analysis literature, including (Jones et al., 2025; Chakrabortty et al., 2018; Shi and Li, 2022; Loh et al., 2022). These papers make

important contributions to the broader literature, but their settings, parameters of interest, or inferential goals are different from those considered in this paper.

First, (Jones et al., 2025) studies a similar mediation analysis problem from the perspective of post-selection inference. Their target parameter is related to the mediation functional considered here, and their modeling strategy is more general in some respects. However, their paper does not consider interactions with treatment and, more importantly, focuses on *fixed-dimensional mediators*. We therefore view (Jones et al., 2025) as addressing an interesting complementary setting rather than being directly comparable to the interaction-rich UHD mediation problem studied in this paper.

Next, Chakrabortty et al. (2018) and Shi and Li (2022) study important high-dimensional mediation analysis problems under a linear structural equation model (LSEM) over a causal graph. In a simplified notation, their model can be written as

$$V - \mu = \mathbf{G}^\top (V - \mu) + \varepsilon,$$

where V collects the treatment, covariates, mediators, and outcome, and $G_{ij} \neq 0$ encodes a directed edge from V_i to V_j. Equivalently, each node is generated as a linear combination of its parent nodes plus an additive error:

$$V_j - \mu_j = \textstyle\sum_{k \in \mathrm{pa}(j)} G_{kj}(V_k - \mu_k) + \varepsilon_j.$$

Mapping this framework to our notation, the corresponding mediator and outcome models are of the form

$$\begin{aligned} Y &= \alpha_0 + \alpha_A A + X^\top \beta + M^\top \gamma + \varepsilon, \\ M &= \delta_0 + \delta_A A + B_X X + HM + U, \end{aligned}$$

where H may encode directed relationships among mediators. Thus, after absorbing constants and coefficients, this falls into the no-treatment-interaction structure discussed in Section D.1. In particular, there are no A–X interactions in the mediator model and no A–X or A–M interactions in the outcome model. As discussed extensively in Section D.1, in such a no-interaction model, the mediation functional, and consequently the NDE and NIE, have a substantially simpler structure and can be reduced to a standard high-dimensional regression/debiasing problem. In contrast, our model requires a multi-step debiasing strategy due to the presence of composite cross-arm nuisance terms such as $\bar{X}^\top B_0^\top \gamma_1$. Furthermore, the primary goals of Chakrabortty et al. (2018) and Shi and Li (2022) are *causal discovery* and the estimation of *individual mediation effects*. For mediation analysis, they focus on controlled direct and indirect effects, which are fundamentally different from the natural direct and indirect effects considered in this paper. Finally, Chakrabortty et al. (2018) allows $p+q$ to grow only polynomially with n, whereas our methodology permits $p+q$ to be exponentially large in n.

Loh et al. (2022) proposes an outcome-regression-based averaging approach for estimating interventional direct and indirect effects in the presence of high-dimensional mediators. Our work differs from theirs in several important respects. First and foremost, Loh et al. (2022) does not provide any theoretical guarantee for their estimation procedure. Second, the causal targets are different: interventional direct and indirect effects are distinct from the natural direct and indirect effects considered in this paper; see VanderWeele et al. (2014) for a detailed discussion of this distinction. Third, their procedure fits outcome-regression models after high-dimensional variable selection and averages the resulting imputed potential outcomes over estimated counterfactual mediator distributions using Monte Carlo integration. In ultra-high-dimensional regimes, where $\dim(M) \gg n$, such regression-based plug-in estimators can suffer from substantial regularization bias unless an additional debiasing step is used; this is precisely one of the main issues addressed by our method. Thus, while Loh et al. (2022) makes an important contribution to estimating interventional direct and indirect effects with high-dimensional mediators, their parameter of interest, methodology, and focus differ from ours.

D.4 Relation to Existing Debiased Inference for High-Dimensional Functionals

In this section, we elaborate on the relationship between our proposed debiasing strategy and the existing literature on debiased inference for high-dimensional functionals. The estimation of debiased inference for linear functionals of high-dimensional regression parameters has a long history, dating back to the seminal works of (Javanmard and Montanari, 2014; Van de Geer et al., 2014; Zhang and Zhang, 2014). More recently Cai et al. (2021) develop debiasing techniques for linear functionals $x^\top(\beta_1 - \beta_2)$ where β_1 and β_2 comes from

two regression model, and Guo et al. (2019) developed debiasing techniques for quadratic/bilinear forms $\beta^\top\gamma$ in high-dimensional linear models. Although these results are highly relevant to the broader literature on high-dimensional inference, the mediation functional considered in our paper is structurally different and cannot be reduced directly to either of these canonical settings. To see the difficulty, consider the mediation functional under our model, suppressing intercepts for simplicity:

$$\theta_0 = \bar{X}^\top(\beta_1 + \mathbf{B}_0^\top\gamma_1).$$

The key nonstandard component is the cross-arm nuisance component $\bar{X}^\top\mathbf{B}_0^\top\gamma_1$, where $\mathbf{B}_0$ is identified from the mediator model in the control arm and γ_1 is identified from the outcome model in the treated arm. There are two natural ways to relate this term to existing debiasing results:

$$\bar{X}^\top\mathbf{B}_0^\top\gamma_1 = (\mathbf{B}_0\bar{X})^\top\gamma_1 \qquad \text{or} \qquad \bar{X}^\top\mathbf{B}_0^\top\gamma_1 = \bar{X}^\top(\mathbf{B}_0^\top\gamma_1).$$

We explain why neither route gives an immediate application of existing linear- or bilinear-functional debiasing methods.

First, consider the representation $\bar{X}^\top\mathbf{B}_0^\top\gamma_1 = (\mathbf{B}_0\bar{X})^\top\gamma_1$. This representation views the target component as an inner product between two high-dimensional vectors, $\mathbf{B}_0\bar{X}$ and γ_1, and therefore appears related to the inner-product debiasing method of Guo et al. (2019). However, their framework requires both vectors in the inner product to be sparse regression coefficient vectors arising from two high-dimensional linear models, which can be estimated consistently with respect to ℓ_1/ℓ_2 norm at a standard Lasso rate. In our case, γ_1 is sparse, but $\mathbf{B}_0\bar{X}$ need not be sparse at all. Indeed, $\mathbf{B}_0\bar{X}$ is not a regression coefficient; it is a derived q-dimensional quantity, with $q \gg n$ in our setup, obtained by multiplying the unknown mediator coefficient matrix $\mathbf{B}_0$ by the sample mean $\bar{X}$. Row-sparsity of $\mathbf{B}_0$ does not imply sparsity of $\mathbf{B}_0\bar{X}$. For example, suppose each row of $\mathbf{B}_0$ has only one nonzero entry, say

$$\mathbf{B}_{0,j1} = 1, \qquad \mathbf{B}_{0,j\ell} = 0 \text{ for all } \ell \neq 1, \qquad j = 1, \ldots, q.$$

Then $\mathbf{B}_0\bar{X} = \bar{X}_1\mathbf{1}_q$, which is dense whenever $\bar{X}_1 \neq 0$, even though each row of $\mathbf{B}_0$ is only 1-sparse. More generally, since $\bar{X}$ is typically dense, the inner products $(\mathbf{B}_0\bar{X})_j = \langle\mathbf{B}_{0,j\cdot}, \bar{X}\rangle$ will generally be nonzero for many, if not all, j's. Although each coordinate $(\mathbf{B}_0\bar{X})_j = \langle\mathbf{B}_{0,j\cdot}, \bar{X}\rangle$ could in principle be estimated by debiasing the row-wise regression of M_j on X in the direction of $\bar{X}$, this would only provide coordinate-wise control. The entire debiased vector $\mathbf{B}_0\bar{X}$ can not be consistently estimated with respect to ℓ_1/ℓ_2 norm as $q \gg n$. Thus, the inner-product debiasing theory for two estimable high-dimensional regression vectors does not apply directly.

Second, one may instead write $\bar{X}^\top\mathbf{B}_0^\top\gamma_1 = \bar{X}^\top b_0$, where $b_0 = \mathbf{B}_0^\top\gamma_1$. If b_0 were the coefficient vector of a single observed high-dimensional linear regression, then standard debiasing methods for linear functionals (Zhang and Zhang, 2014; Van de Geer et al., 2014; Javanmard and Montanari, 2014) would be directly applicable. The key difficulty is that b_0 is not an observed regression coefficient from a single model. It is a composite parameter: the mediator relationship represented by $\mathbf{B}_0$ is associated with the control arm, whereas γ_1 is learned from the treated-arm outcome model. If γ_1 were known, one could regress $M^\top\gamma_1$ on X in the control arm to directly estimate b_0 and then debias in the direction $\bar{X}$. In reality, however, γ_1 is unknown and must be estimated from the treated sample. Replacing γ_1 by $\hat{\gamma}_1$ introduces an additional first-stage estimation error into the control-arm regression. The resulting bias is not present in the standard single-regression linear functional setting. The product-form bias decomposition in Proposition 4.1 shows how the estimation errors from these different nuisance components interact and why additional debiasing is needed. Our multi-step debiasing procedure is designed precisely to address this cross-arm bias propagation.

In summary, our parameter of inference is neither a linear functional of a single sparse regression coefficient nor a bilinear functional of two separately estimable sparse regression coefficients. Rather, it contains a cross-arm and cross-equation component whose relevant parts must be learned and carefully coupled to remove the effect of estimation/regularization in the first order. This structural feature is absent from standard high-dimensional functional inference frameworks and necessitates the multi-step debiasing strategy developed in this paper.

D.5 Relation to Riesz-representer and balancing-based debiasing methods

Our construction is also related to the literature on Riesz-representer and balancing-based approaches to debiasing (Chernozhukov et al., 2022a; Hirshberg and Wager, 2021; Ben-Michael et al., 2021) and to Athey

et al. (2018) for the estimation of ATE in the presence of UHD covariates. In particular, the optimization problems used to compute the weights τ_1 and τ_2 in Steps 1 and 3 of Algorithm 1 are closely related to this line of work; they can be viewed as approximate balancing for estimating the relevant Riesz representers. However, the connection with Riesz regression appears only at the level of the individual linear-functional debiasing subproblems. The full mediation-functional estimation problem, however, is not a standard linear-functional debiasing problem. The key difficulty is the composite term $\bar{X}^\top \mathbf{B}_0^\top \gamma_1$ where the relevant information of B_0 must be learned from the control-arm mediator-covariate relationship, whereas γ_1 must be learned from the treated-arm outcome model, and $\bar{X}$ is computed using observations from both treatment arms. If $\mathbf{B}_0\bar{X}$ were known, one could debias the treated-arm outcome regression in the fixed direction $(\bar{X}^\top, (\mathbf{B}_0\bar{X})^\top)^\top$. If γ_1 were known, one could regress $M^\top\gamma_1$ on X in the control arm and debias the resulting linear functional in the direction $\bar{X}$. However, neither object is known in our setting. Moreover, $\mathbf{B}_0\bar{X}$ can be completely dense and is not generally consistently estimable with respect to the standard ℓ_1/ℓ_2 norm. Consequently, the debiasing directions in the subsequent regressions are themselves estimated from different treatment arms, which introduces additional regularization and plug-in biases that must be carefully removed to obtain a $\sqrt{n}$-CAN estimate of θ_{10}.

The proposed procedure (Algorithm 1) does not estimate the full matrix $\mathbf{B}_0$. Instead, the control-side balancing program directly constructs $\hat{m}_c = \mathbf{M}_c^\top \hat{\tau}_2$ as an approximation of $\mathbf{B}_0\bar{X}$. This estimated quantity enters the direction used in the treated-arm debiasing problem. The procedure subsequently regresses the pseudo-outcome $M^\top\hat{\gamma}_1$ on X in the control arm and reuses the same weight $\hat{\tau}_2$ in the control-side correction. Thus, although the individual optimization problems defining $\hat{\tau}_1$ and $\hat{\tau}_2$ are closely related to standard Riesz-representer or approximate-balancing constructions, these weights are embedded in a coupled multi-step procedure in which both a debiasing direction and a pseudo-outcome contain estimated high-dimensional quantities. Controlling the resulting cross-arm bias propagation and establishing the product-form remainder in Proposition 4.1 are the main roles of our theoretical analysis. Standard Riesz regression or linear/bilinear functional debiasing tools do not directly yield this result; see Section D.4 for further discussion.

The relationship with Liu and Ghassami (2024) is complementary. Their approach also starts from the mediation influence-function representation and uses balancing ideas to stabilize estimation of the nuisance components that enter the leading bias terms. By contrast, we study a sparse UHD linear structural model in which both X and M may substantially exceed the sample size and treatment interactions are present in both the mediator and outcome models. In this setting, the main difficulty is not only instability associated with inverse-probability-type weighting, but also the regularization and cross-arm plug-in biases arising from the estimated direction $\hat{m}_c$ and the estimated pseudo-outcome $M^\top\hat{\gamma}_1$. Thus, although the balancing steps in our method are analogous in spirit to those in Liu and Ghassami (2024), they play a different technical role. Rather than estimating nuisance functions for a fixed influence-function estimator and plugging them into the influence-function expression, our balancing weights are embedded in a multi-step debiasing procedure for penalized high-dimensional regressions. Establishing that the resulting product-form remainder is $o_p(n^{-1/2})$ under sparse UHD assumptions is a principal component of our analysis.

Regarding the connection with Athey et al. (2018), our work extends the high-dimensional debiasing and balancing approach from average treatment effect estimation to causal mediation analysis. Athey et al. (2018) focus on estimating the total effect of a treatment in the presence of high-dimensional covariates. Estimation of the NDE and NIE is more involved because it requires estimating counterfactual mediation functionals rather than only the total effect. In our interaction-rich UHD setting, these functionals contain composite cross-arm nuisance terms such as $\bar{X}^\top \mathbf{B}_0^\top \gamma_1$, which require an estimated debiasing direction, an estimated pseudo-outcome, and coupled treatment- and control-side corrections. Consequently, the proposed construction goes beyond applying a single balancing-weight correction separately to two standard linear functionals.

A nonlinear extension based on Auto-DML (Chernozhukov et al., 2022a) or RieszBoost (Lee and Schuler, 2025) is an interesting but nontrivial direction for future work, rather than a direct consequence of existing theory. In a fully nonparametric mediation model, the target involves the conditional law of the UHD mediator vector M given UHD covariates X, and estimating either this law or the corresponding Riesz representer is generally infeasible without strong structural assumptions. Even if the target can be reduced to a non-linear functional of just the conditional mean $\mathbb{E}[M \mid X, A = 0]$, the problem is still not an immediate application of existing Riesz regression theory. The mediator regression is vector-valued:

$$\nu_0(X) = \mathbb{E}[M \mid X, A = 0] = (\nu_{0,1}(X), \ldots, \nu_{0,q}(X))^\top,$$

$$\nu_{0,j}(x) = \mathbb{E}[M_j \mid X = x, A = 0]\,.$$

The linear construction developed in this paper avoids estimating these q conditional-mean components separately by exploiting the exact identity $\mathbf{M}_c = \mathbf{X}_c B_0^\top + \mathbf{U}_c$ and constructing $\mathbf{M}_c^\top \hat{\tau}_2$ directly. There is no immediate analog of this reduction for a generic nonlinear mediator model. This leads to a growing collection of nuisance regressions, with the number of components of order q, and in our UHD setting $q \gg n$ (can even be exponentially large in n). Existing Auto-DML theory is not directly designed to handle a growing number of nuisance components, especially when that number far exceeds the sample size; the number of nuisance components is always fixed in Chernozhukov et al. (2022a). As a matter of fact, in Example 5 of Chernozhukov et al. (2022a), the authors assumed the mediators are categorical random variables with K different levels, where K is a fixed constant that does not change/grow with n, whereas our model allows the mediators to be discrete or continuous and the number of mediators can be much larger than the sample size. Consequently, a non-linear extension in the UHD regime would require a careful analysis of the interplay between the complexity of UHD nonparametric function classes, an appropriate notion of sparsity (e.g., sparse multi-index or sparse additive structure), and the fact that the ambient dimensions themselves grow with the sample size. The relevant asymptotic framework would no longer be the classical fixed-dimensional setting in which the population model is held fixed while $n \uparrow \infty$. Rather, it would be a triangular-array-type regime, where the statistical model, the nuisance parameter spaces, and the ambient dimensions p and q all vary with n. Therefore, new tools would be needed to handle nonparametric estimation (and debiasing) of mediation functionals in this regime, which we view as an interesting direction for future research.

E Phenotype Features

Table 6: Description of the phenotype variables used in the lung cancer data analysis.

Feature Name	Explanation
`gender`	Patient's gender, such as male or female.
`age`	Age of the patient at diagnosis.
`ABSOLUTE_Purity`	Tumor purity score estimated through molecular analysis.
`number_pack_years_smoked`	Number of cigarette pack-years smoked by the patient.
`initial_weight`	Patient's initial body weight at diagnosis.
`radiation_therapy`	Indicator of whether the patient underwent radiation therapy.
`dlco_predictive_percent`	Predicted percentage of the diffusing capacity of the lungs for carbon monoxide.
`intermediate_dimension`	Intermediate dimension of the tumor, corresponding to a size-related tumor measurement.
`longerest_dimension`	Longest dimension of the tumor.
`karnofsky_performance_score`	Performance status score quantifying the patient's ability to perform daily activities.
`anatomic_neoplasm_subdivision`	Anatomic location of the tumor in the lungs.
`histological_type`	Histological classification of the tumor, such as adenocarcinoma or squamous cell carcinoma.
`days_to_collection`	Number of days from diagnosis to sample collection.
`new_tumor_event after_initial_treatment`	Indicator of whether a new tumor event occurred after initial treatment.

Continued on next page

Feature Name	Explanation
`post_bronchodilator fev1_percent`	Predicted percentage of forced expiratory volume in one second after bronchodilator use.
`shortest_dimension`	Shortest dimension of the tumor.
`pathologic_stage`	Pathologic stage of the cancer, such as Stage I or Stage II.
`sample_type`	Type of the sample, such as primary tumor or normal tissue.
`targeted_molecular_therapy`	Indicator of whether the patient received molecularly targeted therapy.
`status`	Survival status of the patient, such as alive or deceased.

F More on Simulation Study and Sensitivity Analysis

This section provides additional simulation results and sensitivity analyses that complement the main simulation study in Section 6.1. We use the same data-generating mechanisms, estimators, and evaluation metrics described there, and report results for the simulation regimes omitted from the main text for brevity. In particular, the additional experiments examine the performance of the proposed estimator under lower noise levels, increasing sparsity, weak nonzero signals, and conditional heteroskedasticity, as well as its sensitivity to the balancing constants K_1 and K_2.

Table 7 reports the results for the base DGP with a smaller noise variance, $\sigma^2 = 0.1$. Table 8 considers the increasing-sparsity regime, where the number of active coefficients grows with the ambient dimension. Table 9 considers the minimal-signal regime, in which a subset of the nonzero coefficients is generated near the detection boundary. Tables 10 and 11 evaluate the estimator under conditional heteroskedasticity at two different overall noise levels. Finally, Tables 12 and 13 examine the sensitivity of the RMSE for K_1, K_2 at two sample sizes, while Table 14 varies K_1 and K_2 separately to assess the robustness of the estimator to their joint choice.

$s = 3,\ k_1 = k_2 = 3$		$\sigma^2 = 0.1$										
Estimators		$\hat{\theta}_0^{\text{cf}}$ (Debiasing)							$\hat{\theta}^{\text{plug-in}}$			
		Monte Carlo CI				Analytical CI			Monte Carlo CI			
$p+q$	n	RMSE	SD	CI Len	Cov	SE	CI Len	Cov	RMSE	SD	CI Len	Cov
100	250	0.720	0.321	1.260	0.500	0.296	1.159	0.416	0.921	0.229	0.899	0.032
	500	0.182	0.168	0.657	0.786	0.176	0.691	0.804	0.661	0.140	0.548	0.008
	750	0.203	0.130	0.510	0.926	0.139	0.546	0.938	0.575	0.106	0.416	0.000
	1000	0.117	0.115	0.450	0.932	0.121	0.475	0.922	0.519	0.090	0.354	0.000
	1250	0.125	0.119	0.465	0.942	0.114	0.446	0.958	0.490	0.083	0.324	0.000
800	250	0.231	0.230	0.901	0.888	0.392	1.538	0.996	0.366	0.213	0.837	0.694
	500	0.176	0.142	0.557	0.892	0.199	0.781	0.978	0.356	0.045	0.175	0.000
	750	0.150	0.125	0.492	0.942	0.144	0.564	0.946	0.352	0.042	0.166	0.000
	1000	0.115	0.107	0.420	0.942	0.126	0.493	0.966	0.302	0.034	0.132	0.000
	1250	0.090	0.089	0.350	0.944	0.106	0.416	0.984	0.274	0.030	0.118	0.000
1500	250	0.337	0.216	0.848	0.772	0.345	1.352	0.988	0.190	0.181	0.709	0.956
	500	0.159	0.144	0.563	0.890	0.197	0.773	0.994	0.286	0.036	0.140	0.000
	750	0.148	0.118	0.463	0.934	0.142	0.556	0.944	0.243	0.028	0.108	0.000
	1000	0.095	0.093	0.363	0.934	0.115	0.451	0.976	0.216	0.023	0.090	0.000
	1250	0.096	0.089	0.351	0.942	0.102	0.402	0.966	0.199	0.020	0.079	0.000
2500	250	0.635	0.294	1.153	0.551	0.436	1.710	0.831	0.263	0.190	0.747	0.843
	500	0.184	0.151	0.592	0.828	0.226	0.885	0.990	0.211	0.024	0.095	0.000
	750	0.183	0.128	0.502	0.890	0.167	0.655	0.940	0.254	0.025	0.098	0.000
	1000	0.110	0.108	0.424	0.930	0.136	0.532	0.988	0.198	0.020	0.078	0.000
	1250	0.106	0.103	0.404	0.954	0.118	0.461	0.960	0.137	0.014	0.054	0.000

Table 7: Simulation results for the debiased ($\hat{\theta}_0^{\text{cf}}$) and naive ($\hat{\theta}^{\text{plug-in}}$) estimators under the base DGP with $\sigma^2 = 0.1$. Numbers are averaged over 500 replications with $K_1 = K_2 = 2.75$.

increasing sparsity		$\sigma^2 = 1$										
Estimators		$\hat{\theta}_0^{\text{cf}}$ (Debiasing)							$\hat{\theta}^{\text{plug-in}}$			
		Monte Carlo CI				Analytical CI			Monte Carlo CI			
$p+q$	n	RMSE	SD	CI Len	Cov	SE	CI Len	Cov	RMSE	SD	CI Len	Cov
100	250	0.456	0.233	0.914	0.605	0.278	1.088	0.741	1.789	0.303	1.189	0.000
	500	0.264	0.173	0.678	0.790	0.191	0.747	0.830	1.162	0.218	0.854	0.000
	750	0.177	0.152	0.595	0.903	0.159	0.624	0.917	0.939	0.170	0.666	0.000
	1000	0.144	0.129	0.505	0.916	0.136	0.534	0.938	0.815	0.136	0.532	0.000
	1250	0.136	0.119	0.466	0.915	0.126	0.494	0.940	0.792	0.134	0.527	0.000
800	250	0.839	0.304	1.192	0.254	0.462	1.809	0.631	1.361	0.508	1.992	0.326
	500	0.341	0.190	0.747	0.683	0.222	0.869	0.787	1.365	0.181	0.709	0.000
	750	0.211	0.151	0.590	0.834	0.182	0.713	0.913	1.570	0.048	0.187	0.000
	1000	0.138	0.131	0.513	0.933	0.159	0.625	0.978	1.606	0.056	0.219	0.000
	1250	0.123	0.118	0.461	0.943	0.143	0.560	0.973	1.495	0.060	0.236	0.000
1500	250	0.601	0.600	2.351	0.933	0.724	2.838	0.967	2.413	0.709	2.778	0.056
	500	0.677	0.336	1.317	0.580	0.458	1.796	0.774	0.796	0.377	1.478	0.540
	750	0.606	0.246	0.963	0.390	0.288	1.129	0.529	2.187	0.048	0.189	0.000
	1000	0.382	0.223	0.873	0.710	0.241	0.945	0.749	2.286	0.035	0.136	0.000
	1250	0.199	0.188	0.738	0.940	0.223	0.874	0.976	2.243	0.031	0.122	0.000
2500	250	0.806	0.669	2.623	0.375	0.790	3.098	0.625	3.734	0.537	2.105	0.000
	500	0.670	0.404	1.583	0.761	0.553	2.167	0.876	2.061	0.560	2.197	0.052
	750	0.699	0.286	1.122	0.875	0.380	1.490	0.875	1.162	0.296	1.162	0.038
	1000	0.483	0.205	0.802	0.810	0.256	1.004	0.819	2.094	0.042	0.164	0.000
	1250	0.318	0.218	0.853	0.810	0.223	0.872	0.917	2.089	0.031	0.123	0.000

Table 8: Simulation results under the increasing-sparsity DGP with $\sigma^2 = 1$. $s = s_B = 5, 10, 15, 20$ for $p + q = 100, 800, 1500, 2500$; Averaged over 500 replications, $K_1 = K_2 = 2.75$.

minimal signal		$\sigma^2 = 1$										
Estimators		$\hat{\theta}_0^{\mathrm{cf}}$ (Debiasing)							$\hat{\theta}^{\mathrm{plug-in}}$			
		Monte Carlo CI				Analytical CI			Monte Carlo CI			
$p+q$	n	RMSE	SD	CI Len	Cov	SE	CI Len	Cov	RMSE	SD	CI Len	Cov
100	250	0.513	0.304	1.193	0.724	0.300	1.176	0.720	0.832	0.236	0.926	0.066
	500	0.165	0.165	0.648	0.844	0.182	0.714	0.860	0.614	0.147	0.576	0.026
	750	0.197	0.139	0.545	0.944	0.147	0.575	0.972	0.555	0.123	0.481	0.010
	1000	0.118	0.117	0.459	0.946	0.127	0.497	0.954	0.484	0.099	0.389	0.000
	1250	0.121	0.118	0.462	0.954	0.117	0.457	0.960	0.434	0.101	0.396	0.020
800	250	0.208	0.206	0.806	0.954	0.336	1.316	1.000	0.421	0.194	0.762	0.490
	500	0.154	0.142	0.557	0.890	0.185	0.725	0.974	0.345	0.048	0.189	0.000
	750	0.149	0.124	0.485	0.936	0.140	0.549	0.938	0.361	0.050	0.195	0.000
	1000	0.113	0.113	0.442	0.940	0.122	0.480	0.972	0.303	0.039	0.152	0.000
	1250	0.097	0.092	0.361	0.956	0.106	0.414	0.976	0.268	0.034	0.135	0.000
1500	250	0.336	0.188	0.735	0.681	0.270	1.057	0.934	0.215	0.139	0.543	0.768
	500	0.145	0.125	0.491	0.918	0.151	0.592	0.956	0.279	0.035	0.139	0.000
	750	0.109	0.100	0.393	0.932	0.118	0.461	0.958	0.228	0.027	0.104	0.000
	1000	0.083	0.083	0.325	0.952	0.099	0.387	0.982	0.200	0.022	0.087	0.000
	1250	0.089	0.080	0.314	0.920	0.087	0.342	0.942	0.181	0.021	0.083	0.000
2500	250	0.267	0.223	0.872	0.904	0.313	1.227	0.980	0.276	0.124	0.486	0.492
	500	0.135	0.132	0.519	0.934	0.171	0.670	0.992	0.194	0.023	0.091	0.000
	750	0.137	0.113	0.445	0.890	0.130	0.510	0.944	0.226	0.025	0.099	0.000
	1000	0.104	0.095	0.370	0.922	0.109	0.427	0.962	0.176	0.018	0.072	0.000
	1250	0.089	0.088	0.344	0.946	0.096	0.376	0.960	0.112	0.012	0.048	0.000

Table 9: Interaction setting, minimal-signal DGP, $\sigma^2 = 1$: debiased ($\hat{\theta}_0^{\mathrm{cf}}$) vs. naive ($\hat{\theta}^{\mathrm{plug-in}}$). $s = k_1 = k_2 = 3$ fixed; half of nonzero coefficients are weak ($\sim n_{\max}^{-1/2}$). Averaged over 500 replications, $K_1 = K_2 = 2.75$.

heteroscedastic noise		$\sigma^2 = 1$										
Estimators		$\hat{\theta}_0^{\mathrm{cf}}$ (Debiasing)							$\hat{\theta}^{\mathrm{plug-in}}$			
		Monte Carlo CI				Analytical CI			Monte Carlo CI			
$p+q$	n	RMSE	SD	CI Len	Cov	SE	CI Len	Cov	RMSE	SD	CI Len	Cov
100	250	0.691	0.337	1.320	0.598	0.305	1.196	0.500	0.972	0.248	0.970	0.024
	500	0.198	0.178	0.698	0.804	0.189	0.740	0.830	0.717	0.160	0.628	0.010
	750	0.214	0.143	0.561	0.912	0.152	0.596	0.924	0.620	0.126	0.495	0.002
	1000	0.129	0.122	0.479	0.934	0.132	0.518	0.926	0.558	0.109	0.426	0.000
	1250	0.133	0.127	0.496	0.932	0.123	0.481	0.946	0.512	0.107	0.421	0.004
800	250	0.238	0.224	0.879	0.942	0.385	1.508	0.996	0.490	0.216	0.845	0.440
	500	0.169	0.147	0.575	0.898	0.205	0.804	0.978	0.442	0.059	0.232	0.000
	750	0.158	0.130	0.510	0.910	0.153	0.600	0.946	0.437	0.056	0.220	0.000
	1000	0.119	0.119	0.466	0.938	0.133	0.521	0.964	0.358	0.042	0.167	0.000
	1250	0.098	0.098	0.384	0.952	0.114	0.447	0.976	0.329	0.040	0.157	0.000
1500	250	0.446	0.258	1.011	0.696	0.375	1.469	0.961	0.247	0.175	0.684	0.834
	500	0.174	0.150	0.586	0.910	0.200	0.783	0.980	0.377	0.046	0.180	0.000
	750	0.168	0.125	0.491	0.850	0.148	0.579	0.914	0.326	0.037	0.145	0.000
	1000	0.098	0.098	0.383	0.954	0.121	0.475	0.986	0.285	0.030	0.119	0.000
	1250	0.104	0.096	0.374	0.928	0.108	0.423	0.960	0.252	0.027	0.106	0.000
2500	250	0.719	0.312	1.221	0.465	0.464	1.820	0.781	0.446	0.189	0.743	0.405
	500	0.170	0.159	0.623	0.772	0.226	0.884	0.992	0.271	0.030	0.118	0.000
	750	0.207	0.131	0.513	0.936	0.170	0.665	0.902	0.328	0.032	0.127	0.000
	1000	0.118	0.114	0.446	0.942	0.138	0.542	0.984	0.260	0.026	0.102	0.000
	1250	0.110	0.108	0.422	0.936	0.122	0.479	0.970	0.178	0.019	0.075	0.000

Table 10: Interaction setting, heteroscedastic-noise DGP, $\sigma^2 = 1$: debiased ($\hat{\theta}_0^{\mathrm{cf}}$) vs. naive ($\hat{\theta}^{\mathrm{plug-in}}$). The noise variance is $\sigma^2 v(X_i, M_i)$ with $\mathbb{E}[v] = 1$; sparsity fixed at $s = k_1 = k_2 = 3$. Averaged over 500 replications, $K_1 = K_2 = 2.75$.

heteroscedastic noise		$\sigma^2 = 0.5$										
Estimators		$\hat{\theta}_0^{\mathrm{cf}}$ (Debiasing)							$\hat{\theta}^{\mathrm{plug-in}}$			
		Monte Carlo CI				Analytical CI			Monte Carlo CI			
$p+q$	n	RMSE	SD	CI Len	Cov	SE	CI Len	Cov	RMSE	SD	CI Len	Cov
100	250	0.716	0.327	1.280	0.526	0.297	1.164	0.432	0.872	0.235	0.919	0.046
	500	0.186	0.169	0.664	0.922	0.179	0.700	0.936	0.596	0.144	0.566	0.032
	750	0.211	0.133	0.521	0.772	0.142	0.558	0.808	0.519	0.112	0.439	0.008
	1000	0.120	0.116	0.455	0.936	0.124	0.485	0.954	0.456	0.095	0.371	0.002
	1250	0.129	0.121	0.474	0.934	0.116	0.455	0.924	0.420	0.091	0.356	0.010
800	250	0.229	0.224	0.876	0.940	0.384	1.504	0.998	0.396	0.214	0.840	0.626
	500	0.174	0.142	0.556	0.888	0.199	0.781	0.974	0.368	0.048	0.188	0.000
	750	0.153	0.126	0.494	0.894	0.146	0.571	0.946	0.356	0.046	0.182	0.000
	1000	0.117	0.111	0.436	0.938	0.127	0.500	0.966	0.297	0.036	0.141	0.000
	1250	0.093	0.092	0.359	0.950	0.108	0.424	0.986	0.268	0.033	0.129	0.000
1500	250	0.365	0.226	0.887	0.755	0.348	1.364	0.976	0.202	0.178	0.697	0.913
	500	0.162	0.144	0.563	0.866	0.195	0.763	0.990	0.314	0.039	0.151	0.000
	750	0.154	0.119	0.468	0.924	0.142	0.557	0.928	0.263	0.030	0.118	0.000
	1000	0.095	0.093	0.364	0.942	0.116	0.454	0.978	0.231	0.025	0.097	0.000
	1250	0.098	0.091	0.357	0.943	0.103	0.405	0.962	0.207	0.022	0.088	0.000
2500	250	0.656	0.296	1.159	0.509	0.443	1.735	0.823	0.322	0.188	0.738	0.694
	500	0.178	0.152	0.597	0.808	0.223	0.874	0.992	0.226	0.025	0.099	0.000
	750	0.190	0.128	0.500	0.900	0.167	0.653	0.932	0.270	0.027	0.106	0.000
	1000	0.112	0.109	0.429	0.948	0.136	0.532	0.990	0.210	0.022	0.084	0.000
	1250	0.107	0.104	0.407	0.940	0.119	0.465	0.968	0.141	0.015	0.060	0.000

Table 11: Interaction setting, heteroscedastic-noise DGP, $\sigma^2 = 0.5$: debiased ($\hat{\theta}_0^{\mathrm{cf}}$) vs. naive ($\hat{\theta}^{\mathrm{plug-in}}$). The noise variance is $\sigma^2 v(X_i, M_i)$ with $\mathbb{E}[v] = 1$; sparsity fixed at $s = k_1 = k_2 = 3$. Averaged over 500 replications, $K_1 = K_2 = 2.75$.

$n = 750$	$K_1 = K_2 = K$								
$p+q$	1	1.25	1.5	1.75	2	2.25	2.5	2.75	3
100	0.137	0.134	0.131	0.129	0.127	**0.126**	**0.126**	**0.126**	0.126
800	**0.102**	**0.102**	0.102	**0.101**	0.102	0.102	0.103	0.104	0.104
1500	0.139	0.137	**0.136**	0.136	0.136	0.136	0.136	**0.135**	**0.135**
2500	0.130	0.129	0.129	0.129	0.129	0.128	**0.127**	**0.125**	**0.125**

Table 12: Sensitivity of the RMSE of $\hat{\theta}_0^{\mathrm{cf}}$ to the common balancing constant $K_1 = K_2 = K$ under the base DGP ($s = k_1 = k_2 = 3$, $\sigma^2 = 0.5$), $n = 750$. Averaged over 500 replications; the three smallest values in each row are boldfaced.

$n = 1250$	$K_1 = K_2 = K$								
$p + q$	1	1.25	1.5	1.75	2	2.25	2.5	2.75	3
100	0.124	0.122	0.121	0.121	0.121	**0.120**	0.120	**0.119**	**0.119**
800	0.081	0.079	0.077	0.076	0.075	**0.074**	**0.074**	**0.074**	0.074
1500	0.100	0.099	0.098	0.097	**0.096**	**0.096**	**0.096**	0.096	0.096
2500	0.093	0.093	0.093	0.092	**0.091**	0.091	0.091	**0.090**	**0.090**

Table 13: Sensitivity of the RMSE of $\hat{\theta}_0^{\text{cf}}$ to the common balancing constant $K_1 = K_2 = K$ under the base DGP ($s = k_1 = k_2 = 3$, $\sigma^2 = 0.5$), $n = 1250$. Averaged over 500 replications; the three smallest values in each row are boldfaced.

$K_1 \backslash K_2$	1	1.25	1.5	1.75	2	2.25	2.5	2.75	3
1	0.075	0.074	0.074	0.074	0.073	0.073	0.073	0.073	0.073
1.25	0.075	0.075	0.074	0.074	0.074	0.074	0.074	0.074	0.074
1.5	0.075	0.075	0.074	0.074	0.074	0.074	0.074	0.074	0.074
1.75	0.075	0.075	0.074	0.074	0.074	0.073	0.073	0.073	0.073
2	0.075	0.074	0.074	0.074	0.073	0.073	0.073	0.073	0.073
2.25	0.075	0.074	0.074	0.073	0.073	**0.072**	**0.072**	**0.072**	**0.072**
2.5	0.075	0.074	0.074	0.073	0.073	**0.072**	**0.072**	**0.072**	**0.072**
2.75	0.075	0.074	0.073	0.073	0.073	**0.072**	**0.072**	**0.072**	**0.072**
3	0.075	0.074	0.073	0.073	**0.072**	**0.072**	**0.072**	**0.072**	**0.072**

Table 14: Sensitivity of the RMSE of $\hat{\theta}_0^{\text{cf}}$ to (K_1, K_2) under the base DGP ($s = k_1 = k_2 = 3$, $\sigma^2 = 0.5$), $p + q = 1500$, $n = 1250$. Averaged over 500 replications; the five smallest values are boldfaced.