Journal of Machine Learning Research 27 (2026) 1-39 Submitted 6/25; Revised 9/26; Published 9/26

Efficient Estimation of Average Treatment Effect on the Treated under Endogenous Treatment Assignment

Trinetri Ghosh TRINETRIGHOSH.KOLKATA@GMAIL.COM
US Food & Drug Administration
10903 New Hampshire Avenue
Silver Spring, MD 20993, USA

Jiawei Shan JIAWEI.SHAN@WISC.EDU
Departments of Statistics and of Biostatistics & Medical Informatics
University of Wisconsin-Madison
Madison, WI 53706, USA

Menggang Yu MENGGANG@UMICH.EDU
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109, USA

Jiwei Zhao* JIWEI.ZHAO@WISC.EDU
Departments of Statistics and of Biostatistics & Medical Informatics
University of Wisconsin-Madison
Madison, WI 53706, USA

Editor: Jie Peng

Abstract

In this paper, we consider estimation of the average treatment effect on the treated (ATT), an interpretable and relevant causal estimand to policymakers when treatment assignment is endogenous. By considering shadow variables that are unrelated to the treatment assignment but related to the outcomes of interest, we establish identification of the ATT. Then we focus on efficient estimation of the ATT by characterizing the geometric structure of the likelihood, deriving the semiparametric efficiency bound for ATT estimation and proposing an estimator that can achieve this bound. We rigorously establish the theoretical results of the proposed estimator. The finite sample performance of the proposed estimator is studied through comprehensive simulation studies as well as an application to our motivating study.

Keywords: Causal inference, endogeneity, identifiability, semiparametric efficiency, influence function

1. Introduction

Among many causal estimands, the most well studied is the average treatment effect (ATE), which captures the average difference in potential outcomes between treatment and control for all study subjects (Imbens and Rubin, 2015; Hernán and Robins, 2022; Morgan and

*. The authors are listed in alphabetical order. This work was completed while Trinetri Ghosh was affiliated with the University of Wisconsin-Madison.

©2026 Trinetri Ghosh, Jiawei Shan, Menggang Yu, Jiwei Zhao.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/25-1271.html.

Winship, 2014). Another popular estimand is the average treatment effect on the treated (ATT) (Imbens and Rubin, 2015; Hernán and Robins, 2022). Heckman and Smith (1995) and Heckman and Vytlacil (2001) argued that, in many economics and program evaluation studies, ATT might be the implicit quantity sought after, hence more relevant and more interpretable to policymakers. For example, when evaluating whether a smoking cessation campaign intervention implemented by the government can decrease the smoking prevalence in a city, researchers and policymakers might be more interested in the effect of the intervention among those who actually received the campaign (i.e., ATT) but not on those who never received it (Wang et al., 2017). Similar situations also arise in the medical field where a major interest is to assess the treatment effect on actual participants. For instance, patients who agreed to take a new drug with a recommended higher dose, instead of patients who did not, would wonder about the benefit of the higher dose. It is also the case in our motivating study, with details presented in Section 6, where the goal is to assess the effect of the stereotactic body radiation therapy (SBRT) among patients with non-small-cell lung cancer (NSCLC). Radiation oncologists would be more interested in the treatment effect on those who actually received SBRT, but not on those for whom the SBRT was never intended.

A key assumption for identifying ATT is strong ignorability of treatment assignment (Rosenbaum and Rubin, 1983), which means that the treatment assignment mechanism is independent of the potential outcomes given the observed covariates or confounders. This assumption is also known as unconfoundedness, selection on observables, or exogeneity (Imbens, 2004; Imbens and Rubin, 2015). Under this assumption, ATT can be identified and estimated based on propensity scores (Rosenbaum and Rubin, 1983; Hahn, 1998; Dehejia and Wahba, 1999; Lechner, 1999; Hirano et al., 2003; McCaffrey et al., 2004; Hainmueller, 2012; Hartman et al., 2015; Pirracchio et al., 2016; Zhao and Percival, 2017; Moodie et al., 2018), regression (Heckman et al., 1997, 1998; Imbens et al., 2005), matching (Cochran and Rubin, 1973; Rubin, 1973a,b; Abadie and Imbens, 2006, 2011), or combinations of these methods (Abadie and Imbens, 2002; Pirracchio and Carone, 2018). However, in observational studies, the unconfoundedness assumption may be frequently violated (Frölich and Melly, 2013; Mebane and Poast, 2013). That is, the differences in outcome measures between treated and untreated may be due to unobserved differences in selecting treatment status that also predict the outcome. Consequently, ATT estimators that rely on unconfoundedness can be substantially biased, and the corresponding estimand may not even be identifiable at all.

To make progress in the absence of exogeneity, researchers have turned to 'helper' variables, in particular, instrumental variables (IVs) that satisfy three key requirements: independence (from unmeasured confounders), exclusion (from dependence with the potential outcomes), and relevance (to the treatment received) (Baiocchi et al., 2014; Clarke and Windmeijer, 2012; Chen and Briesacher, 2011). When such an IV exists, ATT may be properly estimated in the presence of unmeasured confounding, with further modeling assumptions either on the outcome or on the IVs (Wright, 1928; Robins, 1994; Clarke and Windmeijer, 2012; Liu et al., 2020; Imbens and Rubin, 2015; Hernán and Robins, 2022). It is important to emphasize, however, that full nonparametric identification of ATT generally remains impossible even with valid IVs (Pearl, 1995, 2009). Moreover, credible IVs are often difficult to find (Newhouse and McClellan, 1998; Angrist and Krueger, 2001; Hernán

and Robins, 2006; Martens et al., 2006), because there may be many unmeasured confounders and it can be challenging to rule out the IV's dependence on all of them (Chen and Briesacher, 2011). In addition, negative controls (NCs) provide another class of helper variables for identification under unmeasured confounding. This approach typically requires a pair of NCs: a negative control exposure and a negative control outcome, which together can help establish identifiability of causal effects (Miao et al., 2018; Tchetgen Tchetgen et al., 2024).

In this paper, we consider an alternative kind of 'helper' variables, denoted as $\mathbf{Z}$ throughout, that are (i) independent of the treatment assignment conditional on the (potentially) untreated outcome and other observed covariates, and (ii) related to the (potentially) untreated outcome, possibly through unmeasured confounders. We call such variables shadow variables (SVs) due to their link with the nonresponse instrument in the literature of missing data analysis (Wang et al., 2014; Zhao and Shao, 2015; Miao et al., 2016; Zhao and Ma, 2022; Li et al., 2023; Tian et al., 2026; Miao et al., 2024; Shetty et al., 2023, 2025). SVs may be more readily available in the settings where treatment assignment is more proximal or nearer in time to the measurement of SVs. In such settings, we might have more information about the treatment assignment to rule out dependence of treatment assignment on the SVs. For example, in our motivating study for newly diagnosed NSCLC patients, a new radiation therapy, SBRT, is compared to the standard surgery procedure, lobectomy. According to our collaborating radiation oncologists, lobectomy is the standard treatment for most patients and SBRT has been more commonly given to elderly patients who may be at higher risk of surgical complications. In addition, SBRT is among the most cost-effective treatment options for early-stage NSCLC, requiring fewer visits and generally being covered by most insurance types (Shah et al., 2013; Arruda et al., 2022). Therefore, some variables related to social determinants of health (SDOH) should not drive the treatment assignment and can be considered as $\mathbf{Z}$, such as income level (binary, income$<$\$48,000 versus income$\geq$\$48,000) and insurance level (binary, private versus non-private).

Our main contributions are threefold. First, leveraging a valid SV, we establish the nonparametric identifiability of the ATT, encoded as Δ throughout. SVs have been studied in the missing data literature, but, to the best of our knowledge, this paper provides the first result for the identification of a causal estimand with the aid of an SV. The proposed framework provides a practical alternative to IV-based approaches when credible IVs are difficult to obtain but plausible SVs may be more readily available. Second, under a parametric treatment assignment model indexed by $\boldsymbol{\theta}$, we develop a comprehensive semiparametric theory for inference on both $\boldsymbol{\theta}$ and Δ. Because Δ depends on both the parametric component $\boldsymbol{\theta}$ and the nonparametric nuisance component, its efficient influence function (EIF) must account for variation in both. We characterize the nuisance tangent space and its orthogonal complement to obtain the efficient influence functions (Bickel et al., 1993; Tsiatis, 2006). This geometric characterization clarifies how the nuisance components interact and guides the construction of flexible, efficient estimators. We also propose an alternative estimator $\widehat{\Delta}_{\text{alt}}$ that, although less efficient than $\widehat{\Delta}_{\text{eff}}$ in theory, is substantially easier to implement numerically. We elucidate the relationship between these estimators by connecting their influence functions through projection (Tsiatis, 2006). Third, the derived semiparametric theory enables us to jointly construct Neyman-orthogonal scores for $\boldsymbol{\theta}$ and Δ that account for both the first-order effect of estimating $\boldsymbol{\theta}$ on ATT estimation and the de-

pendence of $\boldsymbol{\theta}$ estimation on nuisance functions. This construction permits flexible machine learning methods for nuisance estimation while retaining valid and efficient inference for the target parameters (Chernozhukov et al., 2018). Simulations and a real-data application demonstrate the effectiveness, scalability, and practical value of our proposed methods.

The remainder of the paper is organized as follows. Section 2 introduces the basic framework and provides sufficient conditions for identifying the likelihood and the parameter of interest. Section 3 develops the semiparametric theory and the corresponding efficient estimators; it also introduces a simpler alternative estimator and clarifies its connection to the efficient one. Section 4 describes implementation details and establishes the asymptotic properties of the proposed estimators. Section 5 presents simulation studies evaluating finite-sample performance and comparing our methods with existing approaches. Section 6 applies the proposed method to the NCDB data. Additional technical details are deferred to the Appendix.

2. Preliminaries

We adopt the potential outcome framework, which implicitly requires the stable unit treatment value assumption (SUTVA) (Neyman, 1923; Rubin, 1974, 1990) throughout this article. Let T denote the treatment where $T = 1$ means treatment/intervention and $T = 0$ control. The observed outcome is then $Y = TY_1 + (1 - T)Y_0$, where Y_t denotes the potential outcome under $T = t$, $t \in \{0, 1\}$. In addition, we observe a vector of pre-treatment covariates $\mathbf{X}$.

The traditional strong ignorability assumption requires $T \perp\!\!\!\perp (Y_1, Y_0) \mid \mathbf{X}$ (Rosenbaum and Rubin, 1983). Consequently, $\mathrm{pr}(T = 1 \mid y_1, y_0, \mathbf{x}) = \mathrm{pr}(T = 1 \mid \mathbf{x})$, so in observational studies the non-randomized treatment assignment mechanism can be modeled based on observed covariates alone, say $w(\mathbf{x}) \equiv \mathrm{pr}(T = 1 \mid \mathbf{x})$. When the ignorability assumption fails, one might have to directly model $\mathrm{pr}(T = 1 \mid y_1, y_0, \mathbf{x})$, which involves the potential outcomes. In this article, we instead propose to model the conditional distribution of T given $\mathbf{X}$ and the potential untreated outcome Y_0, denoted as

$$\pi(y_0, \mathbf{x}) \equiv \mathrm{pr}(T = 1 \mid y_0, \mathbf{x}) = \int \mathrm{pr}(T = 1 \mid y_0, y_1, \mathbf{x}) f(y_1 \mid y_0, \mathbf{x}) \mathrm{d}y_1,$$

where $f(y_1 \mid y_0, \mathbf{x})$ is the conditional density of Y_1 given Y_0 and $\mathbf{X}$. In particular, we allow $\mathrm{pr}(T = 1 \mid y_0, y_1, \mathbf{x}) \neq \mathrm{pr}(T = 1 \mid y_0, \mathbf{x}) \neq \mathrm{pr}(T = 1 \mid \mathbf{x})$.

2.1 Identification

Because for each subject, only one of the potential outcomes is observed, violation of the strong ignorability assumption implies that our parameter of interest is not directly identifiable from the likelihood of the observed data, without imposing extra assumptions. Throughout the ATT is defined as

$$\begin{aligned}\Delta &= \mathrm{E}(Y_1 - Y_0 \mid T = 1) \\ &= \frac{\iint y_1 w(\mathbf{x}) f_1(y_1 \mid \mathbf{x}, 1) g(\mathbf{x}) \mathrm{d}y_1 \mathrm{d}\mathbf{x} - \iint y_0 \frac{\pi(y_0, \mathbf{x})}{1 - \pi(y_0, \mathbf{x})} \{1 - w(\mathbf{x})\} f_0(y_0 \mid \mathbf{x}, 0) g(\mathbf{x}) \mathrm{d}y_0 \mathrm{d}\mathbf{x}}{\int w(\mathbf{x}) g(\mathbf{x}) \mathrm{d}\mathbf{x}},\end{aligned} \tag{1}$$

where $f_1(y_1 \mid \mathbf{x}, 1)$ is the conditional density of Y_1 given $\mathbf{X}$ and $T = 1$, $f_0(y_0 \mid \mathbf{x}, 0)$ the conditional density of Y_0 given $\mathbf{X}$ and $T = 0$, and $g(\mathbf{x})$ is the marginal density of $\mathbf{X}$.

Note that the likelihood function from one single observation in this model is

$$\begin{aligned} &\{f_1(y_1 \mid \mathbf{x}, 1)w(\mathbf{x})g(\mathbf{x})\}^t \left[f_0(y_0 \mid \mathbf{x}, 0)\{1 - w(\mathbf{x})\}g(\mathbf{x})\right]^{1-t} \\ &= w(\mathbf{x})^t\{1 - w(\mathbf{x})\}^{1-t} f_1(y_1 \mid \mathbf{x}, 1)^t f_0(y_0 \mid \mathbf{x}, 0)^{1-t} g(\mathbf{x}). \end{aligned} \tag{2}$$

Therefore one can identify $f_1(y_1 \mid \mathbf{x}, 1)$, $f_0(y_0 \mid \mathbf{x}, 0)$, $g(\mathbf{x})$ and $w(\mathbf{x})$ from (2). However $\pi(y_0, \mathbf{x})$ cannot be identified, which plays a critical role in the numerator of Δ from (1). In other words the parameter of interest Δ cannot be identified from (2) without extra assumptions.

To identify $\pi(y_0, \mathbf{x})$, we write the following transformation of the identifiable $w(\mathbf{x})$ as

$$\{1 - w(\mathbf{x})\}^{-1} = \int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1 - \pi(y_0, \mathbf{x})} \mathrm{d}y_0. \tag{3}$$

Hence, $\pi(y_0, \mathbf{x})$ can be identified if, for any two functions $\pi_1(y_0, \mathbf{x})$ and $\pi_2(y_0, \mathbf{x})$,

$$\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1 - \pi_1(y_0, \mathbf{x})} \mathrm{d}y_0 = \int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1 - \pi_2(y_0, \mathbf{x})} \mathrm{d}y_0 \text{ implies } \pi_1(y_0, \mathbf{x}) = \pi_2(y_0, \mathbf{x}).$$

In particular, we assume that $\mathbf{X}$ can be partitioned into two subsets of variables $\mathbf{U}$ and $\mathbf{Z}$, i.e., $\mathbf{X} = (\mathbf{U}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$, such that the following set of sufficient conditions for identification holds:

(i) *Conditional completeness of the potential outcome distribution*:

$$\mathrm{E}_0\{h(Y_0, \mathbf{U}) \mid \mathbf{X}, 0\} = 0 \text{ implies } h(Y_0, \mathbf{U}) = 0 \text{ almost surely}, \tag{4}$$

(ii) *Exclusion restriction from treatment assignment*:

$$\mathrm{pr}(T = 1 \mid y_0, \mathbf{x}) = \mathrm{pr}(T = 1 \mid y_0, \mathbf{u}). \tag{5}$$

The assumption (4) is the well-known completeness condition, popularly used in missing data analysis, instrumental variable models, and econometrics; see, e.g., Newey and Powell (2003); Hu and Shiu (2018); Zhao and Ma (2022). It is a condition imposed on the distribution $f_0(y_0 \mid \mathbf{x}, 0)$. By writing $\mathrm{E}_0\{h(Y_0, \mathbf{U}) \mid \mathbf{X}, 0\} = \mathrm{E}_0\{h(Y_0, \mathbf{U}) \mid \mathbf{U}, \mathbf{Z}, 0\}$, the condition essentially requires the shadow variable $\mathbf{Z}$ to be correlated with Y_0 and to exhibit sufficient variation relative to Y_0. For instance, in the case of categorical Y_0 and $\mathbf{Z}$, it requires $\mathbf{Z}$ to have at least as many categories as Y_0. In our motivating example where Y_0 is binary, we choose $\mathbf{Z}$ to consist of two binary covariates (hence to have four categories) to ensure that this assumption holds. The potential outcome Y_0 can also be continuous. In that case, several commonly used models for $f_0(y_0 \mid \mathbf{x}, 0)$, such as the full exponential families as illustrated in Example 1 below, satisfy the completeness condition (Newey and Powell, 2003; d'Haultfoeuille, 2011).

Example 1 (Full exponential family) *If the conditional distribution of* Y_0 *given* $(\mathbf{X}, T = 0)$ *is absolutely continuous with density* $f_0(y_0 \mid \mathbf{x}, 0) = s(y_0, \mathbf{u})t(\mathbf{x}) \exp\{\mu(\mathbf{x})^{\mathrm{T}}\tau(y_0, \mathbf{u})\}$, *where* $s(y_0, \mathbf{u}) > 0$, $\tau(y_0, \mathbf{u})$ *is one-to-one in* y_0, *and the support of* $\mu(\mathbf{x})$ *given* $\mathbf{u}$ *contains an open set, then the completeness condition* (4) *holds.*

The assumption (5) requires $\mathbf{Z}$ to be independent of treatment assignment conditional on the untreated potential outcome Y_0 and the other covariates $\mathbf{U}$. Practical justification relies on subject-matter knowledge. For example, when a medical intervention or treatment can be conveniently delivered to patients, a patient's access to transportation or distance to the health care center (e.g. rural vs urban) should not be related to the treatment assignment and can be part of $\mathbf{Z}$. Similarly, a cost-effective treatment may be given to patients purely based on toxicity and efficacy trade-off; hence, some socioeconomic variables or educational background variables can be part of $\mathbf{Z}$ as well. Similar assumptions have been motivated by practical applications in the missing-data literature, including income nonresponse (Wang et al., 2014; Shetty et al., 2025), longitudinal dyspnoea data from cotton-factory workers (Zhao and Shao, 2015), children's mental health (Zhao and Ma, 2022), and house-price nonresponse (Miao et al., 2024).

Under assumptions (4) and (5), the model likelihood (2) becomes

$$\left[\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u})}\mathrm{d}y_0 - 1\right]^t \left[\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u})}\mathrm{d}y_0\right]^{-1} f_1(y_1 \mid \mathbf{x}, 1)^t f_0(y_0 \mid \mathbf{x}, 0)^{1-t} g(\mathbf{x}), \qquad (6)$$

and we have the following identifiability result with its proof contained in Appendix A.

Lemma 1 *Under assumptions (4) and (5), the model likelihood (6) is fully identifiable, hence the parameter Δ is identifiable.*

Remark 1 Identification of Δ using shadow variables is conceptually similar to approaches in the missing-data literature (Wang et al., 2014; Shao and Wang, 2016; Zhao and Ma, 2022; Miao et al., 2024). When ATT is the primary parameter of interest, $\mathrm{E}(Y_1 \mid T = 1)$ is inherently identifiable, and $\mathrm{E}(Y_0 \mid T = 1)$ is the key quantity to be identified. Thus, it suffices to identify the missingness propensity for Y_0, i.e., $\mathrm{pr}(T = 1 \mid y_0, \mathbf{x})$, instead of $\mathrm{pr}(T = 1 \mid y_0, y_1, \mathbf{x})$, as discussed above. It is worth noting that although $\mathrm{E}(Y_1 \mid T = 1)$ is identifiable, its EIF differs from that under ignorability (Hahn, 1998), because the likelihood in the current setting, given by (8), differs from that under ignorability. We derive the corresponding EIFs in Section 3.

Remark 2 Instrumental variables (IVs) are another commonly used class of auxiliary variables in the causal inference literature for identifying causal effects in the presence of unmeasured confounding (Angrist et al., 1996; Wang and Tchetgen Tchetgen, 2018). A valid IV is typically required to be exogenous, relevant to the treatment, and to affect the outcome only through the treatment. These conditions often imply that an IV is a pre-treatment, externally generated variable that does not directly influence the outcome. In contrast, a shadow variable (SV) is assumed not to directly affect treatment assignment while being predictive of the outcome. Consequently, an SV can be post-treatment or measured contemporaneously with the treatment, as long as it does not drive the treatment assignment.

The assumption (5) indicates that we can model T based on y_0 and $\mathbf{u}$. Note that identifying $\pi(\cdot)$ from (3) amounts to solving a Fredholm integral equation of the first kind (Kress, 1989). Such problems are notoriously known to be ill-posed, which makes nonparametric

modeling challenging. In the rest of the paper, we therefore adopt a parametric model for T:

$$\mathrm{pr}(T=1 \mid y_0, \mathbf{x}) = \mathrm{pr}(T=1 \mid y_0, \mathbf{u}) = \pi(y_0, \mathbf{u}; \boldsymbol{\theta}), \tag{7}$$

where the form of $\pi(\cdot)$ is known and $\boldsymbol{\theta}$ is an unknown d-dimensional parameter. Parametric propensity models are widely used in causal inference (Rosenbaum and Rubin, 1983; Imai and Ratkovic, 2014) and nonignorable missing-data problems (Wang et al., 2014; Morikawa and Kim, 2021). We focus on the corresponding efficient estimation of both $\boldsymbol{\theta}$ and Δ, and defer a discussion of the purely nonparametric approach to Appendix E.

In our motivating example, we adopt the logistic regression model

$$\pi(y_0, \mathbf{u}; \boldsymbol{\theta}) = \mathrm{expit}(\theta_1 + \theta_2 y_0 + \boldsymbol{\theta}_3^{\mathrm{T}} \mathbf{u}),$$

and the estimates for θ_2 are highly significant (see Table 4 of Section 6). Indeed, this dependence of the treatment assignment on the potential untreated outcome Y_0 indicates that the traditional strong ignorability assumption may not be sensible in this study.

3. Estimation

Throughout the paper, when there is no ambiguity, we abbreviate $\pi(y_0, \mathbf{u}; \boldsymbol{\theta})$ as $\pi(\boldsymbol{\theta})$ and $\pi(y_{0i}, \mathbf{u}_i; \boldsymbol{\theta})$ as $\pi_i(\boldsymbol{\theta})$. We denote $\partial \pi(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$ by $\dot{\boldsymbol{\pi}}(\boldsymbol{\theta})$. We denote the joint density of Y_1, Y_0 given $\mathbf{X}$ by $f(y_1, y_0 \mid \mathbf{x})$; therefore, $f_1(y_1 \mid \mathbf{x}) = \int f(y_1, y_0 \mid \mathbf{x}) \mathrm{d}y_0$ and $f_0(y_0 \mid \mathbf{x}) = \int f(y_1, y_0 \mid \mathbf{x}) \mathrm{d}y_1$. We write the conditional expectation with respect to $f_0(y_0 \mid \mathbf{x})$ as $\mathrm{E}_0(\cdot \mid \mathbf{x})$ and that with respect to $f_0(y_0 \mid \mathbf{x}, 0)$ as $\mathrm{E}_0(\cdot \mid \mathbf{x}, 0)$. Similar notation also applies to $\mathrm{E}_1(\cdot \mid \mathbf{x})$ and $\mathrm{E}_1(\cdot \mid \mathbf{x}, 1)$. We write $\mathbf{W} = (T, Y, \mathbf{X}^{\mathrm{T}})^{\mathrm{T}}$, $p = \mathrm{pr}(T=1)$, and specify the true value of unknown parameters with superscript ^{0} whenever needed, such as $\boldsymbol{\theta}^0$ and Δ^0. For clarity of presentation, we also adopt the simplified notations summarized in Table 1. These notations leverage the identity $\mathrm{E}_0\{\mathbf{a}(Y_0, \mathbf{x}) \mid \mathbf{x}\} = \{1 - w(\mathbf{x})\}\mathrm{E}_0[\{1 - \pi(\boldsymbol{\theta}^0)\}^{-1}\mathbf{a}(Y_0, \mathbf{x}) \mid \mathbf{x}, 0]$ for any $\mathbf{a}(y_0, \mathbf{x})$, which allows us to identify quantities of the form $\mathrm{E}_0(\cdot \mid \mathbf{x})$.

In this section, our main goal is to derive the efficient influence functions for $(\boldsymbol{\theta}^0, \Delta^0)$ under model assumptions (4)–(5), which in turn motivate an efficient estimation approach for the parameters of interest. In semiparametric theory, an estimator $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\mathbf{w}_1, \ldots, \mathbf{w}_n)$ of $\boldsymbol{\beta}^0$ is called regular and asymptotically linear (RAL) if

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) = n^{-1/2} \sum_{i=1}^{n} \boldsymbol{\psi}(\mathbf{w}_i) + o_p(1),$$

where $\boldsymbol{\psi}(\cdot)$ is a mean-zero function, referred to as the influence function of $\widehat{\boldsymbol{\beta}}$. Under mild regularity conditions, the central limit theorem implies that $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^0) \xrightarrow{d} N(\mathbf{0}, \mathrm{E}(\boldsymbol{\psi}\boldsymbol{\psi}^{\mathrm{T}}))$, provided that $\mathrm{E}(\boldsymbol{\psi}\boldsymbol{\psi}^{\mathrm{T}})$ is finite and nonsingular. Among all RAL estimators of $\boldsymbol{\beta}^0$, the one with the smallest asymptotic variance has the influence function that is the efficient influence function (EIF), denoted as $\boldsymbol{\psi}_{\mathrm{eff}}$, and the corresponding semiparametric efficiency bound is $\mathrm{E}\left(\boldsymbol{\psi}_{\mathrm{eff}}\boldsymbol{\psi}_{\mathrm{eff}}^{\mathrm{T}}\right)$ (Bickel et al., 1993; Tsiatis, 2006).

Table 1: Simplified notations used throughout the paper

Notation	Definition
$w(\mathbf{x})$	$1-\mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1} \mid \mathbf{x}, 0]^{-1}$
$\mathbf{A}(\mathbf{x})$	$\mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0) \mid \mathbf{x}] = \{1-w(\mathbf{x})\}\mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-2}\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0) \mid \mathbf{x}, 0]$
$B(\mathbf{x})$	$\mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\pi(\boldsymbol{\theta}^0) \mid \mathbf{x}] = \{1-w(\mathbf{x})\}\mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-2}\pi(\boldsymbol{\theta}^0) \mid \mathbf{x}, 0]$
$C(\mathbf{x})$	$\mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\pi(\boldsymbol{\theta}^0)^2 Y_0 \mid \mathbf{x}] = \{1-w(\mathbf{x})\}\mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-2}\pi(\boldsymbol{\theta}^0)^2 Y_0 \mid \mathbf{x}, 0]$
$V(\mathbf{x})$	$p\mathrm{E}\left\{\phi_{\mathrm{alt}}(\mathbf{w}; \boldsymbol{\theta}^0, \Delta^0) \mid \mathbf{x}\right\} + \mathrm{E}_0[Y_0\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\pi(\boldsymbol{\theta}^0) \mid \mathbf{x}]$ $= w(\mathbf{x})\{\mathrm{E}_1(Y_1 \mid \mathbf{x}, 1) - \Delta^0\} + \{1-w(\mathbf{x})\}\mathrm{E}_0[Y_0\{1-\pi(\boldsymbol{\theta}^0)\}^{-2}\pi(\boldsymbol{\theta}^0)^2 \mid \mathbf{x}, 0]$
$\mathbf{M}$	$\mathrm{E}\{B(\mathbf{x})^{-1}\mathbf{A}(\mathbf{x})\mathbf{A}(\mathbf{x})^{\mathrm{T}}\}$
$\mathbf{Q}$	$\mathrm{E}[Y_0\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)]$
$\mathbf{D}$	$\mathrm{E}\{B(\mathbf{x})^{-1}V(\mathbf{x})\mathbf{A}(\mathbf{x})\}$

Note: $\phi_{\mathrm{alt}}(\mathbf{w}; \boldsymbol{\theta}, \Delta)$ is defined in (15).

3.1 Efficient Estimation of Δ

Under the assumptions (4) and (7), the model likelihood is

$$\left[\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u}; \boldsymbol{\theta})}\mathrm{d}y_0 - 1\right]^t \left[\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u}; \boldsymbol{\theta})}\mathrm{d}y_0\right]^{-1} f_1(y_1 \mid \mathbf{x}, 1)^t f_0(y_0 \mid \mathbf{x}, 0)^{1-t} g(\mathbf{x}), \quad (8)$$

which contains three nonparametric nuisance parts $f_1(y_1 \mid \mathbf{x}, 1)$, $f_0(y_0 \mid \mathbf{x}, 0)$, $g(\mathbf{x})$, and one d-dimensional parameter $\boldsymbol{\theta}$. We consider the Hilbert space $\mathcal{H}$ of all d-dimensional zero-mean measurable functions of the observed data with finite variance, equipped with the inner product $\langle \mathbf{h}_1, \mathbf{h}_2 \rangle = \mathrm{E}\{\mathbf{h}_1(\cdot)^{\mathrm{T}}\mathbf{h}_2(\cdot)\}$, where $\mathbf{h}_1, \mathbf{h}_2 \in \mathcal{H}$. We first derive the nuisance tangent space and its orthogonal complement, where the nuisance tangent space is defined as the mean squared closure of the nuisance tangent spaces of parametric submodels spanned by the nuisance score vectors. We have the following result whose proof can be found in Appendix B.

Proposition 1 *The space $\mathcal{H}$ can be decomposed as*

$$\mathcal{H} = \Lambda_0 \oplus \Lambda_1 \oplus \Lambda_2 \oplus \Lambda^{\perp},$$

where

$$\Lambda_0 = \left[(1-t)\mathbf{c}(y_0, \mathbf{x}) + \frac{t-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0\{\mathbf{c}(y_0, \mathbf{x}) \mid \mathbf{x}\} : \mathbf{c}(y_0, \mathbf{x}) \in \mathbb{R}^d, \mathrm{E}_0\{\mathbf{c}(y_0, \mathbf{x}) \mid \mathbf{x}, 0\} = \mathbf{0}\right],$$

$$\Lambda_1 = \left[t\mathbf{b}(y_1, \mathbf{x}) \in \mathbb{R}^d : \mathrm{E}_1\{\mathbf{b}(y_1, \mathbf{x}) \mid \mathbf{x}, 1\} = \mathbf{0}\right],$$

$$\Lambda_2 = \left[\mathbf{a}(\mathbf{x}) \in \mathbb{R}^d : \mathrm{E}\{\mathbf{a}(\mathbf{x})\} = \mathbf{0}\right],$$

are the nuisance tangent spaces with respect to $f_0(y_0 \mid \mathbf{x}, 0)$, $f_1(y_1 \mid \mathbf{x}, 1)$, and $g(\mathbf{x})$, respectively, and the orthogonal complement is

$$\Lambda^{\perp} = \left\{\frac{t-\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\mathbf{g}(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \in \mathbb{R}^d\right\}.$$

The notation $\oplus$ represents the direct sum of two spaces that are orthogonal to each other.

The tangent spaces presented in Proposition 1 depend on the parameterization of $\pi(y_0, \mathbf{u})$ and would differ under a nonparametric specification. Proposition 1 indicates that any influence function for estimating Δ can be written as the summation of four pairwise orthogonal elements in spaces Λ_2, Λ_1, Λ_0 and $\Lambda^\perp$, respectively. In particular we study the efficient influence function, which must be of the form

$$\begin{aligned}\phi_{\text{eff}}(\mathbf{w}; \boldsymbol{\theta}^0, \Delta^0) = & \underbrace{a^*(\mathbf{x})}_{\in\Lambda_2} + \underbrace{tb^*(y_1, \mathbf{x})}_{\in\Lambda_1} + \underbrace{(1-t)c^*(y_0, \mathbf{x}) + \frac{t - w(\mathbf{x})}{w(\mathbf{x})} \mathrm{E}_0(c^* \mid \mathbf{x})}_{\in\Lambda_0} + \underbrace{\mathbf{H}^{\mathrm{T}} \mathbf{S}_{\text{eff}}(\mathbf{w}; \boldsymbol{\theta}^0)}_{\in\Lambda^\perp}, \\ \equiv & \phi(\mathbf{w}; \boldsymbol{\theta}^0, \Delta^0) + \mathbf{H}^{\mathrm{T}} \mathbf{S}_{\text{eff}}(\mathbf{w}; \boldsymbol{\theta}^0), \end{aligned} \tag{9}$$

where $\mathrm{E}(a^*) = 0$, $\mathrm{E}_1(b^* \mid \mathbf{x}, 1) = 0$, $\mathrm{E}_0(c^* \mid \mathbf{x}, 0) = 0$, $\mathbf{H}$ is a $d \times 1$ vector and $\mathbf{S}_{\text{eff}}(\mathbf{w}; \boldsymbol{\theta})$ is the efficient score of $\boldsymbol{\theta}$, to be delineated later in Section 3.2. The explicit formulas of $\phi(\mathbf{w}; \boldsymbol{\theta}, \Delta)$ and $\mathbf{H}$ are summarized below, with the proof contained in Appendix C.

Proposition 2 *We can derive that* $\mathbf{H} = p^{-1}\mathbf{M}^{-1}(\mathbf{D} - \mathbf{Q})$ *and* $\phi(\mathbf{w}; \boldsymbol{\theta}, \Delta) =$

$$p^{-1}\left[\frac{t - \pi(\boldsymbol{\theta})}{1 - \pi(\boldsymbol{\theta})} Y - \frac{t - \pi(\boldsymbol{\theta})}{1 - \pi(\boldsymbol{\theta})} \frac{1}{B(\mathbf{x})} \{w(\mathbf{x}) \mathrm{E}_1(Y_1 \mid \mathbf{x}, 1) + C(\mathbf{x})\} - \Delta \left\{ t - \frac{t - \pi(\boldsymbol{\theta})}{1 - \pi(\boldsymbol{\theta})} \frac{w(\mathbf{x})}{B(\mathbf{x})} \right\}\right], \tag{10}$$

where the simplified notations $\{\mathbf{M}, \mathbf{Q}, \mathbf{D}, B(\mathbf{x}), C(\mathbf{x})\}$ *are defined in Table 1.*

It is clear that any estimator of Δ, $\widehat{\Delta}$, depends on a generic estimator of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}$. This is also reflected in the efficient influence function $\phi_{\text{eff}}(\mathbf{w}; \boldsymbol{\theta}, \Delta)$ in (9). For a given $\boldsymbol{\theta}$, the efficient estimator of Δ can be obtained from solving $\frac{1}{n}\sum_{i=1}^n \phi(\mathbf{w}_i; \boldsymbol{\theta}, \Delta) = 0$; i.e.,

$$\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta}) = \frac{\frac{1}{n}\sum_{i=1}^n \left[\frac{t_i - \pi_i(\boldsymbol{\theta})}{1 - \pi_i(\boldsymbol{\theta})} y_i - \frac{t_i - \pi_i(\boldsymbol{\theta})}{1 - \pi_i(\boldsymbol{\theta})} \frac{1}{\widehat{B}(\mathbf{x}_i)} \left\{ \widehat{w}(\mathbf{x}_i) \widehat{\mathrm{E}}_1(Y_1 \mid \mathbf{x}_i, 1) + \widehat{C}(\mathbf{x}_i) \right\}\right]}{\frac{1}{n}\sum_{i=1}^n \left\{ t_i - \frac{t_i - \pi_i(\boldsymbol{\theta})}{1 - \pi_i(\boldsymbol{\theta})} \frac{\widehat{w}(\mathbf{x}_i)}{\widehat{B}(\mathbf{x}_i)} \right\}}.$$

The estimator $\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta})$ literally stands for a family of estimators. If the true value of $\boldsymbol{\theta}$, $\boldsymbol{\theta}^0$, is plugged in, one would obtain $\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta}^0)$. Because $\boldsymbol{\theta}$ is unknown, a natural way is to plug in the efficient estimator of $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\theta}}_{\text{eff}}$, that solves $\frac{1}{n}\sum_{i=1}^n \mathbf{S}_{\text{eff}}(\mathbf{w}_i; \boldsymbol{\theta}) = \mathbf{0}$. Then one would obtain $\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$, which is also abbreviated as $\widehat{\Delta}_{\text{eff}}$ throughout the paper. Its implementation needs modeling of the nuisance parts $\mathrm{E}_0(\cdot \mid \mathbf{x}, 0)$ and $\mathrm{E}_1(\cdot \mid \mathbf{x}, 1)$. The property of the estimator $\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta})$ certainly depends on how well the nuisance models $\widehat{\mathrm{E}}_0(\cdot \mid \mathbf{x}, 0)$ and $\widehat{\mathrm{E}}_1(\cdot \mid \mathbf{x}, 1)$ work. It can also be easily checked that $\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta})$ satisfies the Neyman orthogonality condition (Neyman, 1959, 1979; Chernozhukov et al., 2018). In Section 4, we will show with details that under some regularity conditions, formalized later in Assumptions 1–2,

$$\begin{aligned}\sqrt{n}\left\{\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta}^0) - \Delta^0\right\} &= n^{-1/2}\sum_{i=1}^n \phi(\mathbf{w}_i; \boldsymbol{\theta}^0, \Delta^0) + o_p(1) \xrightarrow{d} N\left[0, \mathrm{E}\{\phi^2(\mathbf{w}; \boldsymbol{\theta}^0, \Delta^0)\}\right], \\ \sqrt{n}\left\{\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}}) - \Delta^0\right\} &= n^{-1/2}\sum_{i=1}^n \phi_{\text{eff}}(\mathbf{w}_i; \boldsymbol{\theta}^0, \Delta^0) + o_p(1) \xrightarrow{d} N\left[0, \mathrm{E}\{\phi_{\text{eff}}^2(\mathbf{w}; \boldsymbol{\theta}^0, \Delta^0)\}\right]. \end{aligned} \tag{11}$$

Note that the estimator $\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta}^0)$ is not of practical interest since it is infeasible, whereas the estimator $\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$ achieves the semiparametric efficiency bound $\mathrm{E}\left\{\phi^2_{\text{eff}}(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0)\right\}$. Because

$$\mathrm{E}\left\{\phi^2_{\text{eff}}(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0)\right\} = \mathrm{E}\left\{\phi^2(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0)\right\} + p^{-2}(\mathbf{D}-\mathbf{Q})^{\mathrm{T}}\mathbf{M}^{-1}(\mathbf{D}-\mathbf{Q}),$$

the component $p^{-2}(\mathbf{D}-\mathbf{Q})^{\mathrm{T}}\mathbf{M}^{-1}(\mathbf{D}-\mathbf{Q})$, or equivalently the component $\mathbf{H}^{\mathrm{T}}\mathbf{S}_{\text{eff}}(\mathbf{w};\boldsymbol{\theta})$ in (9), amounts to the estimation variability brought by $\widehat{\boldsymbol{\theta}}_{\text{eff}}$.

3.2 Efficient Estimation of $\boldsymbol{\theta}$

In this section, we propose an efficient estimator of $\boldsymbol{\theta}$ by characterizing its efficient influence function. Proposition 1 establishes that the nuisance tangent space admits the decomposition $\Lambda = \Lambda_0 \oplus \Lambda_1 \oplus \Lambda_2$. It follows that any element in its perpendicular $\Lambda^{\perp}$ with an arbitrary d-dimensional nondegenerate function $\mathbf{g}(\mathbf{x})$ yields a consistent estimator of $\boldsymbol{\theta}$; i.e., the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\frac{t_i-\pi_i(\boldsymbol{\theta})}{1-\pi_i(\boldsymbol{\theta})}\mathbf{g}(\mathbf{x}_i) = \mathbf{0}. \tag{12}$$

The optimal choice of $\mathbf{g}(\mathbf{x})$ that provides the minimum estimation variability will be given by the efficient score function $\mathbf{S}_{\text{eff}}$, which is defined as the projection of the score function

$$\begin{aligned}\mathbf{S}_{\boldsymbol{\theta}} &= \frac{\{t-w(\mathbf{x})\}\{1-w(\mathbf{x})\}}{w(\mathbf{x})}\mathrm{E}_0[\{1-\pi(\boldsymbol{\theta})\}^{-2}\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}) \mid \mathbf{x}, 0] \\ &= \frac{t-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0\left[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0) \mid \mathbf{x}\right]\end{aligned}$$

derived in Appendix C onto the space $\Lambda^{\perp}$ (Tsiatis, 2006, Theorem 3.5). It can be easily seen that $\mathbf{S}_{\boldsymbol{\theta}} \perp \Lambda_2$ and $\mathbf{S}_{\boldsymbol{\theta}} \perp \Lambda_1$. Therefore, we can decompose $\mathbf{S}_{\boldsymbol{\theta}}$ as

$$\begin{aligned}&\frac{t-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0\left[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0) \mid \mathbf{x}\right] \\ &= \underbrace{(1-t)\mathbf{c}_1(y_0,\mathbf{x}) + \frac{t-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0\{\mathbf{c}_1(y_0,\mathbf{x}) \mid \mathbf{x}\}}_{\in\Lambda_0} + \underbrace{\frac{t-\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\mathbf{g}_1(\mathbf{x})}_{\in\Lambda^{\perp}},\end{aligned}$$

where $\mathrm{E}_0(\mathbf{c}_1 \mid \mathbf{x}, 0) = \mathbf{0}$. Taking $t=1$ and $t=0$ respectively of the above expression, it is straightforward to obtain

$$\begin{aligned}\mathbf{c}_1(y_0,\mathbf{x}) &= \left\{\frac{\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)} - \frac{w(\mathbf{x})}{1-w(\mathbf{x})}\right\}\mathbf{g}_1(\mathbf{x}), \\ \text{and}\quad \mathbf{g}_1(\mathbf{x}) &= \frac{\mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0) \mid \mathbf{x}]}{\mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\pi(\boldsymbol{\theta}^0) \mid \mathbf{x}]} = \frac{\mathbf{A}(\mathbf{x})}{B(\mathbf{x})}.\end{aligned}$$

Hence, the efficient score function is

$$\mathbf{S}_{\text{eff}}(\mathbf{w};\boldsymbol{\theta}) = \frac{t-\pi(\boldsymbol{\theta})}{1-\pi(\boldsymbol{\theta})}\frac{\mathbf{A}(\mathbf{x})}{B(\mathbf{x})}. \tag{13}$$

This implies that, once the nuisance functions have been estimated, the efficient estimator $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ can be obtained by solving $\frac{1}{n}\sum_{i=1}^{n}\mathbf{S}_{\text{eff}}(\mathbf{w}_i;\boldsymbol{\theta})=\mathbf{0}$, which is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{t_i-\pi_i(\boldsymbol{\theta})}{1-\pi_i(\boldsymbol{\theta})}\frac{\widehat{\mathbf{A}}(\mathbf{x}_i)}{\widehat{B}(\mathbf{x}_i)}=\mathbf{0}.$$

Under suitable regularity conditions, we can show that the estimator $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ achieves the semiparametric efficiency bound. That is,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{\text{eff}}-\boldsymbol{\theta}^0)=n^{-1/2}\sum_{i=1}^{n}\mathbf{M}^{-1}\mathbf{S}_{\text{eff}}(\mathbf{w}_i;\boldsymbol{\theta}^0)+o_p(1)\xrightarrow{d}N\left(\mathbf{0},\mathbf{M}^{-1}\right), \tag{14}$$

where $\mathbf{M}=\text{var}(\mathbf{S}_{\text{eff}})=\text{E}(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\text{T}})=-\text{E}\left(\partial\mathbf{S}_{\text{eff}}/\partial\boldsymbol{\theta}^{\text{T}}\right)=\text{E}[B(\mathbf{x})^{-1}\mathbf{A}(\mathbf{x})\mathbf{A}(\mathbf{x})^{\text{T}}]$ and is assumed to be invertible. It can be easily checked that the efficient score $\mathbf{S}_{\text{eff}}(\mathbf{w};\boldsymbol{\theta})$ also satisfies the Neyman orthogonality condition. The proof of this result is given in Lemma A1 of Appendix D.

Last but not least, when $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ is obtained for the model $\pi(y_0,\mathbf{u};\boldsymbol{\theta})=\text{expit}(\theta_1+\theta_2 y_0+\boldsymbol{\theta}_3^{\text{T}}\mathbf{u})$, one can easily develop a hypothesis testing procedure for

$$H_0:\theta_2=0 \text{ versus } H_1:\theta_2\neq 0,$$

based on the corresponding Wald-type statistics, and evaluate whether the treatment assignment indeed depends on the potential untreated outcome.

3.3 A Simpler Estimator for Δ

Though less efficient, a simpler estimator of Δ exists. According to the definition of Δ, i.e., $\Delta=p^{-1}\{\text{E}(TY_1)-\text{E}(TY_0)\}=p^{-1}\left[\text{E}(TY_1)-\text{E}\left\{(1-T)\frac{\pi(\boldsymbol{\theta})}{1-\pi(\boldsymbol{\theta})}Y_0\right\}\right]$, an immediate estimator of Δ is

$$\widehat{\Delta}_{\text{alt}}(\boldsymbol{\theta})=\frac{\frac{1}{n}\sum_{i=1}^{n}\left\{t_i y_{1i}-(1-t_i)\frac{\pi_i(\boldsymbol{\theta})}{1-\pi_i(\boldsymbol{\theta})}y_{0i}\right\}}{\frac{1}{n}\sum_{i=1}^{n}t_i},$$

whose corresponding influence function is

$$\phi_{\text{alt}}(\mathbf{w};\boldsymbol{\theta},\Delta)=p^{-1}\left\{ty_1-(1-t)\frac{\pi(\boldsymbol{\theta})}{1-\pi(\boldsymbol{\theta})}y_0-t\Delta\right\}. \tag{15}$$

It can be easily verified that,

$$\begin{aligned}\sqrt{n}\{\widehat{\Delta}_{\text{alt}}(\boldsymbol{\theta}^0)-\Delta^0\}&=n^{-1/2}\sum_{i=1}^{n}\phi_{\text{alt}}(\mathbf{w}_i;\boldsymbol{\theta}^0,\Delta^0)+o_p(1)\xrightarrow{d}N\left[0,\text{E}\{\phi_{\text{alt}}^2(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0)\}\right],\\ \sqrt{n}\{\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})-\Delta^0\}&=n^{-1/2}\sum_{i=1}^{n}\left\{\phi_{\text{alt}}(\mathbf{w}_i;\boldsymbol{\theta}^0,\Delta^0)-p^{-1}\mathbf{Q}^{\text{T}}\mathbf{M}^{-1}\mathbf{S}_{\text{eff}}(\mathbf{w}_i;\boldsymbol{\theta}^0)\right\}+o_p(1),\end{aligned} \tag{16}$$

under standard regularity conditions (e.g. those in Theorems 2.6 and 3.4 of Newey and McFadden (1994)).

Unfortunately, neither the estimator $\widehat{\Delta}_{\text{alt}}(\boldsymbol{\theta}^0)$ nor the estimator $\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$ is efficient. This can be seen from the fact that $\phi_{\text{alt}}(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0) \notin \Lambda$. Indeed, we have the following projection result

$$\phi_{\text{alt}}(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0) = \underbrace{\phi(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0)}_{\Pi\{\phi_{\text{alt}}(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0)|\Lambda\}} + \underbrace{\frac{t-\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)} p^{-1} B(\mathbf{x})^{-1} V(\mathbf{x})}_{\Pi\{\phi_{\text{alt}}(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0)|\Lambda^\perp\}}. \tag{17}$$

Immediately, we have

$$\mathrm{E}\{\phi^2_{\text{alt}}(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0)\} = \mathrm{E}\{\phi^2(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0)\} + \mathrm{E}\left[\Pi\{\phi_{\text{alt}}(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0) \mid \Lambda^\perp\}^2\right] > \mathrm{E}\{\phi^2(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0)\}.$$

Thus, under regularity conditions that justify the first-order linear representations in (11) and (16), we have the following comparison.

Lemma 2 *Under Assumptions 1–2 presented later, when $\boldsymbol{\theta}^0$ is known, the estimator $\widehat{\Delta}_{\text{alt}}(\boldsymbol{\theta}^0)$ is less efficient than $\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta}^0)$.*

To make the comparison result complete, we also summarize the following result when $\boldsymbol{\theta}$ is estimated by the efficient estimator $\widehat{\boldsymbol{\theta}}_{\text{eff}}$. Its efficiency comparison is guaranteed since $\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$ is semiparametrically efficient among all RAL estimators.

Lemma 3 *Under Assumptions 1–2 presented later, the estimator $\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$ is less efficient than $\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$.*

4. Implementation and Asymptotic Properties

In this section, we provide implementation details for using machine learning tools to construct the efficient estimators $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ and $\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$ and study their asymptotic properties. As mentioned in Section 3, jointly obtaining $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ and $\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$ is equivalent to solving

$$\begin{cases} n^{-1}\sum_{i=1}^n \mathbf{S}_{\text{eff}}(\mathbf{w}_i;\boldsymbol{\theta}) = \mathbf{0} \\ n^{-1}\sum_{i=1}^n \phi(\mathbf{w}_i;\boldsymbol{\theta},\Delta) = 0, \end{cases}$$

which involves the nuisance components

$$\boldsymbol{\eta} = \left\{ \begin{array}{c} \mathrm{E}_1(Y_1 \mid \mathbf{x},1),\ \mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1} \mid \mathbf{x},0],\ \mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-2}\pi(\boldsymbol{\theta}^0) \mid \mathbf{x},0], \\ \mathrm{E}_0[Y_0\{1-\pi(\boldsymbol{\theta}^0)\}^{-2}\pi(\boldsymbol{\theta}^0)^2 \mid \mathbf{x},0],\ \mathrm{E}_0[\{1-\pi(\boldsymbol{\theta}^0)\}^{-2}\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0) \mid \mathbf{x},0] \end{array} \right\} \in \mathcal{T},$$

where $\mathcal{T}$ denotes the nuisance function space. As a special case, when the potential outcomes (Y_0, Y_1) are binary, the evaluation of $\boldsymbol{\eta}$ reduces to evaluating $\boldsymbol{\eta} = \{f_0(1 \mid \mathbf{x},0), f_1(1 \mid \mathbf{x},1)\} = \{\mathrm{E}_0(Y_0 \mid \mathbf{x},0),\ \mathrm{E}_1(Y_1 \mid \mathbf{x},1)\}$. To emphasize the role of the nuisance components in the technical development, we put both parameters together as $\boldsymbol{\psi} = \{\boldsymbol{\theta}^{\mathrm{T}}, \Delta\}^{\mathrm{T}}$, and denote

$$\mathbf{G}(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta}) = \{\mathbf{S}_{\text{eff}}(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta})^{\mathrm{T}}, \phi(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta})\}^{\mathrm{T}}.$$

In Appendix D, we show in Lemma A1 that the score function $\mathbf{G}$ satisfies the Neyman orthogonality condition (Chernozhukov et al., 2018), in the sense that the pathwise (Gateaux) derivative

$$\partial_r \mathrm{E}\left[\mathbf{G}\{\mathbf{W}; \boldsymbol{\psi}^0, \boldsymbol{\eta}^0 + r(\boldsymbol{\eta} - \boldsymbol{\eta}^0)\}\right] |_{r=0} = \mathbf{0}, \quad \text{for any } \boldsymbol{\eta} \in \mathcal{T}.$$

This ensures that small perturbations in the estimation of nuisance functions have only second-order effects on the target parameter. In addition, we apply the sample splitting technique (Bickel, 1982; Schick, 1986; Chernozhukov et al., 2018), widely used to eliminate the first-order overfitting bias induced by the nuisance estimation. Let K be a fixed positive integer that does not change with the sample size n. We take a K-fold random partition $(I_k)_{k=1}^K$ of the observation indices, $[n] = \{1, 2, \ldots, n\}$, such that the size of each fold, I_k, is $|I_k| = n^* = n/K$. For each $k \in [K] = \{1, 2, \ldots, K\}$, let us define $I_k^c = [n] \backslash I_k$ and construct a machine learning (ML) estimator $\widehat{\boldsymbol{\eta}}^{[-k]} = \widehat{\boldsymbol{\eta}}\{(\mathbf{W}_i)_{i \in I_k^c}\} \in \mathcal{T}_n$, where $\mathcal{T}_n \subset \mathcal{T}$ is a nuisance realization set that is a suitably shrinking neighborhood of $\boldsymbol{\eta}^0$. Then we obtain $\widehat{\boldsymbol{\psi}}$ by solving the cross-fitting estimating equation

$$\frac{1}{n} \sum_{k=1}^{K} \sum_{i \in I_k} \mathbf{G}(\mathbf{W}_i; \boldsymbol{\psi}, \widehat{\boldsymbol{\eta}}^{[-k]}) = \mathbf{0}. \tag{18}$$

Remark 3 Note that all nuisance functions in $\boldsymbol{\eta}$ take the form $\mathrm{E}_1(\cdot \mid \mathbf{x}, 1)$ and $\mathrm{E}_0(\cdot \mid \mathbf{x}, 0)$. Therefore, to obtain $\widehat{\boldsymbol{\eta}}$, we can rely on off-the-shelf machine learning methods such as XGBoost or random forests to fit regression-style learners for these conditional means. Because some nuisance functions depend on the true parameter value $\boldsymbol{\theta}^0$, we can simply replace it with a consistent, not necessarily efficient, estimator $\widehat{\boldsymbol{\theta}}$ solving (12) when constructing the regression target. In the special case where (Y_0, Y_1) are binary, this training step simplifies further: it suffices to learn $\mathrm{E}_0(Y_0 \mid \mathbf{x}, 0)$ and $\mathrm{E}_1(Y_1 \mid \mathbf{x}, 1)$.

Let $o(\alpha_n)$, $O(\alpha_n)$, $\omega(\alpha_n)$, $\Omega(\alpha_n)$ and $\Theta(\alpha_n)$ represent sequences that grow at a smaller, equal or smaller, larger, equal or larger, and equal rate relative to α_n, respectively. Further, for any square-integrable vector-valued function $\mathbf{f}(\cdot) = (f_1(\cdot), \ldots, f_p(\cdot))^{\mathrm{T}}$, random variable $\mathbf{W}$, and probability measure P, define $\|\mathbf{f}(\cdot)\|_{P,2} = \|\mathbf{f}(\mathbf{W})\|_{P,2} = \{\int \sum_{j=1}^p |f_j(\mathbf{w})|^2 \mathrm{d}P(\mathbf{w})\}^{1/2}$. We use $\partial_{\mathbf{x}} f$ to denote $(\partial/\partial \mathbf{x}) f$ and $\partial_{\mathbf{x}} f(\mathbf{x}_0)$ for $(\partial/\partial \mathbf{x}) f(\mathbf{x}) \mid_{\mathbf{x}=\mathbf{x}_0}$.

The asymptotic properties for the estimators $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ and $\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$ mainly follow Theorem 3.3 of Chernozhukov et al. (2018), where they presented the general results of the debiased ML estimators for the nonlinear Neyman orthogonal score. Nevertheless, we first provide the necessary assumptions in our own setting.

Assumption 1 *There exists $\delta_n = \Omega(n^{-1/2}\log n)$ such that parameter $\boldsymbol{\theta}$ belongs to a compact set $\mathcal{B}_1 \subset \mathbb{R}^d$, which contains a ball of radius $c_1\delta_n$ centered at the true parameter $\boldsymbol{\theta}^0$, where c_1 is some finite positive constant. Also, parameter Δ belongs to a compact set $\mathcal{B}_2 \subset \mathbb{R}$ and $(\Delta^0 - c_1\delta_n, \Delta^0 + c_1\delta_n) \subseteq \mathcal{B}_2$. In other words, $\mathcal{B} = \mathcal{B}_1 \times \mathcal{B}_2 \subset \mathbb{R}^{d+1}$ contains a ball of radius $\Omega(n^{-1/2}\log n)$ centered at $\boldsymbol{\psi}^0$. The treatment assignment satisfies $c < \pi(\boldsymbol{\theta}) < 1 - c$, where $0 < c \leq 1/2$. Further, $\pi(\boldsymbol{\theta})$ is thrice differentiable with respect to $\boldsymbol{\theta}$ and the derivatives are continuous and bounded. In addition, $Y_t = \Theta(1)$ and $\mathrm{E}_t(Y_t^2 \mid \mathbf{x}, t) = \Theta(1)$ for $t = 0, 1$.*

Assumption 2 *The nuisance estimates achieve an $o_p(n^{-1/4})$ convergence rate, i.e., $\|\widehat{\boldsymbol{\eta}}^{[-k]} - \boldsymbol{\eta}\|_{P,2} = o_p(n^{-1/4})$, for $k \in \{1, 2, \ldots, K\}$.*

Assumption 1 is on regularity conditions. Assumption 2 is about the quality of ML nuisance estimates, which ensures that the second-order residuals of the DML estimator are negligible. This rate is achievable for many ML methods under suitable structural assumptions on the nuisance functions; see, for example, Bickel et al. (2009) for ℓ_1-penalized methods in sparse models, Wager and Athey (2018) for regression trees and random forests, and Farrell et al. (2021) for neural networks. For further discussion of these machine learning methods and the conditions under which they satisfy the nuisance estimation requirements, see Chernozhukov et al. (2018) and the references therein. Now we establish the large sample property of the estimator $\widehat{\boldsymbol{\psi}}$ in the following theorem.

Theorem 1 *Under the Assumptions 1–2, we have*

$$\sqrt{n}(\widehat{\boldsymbol{\psi}} - \boldsymbol{\psi}^0) = -n^{-1/2} \sum_{i=1}^{n} \mathbf{J}^{-1} \mathbf{G}(\mathbf{W}_i; \boldsymbol{\psi}^0, \boldsymbol{\eta}^0) + o_p(1) \xrightarrow{d} N(\mathbf{0}, \mathbf{V}), \text{ where}$$

$$\mathbf{J} = \partial_{\boldsymbol{\psi}^{\mathrm{T}}} \mathrm{E}\{\mathbf{G}(\mathbf{W}; \boldsymbol{\psi}, \boldsymbol{\eta}^0)\} \mid_{\boldsymbol{\psi}=\boldsymbol{\psi}^0} = \begin{pmatrix} -\mathbf{M} & \mathbf{0} \\ p^{-1}(\mathbf{D}-\mathbf{Q})^{\mathrm{T}} & -1 \end{pmatrix}, \text{ and}$$

$$\mathbf{V} = \begin{pmatrix} \mathbf{M}^{-1} & p^{-1}\mathbf{M}^{-1}(\mathbf{D}-\mathbf{Q}) \\ p^{-1}(\mathbf{D}-\mathbf{Q})^{\mathrm{T}}\mathbf{M}^{-1} & p^{-2}(\mathbf{D}-\mathbf{Q})^{\mathrm{T}}\mathbf{M}^{-1}(\mathbf{D}-\mathbf{Q}) + \mathrm{E}\{\phi^2(\mathbf{w}; \boldsymbol{\theta}^0, \Delta^0)\} \end{pmatrix},$$

and the estimator $\widehat{\boldsymbol{\psi}}$ achieves the semiparametric efficiency bound.

Note that here $\widehat{\boldsymbol{\psi}} = (\widehat{\boldsymbol{\theta}}_{\text{eff}}^{\mathrm{T}}, \widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}}))^{\mathrm{T}}$; hence, the results in (14) and (11) follow immediately. The proof of Theorem 1 is provided in Appendix D. In addition, the asymptotic variance $\mathbf{V}$ does not involve any unknown functions beyond $\boldsymbol{\eta}$ and therefore can be estimated using a simple plug-in estimator; see Theorem 3.2 of Chernozhukov et al. (2018) for a general variance estimation formula and a proof of consistency.

5. Simulation Studies

In this section, we present some numerical results. We implement the estimators of $\boldsymbol{\theta}$ and Δ introduced in Section 3. We also compare the finite-sample performance of estimators of Δ with two naive estimators that assume strong ignorability (Hahn, 1998). The first naive estimator is

$$\widehat{\Delta}_{\text{nv1}} = \frac{n^{-1} \sum_{i=1}^{n} \left\{ t_i y_{1i} - (1-t_i) \frac{\widehat{w}(\mathbf{x}_i)}{1-\widehat{w}(\mathbf{x}_i)} y_{0i} \right\}}{n^{-1} \sum_{i=1}^{n} t_i},$$

whose corresponding influence function is $p^{-1} \left\{ TY_1 - (1-T) \frac{w(\mathbf{x})}{1-w(\mathbf{x})} Y_0 - T\Delta \right\}$. The second naive estimator is

$$\widehat{\Delta}_{\text{nv2}} = \frac{n^{-1} \sum_{i=1}^{n} \left\{ t_i y_{1i} - (1-t_i) \frac{\widehat{w}(\mathbf{x}_i)}{1-\widehat{w}(\mathbf{x}_i)} y_{0i} - \frac{t_i - \widehat{w}(\mathbf{x}_i)}{1-\widehat{w}(\mathbf{x}_i)} \widehat{\mathrm{E}}_0(Y_{0i} \mid \mathbf{x}_i) \right\}}{n^{-1} \sum_{i=1}^{n} t_i},$$

which is based on the EIF $p^{-1}\Big[T\{Y_1 - \mathrm{E}_0(Y_0 \mid \mathbf{X}) - \Delta\} - (1-T)\frac{w(\mathbf{x})}{1-w(\mathbf{x})}\{Y_0 - \mathrm{E}_0(Y_0 \mid \mathbf{X})\}\Big]$. To implement the estimator $\widehat{\Delta}_{\mathrm{nv1}}$, one needs to correctly specify the propensity score model $w(\mathbf{x})$. The estimator $\widehat{\Delta}_{\mathrm{nv2}}$ enjoys the so-called double robustness property in that one only needs to correctly specify either $w(\mathbf{x})$ or $\mathrm{E}_0(Y_0 \mid \mathbf{x})$, both of which are essentially conditional expectations.

In the simulations, we consider a random sample with 600 observations with the following data generating process. We first generate $\mathbf{X} \sim N_2(\mathbf{0}, \mathbf{I}_2)$. For the potential outcomes, Y_1 follows a Bernoulli distribution with success probability $\mathrm{expit}(x_1)$, whereas Y_0 follows a Bernoulli distribution with probability of success $\mathrm{expit}(x_2)$. We generate a binary treatment T according to

$$\mathrm{pr}(T = 1 \mid y_0, \mathbf{x}) = \mathrm{pr}(T = 1 \mid y_0, x_1) = \mathrm{expit}(\theta_1 + \theta_2 y_0 + \theta_3 x_1), \tag{19}$$

where $\boldsymbol{\theta} = (0.3, -0.3, -0.25)^{\mathrm{T}}$. It can be computed that the true value of ATT is around 0.012.

We solve (13) to obtain the estimator of $\boldsymbol{\theta}$, and solve (15) and (10) to obtain estimates for Δ respectively. We also fit random forest (rforest, via R package `ranger`) and XGBoost (xgboost, via R package `xgboost`) models using all simulated data for the nuisance models $\mathrm{E}_0(Y_0 \mid \mathbf{X})$ and $\mathrm{E}_1(Y_1 \mid \mathbf{X})$. The results based on 500 simulation replications are reported in Table 2. In this table, $\widehat{\mathrm{sd}}_p$ is the estimated standard deviation (SD) based on 200 perturbation samples (Jin et al., 2001). The conclusions are quite clear. For $\boldsymbol{\theta}$, we can see that its estimate yields very small bias, and the estimated SD $\widehat{\mathrm{sd}}_p$ is close to the empirical SD. For Δ, the naive estimators are biased and should not be used in practice; in contrast, the proposed estimators all yield almost zero bias. Regarding different machine learning models for estimating the nuisances, we note that the performances of the corresponding estimators of $\boldsymbol{\theta}$ and Δ, albeit varying slightly, are overall very similar and reach the same conclusion.

To further investigate the efficiency gain of the estimator $\widehat{\Delta}_{\mathrm{eff}}(\boldsymbol{\theta})$ over $\widehat{\Delta}_{\mathrm{alt}}(\boldsymbol{\theta})$ for a particular $\boldsymbol{\theta}$, we use the following four choices of $\boldsymbol{\theta}$: the efficient estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{eff}}$, the true value $\boldsymbol{\theta}^0$, $\widehat{\boldsymbol{\theta}}_{\mathrm{arb},1}$ corresponding to the solution of (12) with $\mathbf{g}(\mathbf{x}) = \{w(\mathbf{x}), w(\mathbf{x})x_1, w(\mathbf{x})x_1^2\}^{\mathrm{T}}$, and $\widehat{\boldsymbol{\theta}}_{\mathrm{arb},2}$ corresponding to the solution of (12) with $\mathbf{g}(\mathbf{x}) = \{w(\mathbf{x}), \mathrm{expit}(\theta_1 + \theta_2 + \theta_3 x_1)\mathrm{E}_0(Y_0 \mid \mathbf{x}), w(\mathbf{x})x_1\}^{\mathrm{T}}$. The results are summarized in Table 3. When using the efficient estimate $\widehat{\boldsymbol{\theta}}_{\mathrm{eff}}$, the estimator $\widehat{\Delta}_{\mathrm{eff}}(\widehat{\boldsymbol{\theta}}_{\mathrm{eff}})$ is still advantageous, but the efficiency gain is quite minimal. On the other hand, when using the true value $\boldsymbol{\theta}^0$, the estimator $\widehat{\Delta}_{\mathrm{eff}}(\boldsymbol{\theta}^0)$ is a lot more efficient compared to $\widehat{\Delta}_{\mathrm{alt}}(\boldsymbol{\theta}^0)$. This indicates, because of the projection result (17), the estimator based on the influence function $\phi(\mathbf{w}; \boldsymbol{\theta}, \Delta)$ does improve the estimation efficiency compared to the one using $\phi_{\mathrm{alt}}(\mathbf{w}; \boldsymbol{\theta}, \Delta)$, but this gain can be largely mitigated by efficiently estimating $\widehat{\boldsymbol{\theta}}_{\mathrm{eff}}$. In general, when we arbitrarily choose $\mathbf{g}(\mathbf{x})$ functions in (12) which lead to estimators $\widehat{\boldsymbol{\theta}}_{\mathrm{arb},1}$ and $\widehat{\boldsymbol{\theta}}_{\mathrm{arb},2}$ in Table 3, the efficiency gain could be somewhere between 15% and 70%.

6. Application to NCDB

In this section, we analyze the National Cancer Database (NCDB), a clinical oncology database sourced from hospital registry data that are collected in more than 1,500 Commission on Cancer (CoC)-accredited facilities.

Table 2: Simulation: results for $\boldsymbol{\theta}$ and ATT. The notation sd, $\widehat{\text{sd}}_p$ and CP refer to the empirical standard deviation, the standard deviation based on 200 perturbation samples, and 95% coverage probability, respectively. The notation $\widehat{\theta}_{\text{eff},1}$, $\widehat{\theta}_{\text{eff},2}$ and $\widehat{\theta}_{\text{eff},3}$ refer to the three coordinates of the parameter estimate $\widehat{\boldsymbol{\theta}}_{\text{eff}}$, respectively.

		$w(\mathbf{X})$	$\text{E}_0(Y_0 \mid \mathbf{X})$	$\text{E}_1(Y_1 \mid \mathbf{X})$	Bias	sd	$\widehat{\text{sd}}_p$	CP
Naive	$\widehat{\Delta}_{\text{nv1}}$	logistic	-	-	-0.063	0.044	0.041	0.636
		rforest	-	-	0.228	0.035	0.038	0.000
		xgboost	-	-	-0.004	0.047	0.043	0.946
	$\widehat{\Delta}_{\text{nv2}}$	logistic	logistic	-	-0.063	0.044	0.041	0.646
		rforest	rforest	-	-0.038	0.039	0.038	0.818
		xgboost	xgboost	-	-0.055	0.043	0.040	0.698
		$\pi(Y_0, \mathbf{U}; \boldsymbol{\theta})$	$\text{E}_0(Y_0 \mid \mathbf{X})$	$\text{E}_1(Y_1 \mid \mathbf{X})$	Bias	sd	$\widehat{\text{sd}}_p$	CP
Proposed	$\widehat{\theta}_{\text{eff},1}$	model (19)	logistic	-	-0.017	0.226	0.253	0.968
	$\widehat{\theta}_{\text{eff},2}$		logistic	-	0.001	0.452	0.442	0.986
	$\widehat{\theta}_{\text{eff},3}$		logistic	-	-0.008	0.088	0.092	0.964
	$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$		logistic	-	-0.003	0.106	0.104	0.936
	$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$		logistic	logistic	-0.003	0.106	0.111	0.946
	$\widehat{\theta}_{\text{eff},1}$	model (19)	rforest	-	0.002	0.136	0.144	0.960
	$\widehat{\theta}_{\text{eff},2}$		rforest	-	0.001	0.222	0.239	0.964
	$\widehat{\theta}_{\text{eff},3}$		rforest	-	-0.008	0.088	0.089	0.958
	$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$		rforest	-	-0.001	0.052	0.052	0.944
	$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$		rforest	rforest	-0.001	0.051	0.052	0.956
	$\widehat{\theta}_{\text{eff},1}$	model (19)	xgboost	-	-0.008	0.179	0.172	0.944
	$\widehat{\theta}_{\text{eff},2}$		xgboost	-	0.007	0.328	0.312	0.946
	$\widehat{\theta}_{\text{eff},3}$		xgboost	-	-0.007	0.088	0.089	0.956
	$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$		xgboost	-	-0.005	0.078	0.071	0.924
	$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$		xgboost	xgboost	-0.004	0.078	0.082	0.932

Table 3: Simulation: results for the relative efficiency (RE) of $\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta})$ over $\widehat{\Delta}_{\text{alt}}(\boldsymbol{\theta})$ with different estimates of $\boldsymbol{\theta}$. Here RE is defined as the mean squared error (MSE) of the estimator $\widehat{\Delta}_{\text{alt}}(\boldsymbol{\theta})$ divided by that of $\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta})$. RE>1 indicates that $\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta})$ is more efficient. The estimator $\widehat{\boldsymbol{\theta}}_{\text{arb},1}$ corresponds to the choice of $\mathbf{g}(\mathbf{x}) = \{w(\mathbf{x}), w(\mathbf{x})x_1, w(\mathbf{x})x_1^2\}^{\mathrm{T}}$ and the estimator $\widehat{\boldsymbol{\theta}}_{\text{arb},2}$ corresponds to the choice of $\mathbf{g}(\mathbf{x}) = \{w(\mathbf{x}), \text{expit}(\theta_1 + \theta_2 + \theta_3 x_1)\mathrm{E}_0(Y_0 \mid \mathbf{x}), w(\mathbf{x})x_1\}^{\mathrm{T}}$, in (12).

	$\pi(Y_0, \mathbf{U}; \boldsymbol{\theta})$	$\mathrm{E}_0(Y_0 \mid \mathbf{X})$	$\mathrm{E}_1(Y_1 \mid \mathbf{X})$	Bias	sd	MSE	RE
$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$	model (19)	logistic	-	-0.003	0.106	0.011	
$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$		logistic	logistic	-0.003	0.106	0.011	
							1.004
$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$	model (19)	rforest	-	-0.001	0.052	0.003	
$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$		rforest	rforest	-0.001	0.051	0.003	
							1.059
$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$	model (19)	xgboost	-	-0.005	0.078	0.006	
$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$		xgboost	xgboost	-0.004	0.078	0.006	
							1.014
$\widehat{\Delta}_{\text{alt}}(\boldsymbol{\theta}^0)$	model (19)	logistic	-	-0.000	0.060	0.004	
$\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta}^0)$		logistic	logistic	-0.000	0.043	0.002	
							1.908
$\widehat{\Delta}_{\text{alt}}(\boldsymbol{\theta}^0)$	model (19)	rforest	-	-0.000	0.060	0.004	
$\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta}^0)$		rforest	rforest	-0.000	0.037	0.001	
							2.658
$\widehat{\Delta}_{\text{alt}}(\boldsymbol{\theta}^0)$	model (19)	xgboost	-	-0.000	0.060	0.004	
$\widehat{\Delta}_{\text{eff}}(\boldsymbol{\theta}^0)$		xgboost	xgboost	-0.000	0.042	0.002	
							2.040
$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{arb},1})$	model (19)	logistic	-	-0.019	0.152	0.023	
$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{arb},1})$		logistic	logistic	-0.017	0.141	0.020	
							1.171
$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{arb},1})$	model (19)	rforest	-	-0.014	0.145	0.021	
$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{arb},1})$		rforest	rforest	-0.011	0.123	0.015	
							1.407
$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{arb},1})$	model (19)	xgboost	-	-0.019	0.147	0.022	
$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{arb},1})$		xgboost	xgboost	-0.016	0.131	0.017	
							1.267
$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{arb},2})$	model (19)	logistic	-	-0.002	0.047	0.002	
$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{arb},2})$		logistic	logistic	-0.002	0.044	0.002	
							1.146
$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{arb},2})$	model (19)	rforest	-	-0.002	0.048	0.002	
$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{arb},2})$		rforest	rforest	-0.001	0.037	0.001	
							1.672
$\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{arb},2})$	model (19)	xgboost	-	-0.002	0.047	0.002	
$\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{arb},2})$		xgboost	xgboost	-0.002	0.043	0.002	
							1.238

In our application, we focus on 7,233 stage T2 lung cancer patients out of whom 1,109 were treated with stereotactic body radiation therapy (SBRT, $T = 1$) and 6,124 were treated with lobectomy surgery ($T = 0$). The estimand of interest is the ATT on 3-year survival rates; that is, the binary outcome Y equals 1 when a patient was recorded as alive at least 3 years after being diagnosed with lung cancer and 0 otherwise. There are five $\mathbf{X}$ variables in this database. Out of them, three are continuous—age, size of tumor (logarithmic scale), year of diagnosis, and the other two are categorical—income level (binary, income<\$48,000 versus income≥\$48,000), and insurance level (binary, private versus non-private). An important prognostic variable, Charlson-Deyo comorbidity index score (CDCI), is also recorded in the dataset. We conduct our analyses stratified on CDCI=0, and CDCI>0. We standardize all the continuous variables throughout the analysis.

In our analysis, we consider two naive estimators for Δ: $\widehat{\Delta}_{\text{nv1}}$ and $\widehat{\Delta}_{\text{nv2}}$ which involve both $w(\mathbf{x})$ and $\mathrm{E}_0(Y_0 \mid \mathbf{x})$. We consider three methods to fit each of them: logistic regression (logistic), random forest (rforest), and XGBoost (xgboost). The logistic regression model for $w(\mathbf{x})$ is comparable to the modeling of the propensity score in our proposed method. Also note that the efficient influence function under strong ignorability was first derived in Hahn (1998), and can be regarded as a special case of the efficient influence function derived in this paper.

The strong ignorability assumption assumes that the choice of treatment does not depend on potential outcomes Y_1 or Y_0 but on all other covariates. This assumption might not be realistic for applications. In our application, when the physician assigned the patients to either radiation therapy or surgery, their decision should not depend on the patients' income or insurance. On the other hand, the decision might potentially depend on Y_1 and Y_0. As discussed in the introduction section, $\mathbf{Z}$ consists of income and insurance levels. Therefore, to apply our method, we assume

$$\mathrm{pr}(T = 1 \mid Y_0, \mathbf{X}) = \pi(y_0, \mathbf{u}; \boldsymbol{\theta}) = \mathrm{expit}(\theta_1 + \theta_2 y_0 + \xi^{\mathrm{T}}\mathbf{u}), \tag{20}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \xi^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{u}$ includes year of diagnosis, age, and size of tumor. We first implement the proposed method $\widehat{\Delta}_{\text{alt}}$, relying on $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ that is obtained from solving (13) which only needs a model for $\mathrm{E}_0(\cdot \mid \mathbf{x}, 0)$. We then implement the proposed method $\widehat{\Delta}_{\text{eff}}$ according to the efficient influence function of Δ. The implementation of $\widehat{\Delta}_{\text{eff}}$ relies on $\widehat{\boldsymbol{\theta}}_{\text{eff}}$, $\mathrm{E}_0(\cdot \mid \mathbf{x}, 0)$ and $\mathrm{E}_1(\cdot \mid \mathbf{x}, 1)$. For comparison, we also fit logistic regression (logistic), random forest (rforest), and XGBoost (xgboost) methods to estimate $\mathrm{E}_0(\cdot \mid \mathbf{x}, 0)$ or $\mathrm{E}_1(\cdot \mid \mathbf{x}, 1)$.

The results of $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ for estimating $\boldsymbol{\theta}$ in model (20) are contained in Table 4. Similar to our simulation studies, different machine learning methods all reach the same conclusion. It is clear that the effect of Y_0 is statistically significant, no matter which model is used for estimating $\mathrm{E}_0(Y_0 \mid \mathbf{x}, 0)$, and regardless of CDCI=0 or CDCI>0. This indicates, in this application, the assumption (20) that the treatment assignment mechanism does depend on the untreated potential outcome Y_0 is sensible. This analytical result also echoes our earlier arguments that are solely based on clinical experience.

Different ATT estimation results using naive methods and proposed methods are summarized in Table 5. From Table 5, among treated patients with CDCI=0, the naive methods suggest that the 3-year survival rate would have been 5%–20% higher had these patients received $T = 0$ instead. In contrast, the proposed methods suggest little to no difference in 3-year survival between the two treatments for this subgroup. Among treated patients

Table 4: Real Data Application: estimation results $\widehat{\boldsymbol{\theta}}_{\text{eff}}$ by fitting logistic regression models for the propensity score by solving (13), when the data are stratified based on CDCI=0 and CDCI>0. For comparison purposes, the model $\mathrm{E}_0(Y_0 \mid \mathbf{x}, 0)$ is fitted using logistic regression, random forest and XGBoost, respectively.

$\mathrm{E}_0(Y_0 \mid \mathbf{x}, 0)$		Intercept	Y_0 Year of diagnosis	Age	Size of tumor
CDCI=0					
logistic	Est	-1.369	-0.848	0.595	0.568	-0.053
	$\widehat{\mathrm{sd}}_p$	0.054	0.006	0.049	0.045	0.047
	p-value	0.000	0.000	0.000	0.000	0.262
rforest	Est	-1.396	-0.723	0.593	0.570	-0.051
	$\widehat{\mathrm{sd}}_p$	0.081	0.146	0.049	0.042	0.044
	p-value	0.000	0.000	0.000	0.000	0.243
xgboost	Est	-1.366	-0.840	0.594	0.567	-0.052
	$\widehat{\mathrm{sd}}_p$	0.057	0.012	0.051	0.046	0.049
	p-value	0.000	0.000	0.000	0.000	0.284
CDCI>0					
logistic	Est	-1.830	-0.876	0.833	0.476	-0.147
	$\widehat{\mathrm{sd}}_p$	0.067	0.005	0.070	0.053	0.064
	p-value	0.000	0.000	0.000	0.000	0.021
rforest	Est	-1.846	-0.765	0.823	0.477	-0.144
	$\widehat{\mathrm{sd}}_p$	0.066	0.064	0.068	0.055	0.059
	p-value	0.000	0.000	0.000	0.000	0.016
xgboost	Est	-1.826	-0.873	0.830	0.477	-0.148
	$\widehat{\mathrm{sd}}_p$	0.067	0.006	0.066	0.055	0.063
	p-value	0.000	0.000	0.000	0.000	0.019

with CDCI>0, the naive methods suggest a larger increase of 15%–31% in 3-year survival under $T = 0$, whereas the proposed methods indicate a more modest but still statistically significant increase of about 7%–11%.

To put it in a nutshell, if the patients in $T = 1$ group were assigned to $T = 0$, naive methods that assume strong ignorability show that the average 3-year survival rate could increase 20-30%. This seems unrealistic. In contrast, the proposed methods indicate that the conclusion varies according to the level of CDCI: the 3-year survival rate does increase for CDCI>0 patients but only about 7-11%, whereas it almost stays the same for CDCI=0 patients.

Finally, we primarily use perturbation resampling with 500 replications to estimate the standard deviation in this application, reported as $\widehat{\text{sd}}_p$ in the tables. We also validated these SD estimates using a separate nonparametric bootstrap with 500 bootstrap samples; the results were similar and are therefore omitted. In our experience, 500 resamples are sufficient for this application. Below, in Figures 1 and 2, we provide the sensitivity plots of the number of subsamples in the perturbation resampling method for implementing $\widehat{\Delta}_{\text{alt}}$ and $\widehat{\Delta}_{\text{eff}}$ in Table 5, respectively.

Acknowledgments

This work is supported in part by U.S. National Science Foundation (DMS 2310942 and 2610561). The authors would like to thank the Action Editor and the three reviewers for their insightful comments which have helped improve the manuscript substantially.

Table 5: Real Data Application: comparison of estimated ATTs by naive methods (assuming strong ignorability) and proposed methods, using different estimation procedures for nuisance models, after stratifying the data based on CDCI=0 and CDCI>0.

CDCI=0							
		$w(\mathbf{X})$	$\mathrm{E}_0(Y_0 \mid \mathbf{X})$	$\mathrm{E}_1(Y_1 \mid \mathbf{X})$	Est	$\widehat{\mathrm{sd}}_p$	p-value
Naive	$\widehat{\Delta}_{\mathrm{nv1}}$	logistic	-	-	-0.1908	0.0222	0.0000
		rforest	-	-	-0.0539	0.0209	0.0101
		xgboost	-	-	-0.2025	0.0274	0.0000
	$\widehat{\Delta}_{\mathrm{nv2}}$	logistic	logistic	-	-0.1942	0.0235	0.0000
		rforest	rforest	-	-0.2045	0.0218	0.0000
		xgboost	xgboost	-	-0.1882	0.0256	0.0000
		$\pi(Y_0, \mathbf{U}; \boldsymbol{\theta})$	$\mathrm{E}_0(Y_0 \mid \mathbf{X}, 0)$	$\mathrm{E}_1(Y_1 \mid \mathbf{X}, 1)$	Est	$\widehat{\mathrm{sd}}_p$	p-value
Proposed	$\widehat{\Delta}_{\mathrm{alt}}(\widehat{\boldsymbol{\theta}}_{\mathrm{eff}})$	model (20)	logistic	-	0.0118	0.0232	0.6095
			rforest	-	-0.0270	0.0396	0.4948
			xgboost	-	0.0084	0.0222	0.7044
	$\widehat{\Delta}_{\mathrm{eff}}(\widehat{\boldsymbol{\theta}}_{\mathrm{eff}})$	model (20)	logistic	logistic	0.0091	0.0229	0.6900
			rforest	rforest	-0.0465	0.0321	0.1475
			xgboost	xgboost	-0.0103	0.0224	0.6440
CDCI>0							
		$w(\mathbf{X})$	$\mathrm{E}_0(Y_0 \mid \mathbf{X})$	$\mathrm{E}_1(Y_1 \mid \mathbf{X})$	Est	$\widehat{\mathrm{sd}}_p$	p-value
Naive	$\widehat{\Delta}_{\mathrm{nv1}}$	logistic	-	-	-0.2841	0.0249	0.0000
		rforest	-	-	-0.1554	0.0239	0.0000
		xgboost	-	-	-0.3154	0.0307	0.0000
	$\widehat{\Delta}_{\mathrm{nv2}}$	logistic	logistic	-	-0.2838	0.0236	0.0000
		rforest	rforest	-	-0.2872	0.0248	0.0000
		xgboost	xgboost	-	-0.2892	0.0275	0.0000
		$\pi(Y_0, \mathbf{U}; \boldsymbol{\theta})$	$\mathrm{E}_0(Y_0 \mid \mathbf{X}, 0)$	$\mathrm{E}_1(Y_1 \mid \mathbf{X}, 1)$	Est	$\widehat{\mathrm{sd}}_p$	p-value
Proposed	$\widehat{\Delta}_{\mathrm{alt}}(\widehat{\boldsymbol{\theta}}_{\mathrm{eff}})$	model (20)	logistic	-	-0.0726	0.0231	0.0017
			rforest	-	-0.1050	0.0294	0.0004
			xgboost	-	-0.0747	0.0241	0.0020
	$\widehat{\Delta}_{\mathrm{eff}}(\widehat{\boldsymbol{\theta}}_{\mathrm{eff}})$	model (20)	logistic	logistic	-0.0750	0.0240	0.0018
			rforest	rforest	-0.1072	0.0226	0.0000
			xgboost	xgboost	-0.1125	0.0262	0.0000

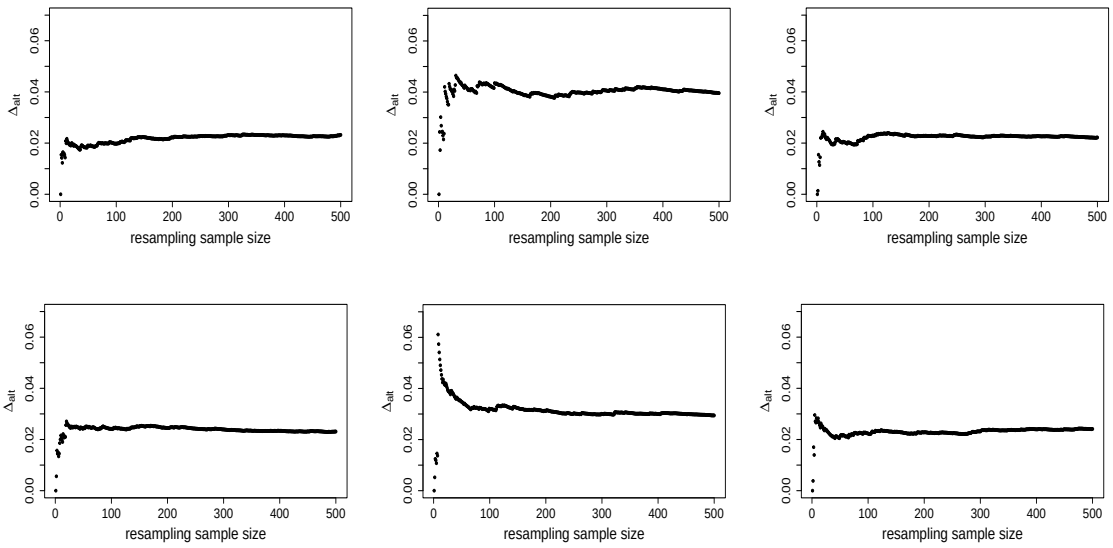


Figure 1: Real Data Application: standard deviations for $\widehat{\Delta}_{\text{alt}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$ based on 500 perturbed resamplings. The first row illustrates the standard deviations for the group CDCI=0 and the second row for CDCI>0. The first, second and third columns summarize the results when logistic regression, random forest and XGBoost are implemented to estimate $\text{E}_0(Y_0 \mid \mathbf{X}, 0)$, respectively.

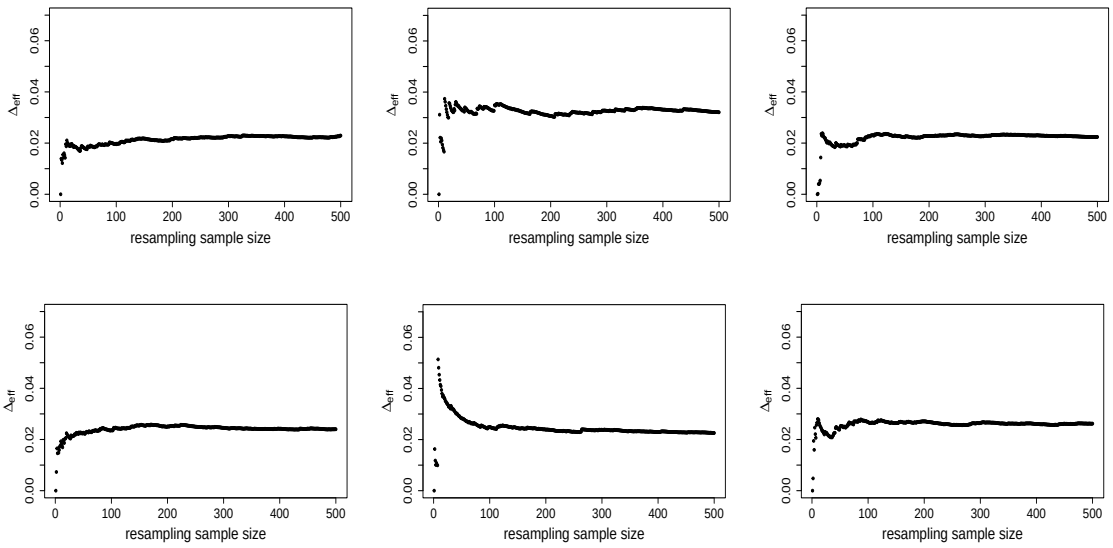


Figure 2: Real Data Application: standard deviations for $\widehat{\Delta}_{\text{eff}}(\widehat{\boldsymbol{\theta}}_{\text{eff}})$ based on 500 perturbed resamplings. The first row illustrates the standard deviations for the group CDCI=0 and the second row for CDCI>0. The first, second and third columns summarize the results when logistic regression, random forest and XGBoost are implemented to estimate $\mathrm{E}_0(Y_0 \mid \mathbf{X}, 0)$ and $\mathrm{E}_1(Y_1 \mid \mathbf{X}, 1)$, respectively.

Appendix A. Proof of Lemma 1

Assume we have two different sets of models $f_1(y_1 \mid \mathbf{x}, 1)$, $f_0(y_0 \mid \mathbf{x}, 0)$, $g(\mathbf{x})$, $\pi(y_0, \mathbf{u})$ and $\widetilde{f}_1(y_1 \mid \mathbf{x}, 1)$, $\widetilde{f}_0(y_0 \mid \mathbf{x}, 0)$, $\widetilde{g}(\mathbf{x})$, $\widetilde{\pi}(y_0, \mathbf{u})$, such that

$$\begin{aligned} & \left[\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u})} \mathrm{d}y_0 - 1\right]^t \left[\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u})} \mathrm{d}y_0\right]^{-1} f_1(y_1 \mid \mathbf{x}, 1)^t f_0(y_0 \mid \mathbf{x}, 0)^{1-t} g(\mathbf{x}) \\ = & \left[\int \frac{\widetilde{f}_0(y_0 \mid \mathbf{x}, 0)}{1-\widetilde{\pi}(y_0, \mathbf{u})} \mathrm{d}y_0 - 1\right]^t \left[\int \frac{\widetilde{f}_0(y_0 \mid \mathbf{x}, 0)}{1-\widetilde{\pi}(y_0, \mathbf{u})} \mathrm{d}y_0\right]^{-1} \widetilde{f}_1(y_1 \mid \mathbf{x}, 1)^t \widetilde{f}_0(y_0 \mid \mathbf{x}, 0)^{1-t} \widetilde{g}(\mathbf{x}). \end{aligned}$$

Taking $t = 0$ and taking integration with respect to Y_0 on both sides of the above equation, one would have

$$\left[\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u})} \mathrm{d}y_0\right]^{-1} g(\mathbf{x}) = \left[\int \frac{\widetilde{f}_0(y_0 \mid \mathbf{x}, 0)}{1-\widetilde{\pi}(y_0, \mathbf{u})} \mathrm{d}y_0\right]^{-1} \widetilde{g}(\mathbf{x}).$$

Taking $t = 1$ and taking integration with respect to Y_1, one would obtain

$$\begin{aligned} & \left[\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u})} \mathrm{d}y_0 - 1\right] \left[\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u})} \mathrm{d}y_0\right]^{-1} g(\mathbf{x}) \\ = & \left[\int \frac{\widetilde{f}_0(y_0 \mid \mathbf{x}, 0)}{1-\widetilde{\pi}(y_0, \mathbf{u})} \mathrm{d}y_0 - 1\right] \left[\int \frac{\widetilde{f}_0(y_0 \mid \mathbf{x}, 0)}{1-\widetilde{\pi}(y_0, \mathbf{u})} \mathrm{d}y_0\right]^{-1} \widetilde{g}(\mathbf{x}). \end{aligned}$$

Hence, immediately, we have

$$\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u})} \mathrm{d}y_0 = \int \frac{\widetilde{f}_0(y_0 \mid \mathbf{x}, 0)}{1-\widetilde{\pi}(y_0, \mathbf{u})} \mathrm{d}y_0, \text{ and } g(\mathbf{x}) = \widetilde{g}(\mathbf{x}).$$

Then, by taking $t = 0$ and $t = 1$ respectively, we also obtain

$$f_0(y_0 \mid \mathbf{x}, 0) = \widetilde{f}_0(y_0 \mid \mathbf{x}, 0), \text{ and } f_1(y_1 \mid \mathbf{x}, 1) = \widetilde{f}_1(y_1 \mid \mathbf{x}, 1).$$

Finally, from

$$\int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\pi(y_0, \mathbf{u})} \mathrm{d}y_0 = \int \frac{f_0(y_0 \mid \mathbf{x}, 0)}{1-\widetilde{\pi}(y_0, \mathbf{u})} \mathrm{d}y_0,$$

and the completeness condition imposed in (4), it is immediate that

$$\pi(y_0, \mathbf{u}) = \widetilde{\pi}(y_0, \mathbf{u}),$$

and this completes the proof.

Appendix B. Proof of Proposition 1

Firstly, it can be easily verified, with the proofs omitted, that $\Lambda_0 \perp \Lambda_1$, $\Lambda_0 \perp \Lambda_2$ and $\Lambda_1 \perp \Lambda_2$, hence the nuisance tangent space $\Lambda = \Lambda_0 \oplus \Lambda_1 \oplus \Lambda_2$. Now we derive $\Lambda^{\perp}$. Denote a generic element as $t\mathbf{g}_1(y_1, \mathbf{x}) + (1-t)\mathbf{g}_0(y_0, \mathbf{x})$. It is straightforward that

$$\Lambda_2^{\perp} = \left[t\mathbf{g}_1(y_1, \mathbf{x}) + (1-t)\mathbf{g}_0(y_0, \mathbf{x}) : \mathrm{E}_1(\mathbf{g}_1 \mid \mathbf{x}, 1)w(\mathbf{x}) + \mathrm{E}_0(\mathbf{g}_0 \mid \mathbf{x}, 0)\{1 - w(\mathbf{x})\} = \mathbf{0}\right],$$

$$\begin{aligned}\Lambda_1^{\perp} &= \{t\mathbf{g}_1(y_1,\mathbf{x}) + (1-t)\mathbf{g}_0(y_0,\mathbf{x}) : \mathbf{g}_1(y_1,\mathbf{x}) = \mathrm{E}_1(\mathbf{g}_1 \mid \mathbf{x},1)\}, \\ &= \{t\mathbf{g}(\mathbf{x}) + (1-t)\mathbf{g}_0(y_0,\mathbf{x})\}.\end{aligned}$$

Hence

$$\Lambda_2^{\perp} \cap \Lambda_1^{\perp} = \left[t\mathbf{g}(\mathbf{x}) + (1-t)\mathbf{g}_0(y_0,\mathbf{x}) : \mathrm{E}_0(\mathbf{g}_0 \mid \mathbf{x},0) = -w(\mathbf{x})\{1-w(\mathbf{x})\}^{-1}\mathbf{g}(\mathbf{x})\right].$$

Keeping this form in mind, we consider the element $t\mathbf{g}(\mathbf{x}) + (1-t)\mathbf{g}_0(y_0,\mathbf{x})$ which satisfies that, for any $\mathbf{c}(y_0,\mathbf{x})$ such that $\mathrm{E}_0\{\mathbf{c}(y_0,\mathbf{x}) \mid \mathbf{x},0\} = \mathbf{0}$, we have

$$\begin{aligned}0 &= \mathrm{E}\left[\{t\mathbf{g}(\mathbf{x}) + (1-t)\mathbf{g}_0(y_0,\mathbf{x})\}^{\mathrm{T}}\left\{(1-t)\mathbf{c}(y_0,\mathbf{x}) + \frac{t-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0(\mathbf{c} \mid \mathbf{x})\right\}\right] \\ &= \mathrm{E}\left[\{1-w(\mathbf{x})\}\mathbf{g}(\mathbf{x})^{\mathrm{T}}\mathrm{E}_0(\mathbf{c} \mid \mathbf{x}) + \{1-w(\mathbf{x})\}\mathrm{E}_0(\mathbf{g}_0^{\mathrm{T}}\mathbf{c} \mid \mathbf{x},0)\right. \\ &\quad \left. -\{1-w(\mathbf{x})\}\mathrm{E}_0(\mathbf{g}_0 \mid \mathbf{x},0)^{\mathrm{T}}\mathrm{E}_0(\mathbf{c} \mid \mathbf{x})\right] \\ &= \mathrm{E}\left[\mathbf{g}(\mathbf{x})^{\mathrm{T}}\mathrm{E}_0(\mathbf{c} \mid \mathbf{x}) + \{1-w(\mathbf{x})\}\mathrm{E}_0(\mathbf{g}_0^{\mathrm{T}}\mathbf{c} \mid \mathbf{x},0)\right] \\ &= \mathrm{E}\left\{\mathrm{E}_0\left(\{1-w(\mathbf{x})\}[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\mathbf{g}(\mathbf{x}) + \mathbf{g}_0(y_0,\mathbf{x})]^{\mathrm{T}}\mathbf{c}(y_0,\mathbf{x}) \mid \mathbf{x},0\right)\right\}.\end{aligned}$$

Therefore, it must be the situation that $\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\mathbf{g}(\mathbf{x}) + \mathbf{g}_0(y_0,\mathbf{x})$ is a function of $\mathbf{x}$ only, say, $\mathbf{v}(\mathbf{x})$:

$$\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\mathbf{g}(\mathbf{x}) + \mathbf{g}_0(y_0,\mathbf{x}) = \mathbf{v}(\mathbf{x}).$$

Taking conditional expectations on both sides, we can solve $\mathbf{v}(\mathbf{x}) = \mathbf{g}(\mathbf{x})$, hence,

$$\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\mathbf{g}(\mathbf{x}) + \mathbf{g}_0(y_0,\mathbf{x}) = \mathbf{g}(\mathbf{x}).$$

Finally,

$$\begin{aligned}\Lambda^{\perp} &= \left([t-\pi(\boldsymbol{\theta}^0)\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}(1-t)]\mathbf{g}(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \in \mathbb{R}^d\right) \\ &= \left\{\frac{t-\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\mathbf{g}(\mathbf{x}) : \mathbf{g}(\mathbf{x}) \in \mathbb{R}^d\right\}.\end{aligned}$$

Appendix C. Proof of Proposition 2

The efficient influence function of Δ, ϕ_{eff}, must be of the form

$$\phi_{\mathrm{eff}} = a_2(\mathbf{x}) + tb_2(y_1,\mathbf{x}) + (1-t)c_2(y_0,\mathbf{x}) + \frac{t-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0(c_2 \mid \mathbf{x}) + \mathbf{H}^{\mathrm{T}}\mathbf{S}_{\mathrm{eff}},$$

where $\mathrm{E}(a_2) = \mathbf{0}$, $\mathrm{E}_1(b_2 \mid \mathbf{x},1) = \mathbf{0}$, $\mathrm{E}_0(c_2 \mid \mathbf{x},0) = \mathbf{0}$ and $\mathbf{H}$ is a $d \times 1$ vector.

We consider parametric submodels $f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\alpha})$, $f_1(y_1 \mid \mathbf{x},1;\boldsymbol{\beta})$ and $f_0(y_0 \mid \mathbf{x},0;\boldsymbol{\gamma})$, which coincide with the true models when $\boldsymbol{\alpha} = \boldsymbol{\alpha}^0$, $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ and $\boldsymbol{\gamma} = \boldsymbol{\gamma}^0$, respectively. Then, the likelihood under the submodel is

$$w(\mathbf{x};\boldsymbol{\gamma},\boldsymbol{\theta})^t\{1-w(\mathbf{x};\boldsymbol{\gamma},\boldsymbol{\theta})\}^{1-t}f_0(y \mid \mathbf{x},0;\boldsymbol{\gamma})^{1-t}f_1(y \mid \mathbf{x},1;\boldsymbol{\beta})^t f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\alpha}).$$

We can derive

$$\begin{aligned}
\mathbf{S}_{\boldsymbol{\alpha}}(\mathbf{x}) &= \frac{\partial \log f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} \in \Lambda_2, \\
\mathbf{S}_{\boldsymbol{\beta}}(t,y,\mathbf{x}) &= t\frac{\partial \log f_1(y_1 \mid \mathbf{x},1;\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \in \Lambda_1, \\
\mathbf{S}_{\boldsymbol{\gamma}}(t,y,\mathbf{x}) &= (1-t)\frac{\partial \log f_0(y_0 \mid \mathbf{x},0;\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \\
&\quad + \frac{t-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0\left\{\frac{\partial \log f_0(y_0 \mid \mathbf{x},0;\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \mid \mathbf{x}\right\} \in \Lambda_0, \\
\mathbf{S}_{\boldsymbol{\theta}}(t,\mathbf{x}) &= \frac{t-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0\left[\{1-\pi(\boldsymbol{\theta}^0)\}^{-1}\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0) \mid \mathbf{x}\right],
\end{aligned}$$

where the partial derivatives are evaluated at the true values if not otherwise specified. Next, we can derive

$$\begin{aligned}
\mathrm{E}(\phi_{\mathrm{eff}}\mathbf{S}_{\boldsymbol{\alpha}}) &= \mathrm{E}\left\{a_2(\mathbf{x})\,\partial \log f_{\mathbf{X}}/\partial \boldsymbol{\alpha}\right\}, \\
\mathrm{E}(\phi_{\mathrm{eff}}\mathbf{S}_{\boldsymbol{\beta}}) &= \mathrm{E}\{\mathrm{E}_1(b_2\,\partial \log f_1/\partial \boldsymbol{\beta}^{\mathrm{T}} \mid \mathbf{x},1)w(\mathbf{x})\} = \mathrm{E}\{w(\mathbf{x})b_2\,\partial \log f_1/\partial \boldsymbol{\beta}^{\mathrm{T}}\}, \\
\mathrm{E}(\phi_{\mathrm{eff}}\mathbf{S}_{\boldsymbol{\gamma}}) &= \mathrm{E}\left(\left[\{1-\pi(\boldsymbol{\theta}^0)\}c_2 + \frac{1-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0(c_2 \mid \mathbf{x})\right]\partial \log f_0/\partial \boldsymbol{\gamma}\right), \\
\mathrm{E}(\phi_{\mathrm{eff}}\mathbf{S}_{\boldsymbol{\theta}}^{\mathrm{T}}) &= \mathrm{E}\left[\frac{1-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0\left(c_2 \mid \mathbf{x}\right)\mathrm{E}_0\left\{\frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)^{\mathrm{T}}}{1-\pi(\boldsymbol{\theta}^0)} \mid \mathbf{x}\right\}\right] + \mathbf{H}^{\mathrm{T}}\mathbf{M}.
\end{aligned}$$

Now, we rewrite

$$\Delta = \frac{A(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta}) - B(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\theta})}{C(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\theta})},$$

where

$$\begin{aligned}
A(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\theta}) &= \iint y_1 w(\mathbf{x};\boldsymbol{\gamma},\boldsymbol{\theta}) f_1(y_1 \mid \mathbf{x},1;\boldsymbol{\beta}) f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\alpha})\mathrm{d}y_1\mathrm{d}\mathbf{x} = \mathrm{E}(TY_1), \\
B(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\theta}) &= \iint y_0 \pi(\boldsymbol{\theta})\{1-\pi(\boldsymbol{\theta})\}^{-1}\{1-w(\mathbf{x};\boldsymbol{\gamma},\boldsymbol{\theta})\} f_0(y_0 \mid \mathbf{x},0;\boldsymbol{\gamma}) f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\alpha})\mathrm{d}y_0\mathrm{d}\mathbf{x} \\
&= \mathrm{E}(TY_0), \\
C(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\theta}) &= \int w(\mathbf{x};\boldsymbol{\gamma},\boldsymbol{\theta}) f_{\mathbf{X}}(\mathbf{x};\boldsymbol{\alpha})\mathrm{d}\mathbf{x} = \mathrm{E}(T) = p.
\end{aligned}$$

For $\partial\Delta/\partial\boldsymbol{\alpha}$, we have

$$\frac{\partial A}{\partial \boldsymbol{\alpha}} = \mathrm{E}\left\{TY_1\frac{\partial \log f_{\mathbf{X}}}{\partial \boldsymbol{\alpha}}\right\}, \frac{\partial B}{\partial \boldsymbol{\alpha}} = \mathrm{E}\left\{TY_0\frac{\partial \log f_{\mathbf{X}}}{\partial \boldsymbol{\alpha}}\right\}, \frac{\partial C}{\partial \boldsymbol{\alpha}} = \mathrm{E}\left\{T\frac{\partial \log f_{\mathbf{X}}}{\partial \boldsymbol{\alpha}}\right\}.$$

Hence

$$\frac{\partial \Delta}{\partial \boldsymbol{\alpha}} = \mathrm{E}\left[\left\{\frac{T}{p}(Y_1 - Y_0) - \frac{T}{p}\Delta^0\right\}\frac{\partial \log f_{\mathbf{X}}}{\partial \boldsymbol{\alpha}}\right].$$

This implies

$$\begin{aligned} a_2(\mathbf{x}) &= \mathrm{E}\left\{\frac{T}{p}(Y_1 - Y_0) - \frac{T}{p}\Delta^0 \mid \mathbf{x}\right\} \\ &= \frac{1}{p}\mathrm{E}_1(Y_1 \mid \mathbf{x}, 1)w(\mathbf{x}) - \frac{1}{p}\mathrm{E}_0\left\{\frac{\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}Y_0 \mid \mathbf{x}, 0\right\}\{1 - w(\mathbf{x})\} - \frac{\Delta^0}{p}w(\mathbf{x}), \\ &= \frac{1}{p}\mathrm{E}_1(Y_1 \mid \mathbf{x}, 1)w(\mathbf{x}) - \frac{1}{p}\mathrm{E}_0\{\pi(\boldsymbol{\theta}^0)Y_0 \mid \mathbf{x}\} - \frac{\Delta^0}{p}w(\mathbf{x}). \end{aligned}$$

For $\partial\Delta/\partial\boldsymbol{\beta}$, we have

$$\frac{\partial\Delta}{\partial\boldsymbol{\beta}} = \mathrm{E}\left\{\frac{w(\mathbf{X})}{p}Y_1\frac{\partial \log f_1}{\partial\boldsymbol{\beta}}\right\},$$

hence $b_2 - y_1/p$ must be a function of $\mathbf{x}$ alone. From this, we can solve

$$b_2 = \frac{1}{p}y_1 - \frac{1}{p}\mathrm{E}_1(Y_1 \mid \mathbf{x}, 1).$$

For $\partial\Delta/\partial\boldsymbol{\gamma}$, we can derive

$$\begin{aligned} \frac{\partial A}{\partial\boldsymbol{\gamma}} &= \mathrm{E}[\{1 - w(\mathbf{X})\}\mathrm{E}_1(Y_1 \mid \mathbf{X}, 1)\ \partial \log f_0/\partial\boldsymbol{\gamma}], \\ \frac{\partial B}{\partial\boldsymbol{\gamma}} &= \mathrm{E}\left([\pi(\boldsymbol{\theta}^0)Y_0 - \mathrm{E}_0\{\pi(\boldsymbol{\theta}^0)Y_0 \mid \mathbf{X}\}]\ \partial \log f_0/\partial\boldsymbol{\gamma}\right), \\ \frac{\partial p}{\partial\boldsymbol{\gamma}} &= \mathrm{E}[\{1 - w(\mathbf{X})\}\ \partial \log f_0/\partial\boldsymbol{\gamma}]. \end{aligned}$$

Since $\partial \log f_0/\partial\boldsymbol{\gamma}$ satisfies $\mathrm{E}_0[\{1 - \pi(\boldsymbol{\theta}^0)\}\ \partial \log f_0/\partial\boldsymbol{\gamma} \mid \mathbf{x}] = \mathbf{0}$, it must be the situation that

$$\begin{aligned} &\frac{1 - w(\mathbf{x})}{1 - \pi(\boldsymbol{\theta}^0)}\mathrm{E}_1(Y_1 \mid \mathbf{x}, 1) - \frac{\pi(\boldsymbol{\theta}^0)}{1 - \pi(\boldsymbol{\theta}^0)}Y_0 + \frac{\mathrm{E}_0\{\pi(\boldsymbol{\theta}^0)Y_0 \mid \mathbf{x}\}}{1 - \pi(\boldsymbol{\theta}^0)} - \frac{1 - w(\mathbf{x})}{1 - \pi(\boldsymbol{\theta}^0)}\Delta^0 - pc_2 \\ &-\frac{p}{1 - \pi(\boldsymbol{\theta}^0)}\frac{1 - w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0(c_2 \mid \mathbf{x}) \end{aligned}$$

is a function of $\mathbf{x}$ only, say, $e(\mathbf{x})$, then taking conditional expectations on both sides, we must have

$$e(\mathbf{x}) = \mathrm{E}_1(Y_1 \mid \mathbf{x}, 1) - \Delta^0 - \frac{p}{w(\mathbf{x})}\mathrm{E}_0(c_2 \mid \mathbf{x}).$$

Then taking expectations on both sides again, and after some tedious calculations, we will have

$$\begin{aligned} \mathrm{E}_0(c_2 \mid \mathbf{x}) &= \frac{w(\mathbf{x})}{1 - w(\mathbf{x})}\frac{1}{B(\mathbf{x})}\frac{1}{p}[\{1 - w(\mathbf{x})\}B(\mathbf{x}) - w(\mathbf{x})]\{\mathrm{E}_1(Y_1 \mid \mathbf{x}, 1) - \Delta^0\} \\ &\quad -\frac{w(\mathbf{x})}{1 - w(\mathbf{x})}\frac{1}{B(\mathbf{x})}\frac{1}{p}\mathrm{E}_0\left\{\frac{\pi(\boldsymbol{\theta}^0)}{1 - \pi(\boldsymbol{\theta}^0)}Y_0 \mid \mathbf{x}\right\} \end{aligned}$$

$$
\begin{aligned}
& + \frac{w(\mathbf{x})}{1-w(\mathbf{x})}\frac{1}{B(\mathbf{x})}\frac{1}{p}\{1+B(\mathbf{x})\}\mathrm{E}_0\{\pi(\boldsymbol{\theta})Y_0 \mid \mathbf{x}\}, \text{ and} \\
c_2 \;=\; & \frac{1}{p}\frac{\pi(\boldsymbol{\theta}^0)-w(\mathbf{x})}{1-\pi(\boldsymbol{\theta}^0)}\frac{w(\mathbf{x})}{1-w(\mathbf{x})}\frac{1}{B(\mathbf{x})}\{\mathrm{E}_1(Y_1 \mid \mathbf{x},1)-\Delta^0\} \\
& + \frac{1}{p}\mathrm{E}_0\{\pi(\boldsymbol{\theta}^0)Y_0 \mid \mathbf{x}\}\left\{\frac{1}{1-w(\mathbf{x})} - \frac{\pi(\boldsymbol{\theta}^0)-w(\mathbf{x})}{\{1-\pi(\boldsymbol{\theta}^0)\}\{1-w(\mathbf{x})\}B(\mathbf{x})}\right\} \\
& - \frac{1}{p}\frac{\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}Y_0 + \frac{1}{p}\frac{\pi(\boldsymbol{\theta}^0)-w(\mathbf{x})}{\{1-\pi(\boldsymbol{\theta}^0)\}\{1-w(\mathbf{x})\}B(\mathbf{x})}\mathrm{E}_0\left\{\frac{\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}Y_0 \mid \mathbf{x}\right\}.
\end{aligned}
$$

Now we put the first three terms in ϕ_{eff}, $a_2(\mathbf{x})+tb_2(y_1,\mathbf{x})+(1-t)c_2(y_0,\mathbf{x})+\frac{t-w(\mathbf{x})}{w(\mathbf{x})}\mathrm{E}_0(c_2 \mid \mathbf{x})$, together, it equals

$$
\begin{aligned}
& \frac{1}{p}\frac{t-\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\left(y - \frac{1}{B(\mathbf{x})}\left[w(\mathbf{x})\mathrm{E}_1(Y_1 \mid \mathbf{x},1) + \mathrm{E}_0\left\{\frac{\pi(\boldsymbol{\theta}^0)^2}{1-\pi(\boldsymbol{\theta}^0)}Y_0 \mid \mathbf{x}\right\}\right]\right) \\
& - \frac{\Delta^0}{p}\left\{t - \frac{t-\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\frac{w(\mathbf{x})}{B(\mathbf{x})}\right\} \\
= & \frac{1}{p}\frac{t-\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\left(y - \frac{1}{B(\mathbf{x})}\left[w(\mathbf{x})\mathrm{E}_1(Y_1 \mid \mathbf{x},1) + \{1-w(\mathbf{x})\}\mathrm{E}_0\left\{\frac{\pi(\boldsymbol{\theta}^0)^2}{\{1-\pi(\boldsymbol{\theta}^0)\}^2}Y_0 \mid \mathbf{x},0\right\}\right]\right) \\
& - \frac{\Delta^0}{p}\left\{t - \frac{t-\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\frac{w(\mathbf{x})}{B(\mathbf{x})}\right\} \\
= & \phi(\mathbf{w};\boldsymbol{\theta}^0,\Delta^0).
\end{aligned}
$$

For $\partial\Delta/\partial\boldsymbol{\theta}$, we can derive

$$
\begin{aligned}
\frac{\partial A}{\partial \boldsymbol{\theta}} &= \mathrm{E}\left[\{1-w(\mathbf{x})\}\,\mathrm{E}_1(Y_1 \mid \mathbf{x},1)\frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\right], \\
\frac{\partial B}{\partial \boldsymbol{\theta}} &= \mathrm{E}\left\{Y_0\frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\right\} - \mathrm{E}\left[\mathrm{E}_0\left\{Y_0\pi(\boldsymbol{\theta}^0) \mid \mathbf{x}\right\}\frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\right], \\
\frac{\partial p}{\partial \boldsymbol{\theta}} &= \mathrm{E}\left[\frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\{1-w(\mathbf{x})\}\right].
\end{aligned}
$$

Hence,

$$
\frac{\partial\Delta}{\partial\boldsymbol{\theta}} = p^{-1}\mathrm{E}\left(\left[\{1-w(\mathbf{x})\}\left\{\mathrm{E}_1(Y_1 \mid \mathbf{x},1)-\Delta^0\right\} - Y_0 + \mathrm{E}_0\left\{Y_0\pi(\boldsymbol{\theta}^0) \mid \mathbf{x}\right\}\right]\frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}\right).
$$

Then plugging into $\mathrm{E}_0(c_2 \mid \mathbf{x})$ in $\mathrm{E}(\phi_{\text{eff}}\mathbf{S}_{\boldsymbol{\theta}}^{\mathrm{T}}) = \partial\Delta/\partial\boldsymbol{\theta}^{\mathrm{T}}$, we get

$$
\begin{aligned}
& \mathrm{E}\left\{\left(\frac{1}{B(\mathbf{x})}\frac{1}{p}\left[\{1-w(\mathbf{x})\}B(\mathbf{x}) - w(\mathbf{x})\right]\left\{\mathrm{E}_1(Y_1 \mid \mathbf{x},1)-\Delta^0\right\} - \frac{1}{B(\mathbf{x})}\frac{1}{p}\mathrm{E}_0\left\{\frac{\pi(\boldsymbol{\theta}^0)}{1-\pi(\boldsymbol{\theta}^0)}Y_0 \mid \mathbf{x}\right\}\right.\right. \\
& \left.\left. + \frac{1}{B(\mathbf{x})}\frac{1}{p}\{1+B(\mathbf{x})\}\,\mathrm{E}_0\left\{\pi(\boldsymbol{\theta}^0)Y_0 \mid \mathbf{x}\right\}\right)\mathbf{A}(\mathbf{x})^{\mathrm{T}}\right\} + \mathbf{H}^{\mathrm{T}}\mathbf{M} \\
= \; & \frac{1}{p}\mathrm{E}\left(\left[\{1-w(\mathbf{x})\}\left\{\mathrm{E}_1(Y_1 \mid \mathbf{x},1)-\Delta^0\right\} - Y_0 + \mathrm{E}_0\left\{Y_0\pi(\boldsymbol{\theta}^0) \mid \mathbf{x}\right\}\right]\frac{\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0)^{\mathrm{T}}}{1-\pi(\boldsymbol{\theta}^0)}\right).
\end{aligned}
$$

After simplifying the above equation, we get $p\mathbf{H}^{\mathrm{T}}\mathbf{M} = (\mathbf{D}-\mathbf{Q})^{\mathrm{T}}$, i.e., $p\mathbf{H}^{\mathrm{T}} = (\mathbf{D}-\mathbf{Q})^{\mathrm{T}}\mathbf{M}^{-1}$.

Appendix D. Proof of Theorem 1

Lemma A1 *The score function* $\mathbf{G}$ *satisfies the Neyman orthogonality condition, i.e.,*

$$\partial_r \mathrm{E}\left[\mathbf{G}\{\mathbf{W}; \boldsymbol{\psi}^0, \boldsymbol{\eta}^0 + r(\boldsymbol{\eta} - \boldsymbol{\eta}^0)\}\right]|_{r=0} = \mathbf{0}, \quad \textit{for any } \boldsymbol{\eta} \in \mathcal{T}.$$

Proof of Lemma A1. Let

$$\begin{aligned}
\eta_1^0(\mathbf{x}) &= \mathrm{E}_1(Y_1 \mid \mathbf{x}, 1), \\
\eta_2^0(\mathbf{x}) &= \mathrm{E}_0[\{1 - \pi(\boldsymbol{\theta}^0)\}^{-1} \mid \mathbf{x}, 0], \\
\eta_3^0(\mathbf{x}) &= \mathrm{E}_0[\{1 - \pi(\boldsymbol{\theta}^0)\}^{-2}\pi(\boldsymbol{\theta}^0) \mid \mathbf{x}, 0], \\
\eta_4^0(\mathbf{x}) &= \mathrm{E}_0[Y_0\{1 - \pi(\boldsymbol{\theta}^0)\}^{-2}\pi(\boldsymbol{\theta}^0)^2 \mid \mathbf{x}, 0], \\
\boldsymbol{\eta}_5^0(\mathbf{x}) &= \mathrm{E}_0[\{1 - \pi(\boldsymbol{\theta}^0)\}^{-2}\dot{\boldsymbol{\pi}}(\boldsymbol{\theta}^0) \mid \mathbf{x}, 0].
\end{aligned}$$

Note that

$$\mathbf{S}_{\mathrm{eff}}(\mathbf{w}; \boldsymbol{\theta}, \boldsymbol{\eta}) = \frac{t - \pi(\boldsymbol{\theta})}{1 - \pi(\boldsymbol{\theta})} \frac{\boldsymbol{\eta}_5(\mathbf{x})}{\eta_3(\mathbf{x})}.$$

We have

$$\begin{aligned}
&\partial_r \mathrm{E}\left[\mathbf{S}_{\mathrm{eff}}(\mathbf{W}; \boldsymbol{\theta}, \boldsymbol{\eta}^0 + r(\boldsymbol{\eta} - \boldsymbol{\eta}^0))\right]|_{r=0} \\
&= \mathrm{E}\left[\frac{T - \pi(\boldsymbol{\theta}^0)}{1 - \pi(\boldsymbol{\theta}^0)}\left(\frac{\boldsymbol{\eta}_5(\mathbf{X}) - \boldsymbol{\eta}_5^0(\mathbf{X})}{\eta_3^0(\mathbf{X})} + \frac{\boldsymbol{\eta}_5^0(\mathbf{X})}{\{\eta_3^0(\mathbf{X})\}^2}\left\{\eta_3(\mathbf{X}) - \eta_3^0(\mathbf{X})\right\}\right)\right] \\
&= \mathrm{E}\left[\frac{\mathrm{pr}(T = 1 \mid \mathbf{X}, Y_0) - \pi(\boldsymbol{\theta}^0)}{1 - \pi(\boldsymbol{\theta}^0)}\left(\frac{\boldsymbol{\eta}_5(\mathbf{X}) - \boldsymbol{\eta}_5^0(\mathbf{X})}{\eta_3^0(\mathbf{X})} + \frac{\boldsymbol{\eta}_5^0(\mathbf{X})}{\{\eta_3^0(\mathbf{X})\}^2}\left\{\eta_3(\mathbf{X}) - \eta_3^0(\mathbf{X})\right\}\right)\right] \\
&= \mathbf{0},
\end{aligned}$$

for any $\boldsymbol{\eta} \in \mathcal{T}$, where the second equality is by the law of total expectation, and the last equality is by (5). We can also rewrite ϕ as

$$\phi(\mathbf{w}; \boldsymbol{\psi}, \boldsymbol{\eta}) = p^{-1}\left[\frac{t - \pi(\boldsymbol{\theta})}{1 - \pi(\boldsymbol{\theta})}\left(Y - \frac{\{\eta_1(\mathbf{x}) - \Delta\}\{\eta_2(\mathbf{x}) - 1\} + \eta_4(\mathbf{x})}{\eta_3(\mathbf{x})}\right) - \Delta t\right].$$

Then, we have

$$\begin{aligned}
&\partial_r \mathrm{E}\left[\phi(\mathbf{W}; \boldsymbol{\psi}^0, \boldsymbol{\eta}^0 + r(\boldsymbol{\eta} - \boldsymbol{\eta}^0))\right]|_{r=0} \\
&= \mathrm{E}\Bigg[\frac{T - \pi(\boldsymbol{\theta}^0)}{1 - \pi(\boldsymbol{\theta}^0)}\Bigg(-\frac{\eta_2^0(\mathbf{X}) - 1}{\eta_3^0(\mathbf{X})}\left\{\eta_1(\mathbf{X}) - \eta_1^0(\mathbf{X})\right\} - \frac{\eta_1^0(\mathbf{X}) - \Delta^0}{\eta_3^0(\mathbf{X})}\left\{\eta_2(\mathbf{X}) - \eta_2^0(\mathbf{X})\right\} \\
&\qquad + \frac{\{\eta_1^0(\mathbf{X}) - \Delta^0\}\{\eta_2^0(\mathbf{X}) - 1\} + \eta_4^0(\mathbf{X})}{\eta_3^0(\mathbf{X})^2}\left\{\eta_3(\mathbf{X}) - \eta_3^0(\mathbf{X})\right\} - \frac{1}{\eta_3^0(\mathbf{X})}\left\{\eta_4(\mathbf{X}) - \eta_4^0(\mathbf{X})\right\}\Bigg)\Bigg] \\
&= 0,
\end{aligned}$$

for any $\boldsymbol{\eta} \in \mathcal{T}$, where the last equality is due to the fact

$$\mathrm{E}\left\{\frac{T - \pi(\boldsymbol{\theta}^0)}{1 - \pi(\boldsymbol{\theta}^0)} \mid \mathbf{X}\right\} = 1 - \mathrm{E}\left\{\frac{1 - T}{1 - \pi(\boldsymbol{\theta}^0)} \mid \mathbf{X}\right\} = 0.$$

This completes the proof.

□

We also introduce the following high-level regularity conditions, parallel to Assumptions 3.3 and 3.4 in Section 3.3 of Chernozhukov et al. (2018). Specifically, Assumption A1 is mild and fairly standard in problems with moment conditions. It concerns the smoothness of the estimating equations and the unique identification of the parameter, and can be viewed as a restatement and strengthening of Assumption 1 for the purpose of inference. Assumption A2 provides a high-level characterization of the quality of the nuisance estimation, in that it is formulated in terms of the intrinsic rates most closely tied to the statistical problem at hand, rather than directly in terms of the norm $\|\cdot\|_{P,2}$ for $\widehat{\boldsymbol{\eta}}$. However, as remarked by Chernozhukov et al. (2018), the rate conditions in (A.1)–(A.3) can generally be implied from the crude but more intuitive requirement on the convergence rate of $\widehat{\boldsymbol{\eta}}$, namely that given in Assumption 2. As discussed there, such convergence rate is satisfied by many commonly used machine learning methods.

Assumption A1 *(Non-linear moment condition problem with Neyman orthogonality) It holds that:*

(a) $\mathrm{E}\left\{\mathbf{G}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\right\}=\mathbf{0}$ *and* $\mathcal{B}$ *contains a ball of radius* $\Omega(n^{-1/2}\log n)$ *centered at* $\boldsymbol{\psi}^0$.

(b) The map $(\boldsymbol{\psi},\boldsymbol{\eta})\to \mathrm{E}\left\{\mathbf{G}(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta})\right\}$ *is twice continuously Gateaux-differentiable on* $\mathcal{B}\times\mathcal{T}$.

(c) For all $\boldsymbol{\psi}\in\mathcal{B}$, $2\|\mathrm{E}\left\{\mathbf{G}(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta}^0)\right\}\|\geq\|\mathbf{J}(\boldsymbol{\psi}-\boldsymbol{\psi}^0)\|\wedge c_0$ *is satisfied, and*

$$\mathbf{J}=\partial_{\boldsymbol{\psi}^{\mathrm{T}}}\mathrm{E}\left\{\mathbf{G}(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta}^0)\right\}\mid_{\boldsymbol{\psi}=\boldsymbol{\psi}^0}$$

has singular values between c_0 *and* c_1. *Here* c_0 *and* c_1 *are some finite positive constants.*

Assumption A2 *(Score regularity and requirements on the quality of estimation of nuisance parameters) It holds that:*

(a) $\widehat{\boldsymbol{\eta}}^{[-k]}$ *belongs to the realization set* $\mathcal{T}_n$ *for each* $k\in\{1,2,\ldots,K\}$, *with probability tending to 1 where* $\mathcal{T}_n$ *satisfies* $\boldsymbol{\eta}^0\in\mathcal{T}_n$ *and conditions provided as below.*

(b) The space of $\boldsymbol{\psi}$, $\mathcal{B}$ *is bounded and for each* $\boldsymbol{\eta}\in\mathcal{T}_n$, *the function class* $\mathcal{F}_{1,\boldsymbol{\eta}}=\{G_j(\cdot,\boldsymbol{\psi},\boldsymbol{\eta}):j=1,2,\ldots,(d+1);\boldsymbol{\psi}\in\mathcal{B}\}$ *is suitably measurable and its uniform covering entropy obeys*

$$\sup_Q \log N(\epsilon\|\mathbf{F}_{1,\boldsymbol{\eta}}\|_{Q,2},\mathcal{F}_{1,\boldsymbol{\eta}},\|\cdot\|_{Q,2})\leq\nu\log(a/\epsilon),\quad \textit{for all}\quad 0\leq\epsilon\leq 1,$$

where $\mathbf{F}_{1,\boldsymbol{\eta}}$ *is a measurable envelope for* $\mathcal{F}_{1,\boldsymbol{\eta}}$ *satisfying* $\|\mathbf{F}_{1,\boldsymbol{\eta}}(\cdot)\|_{P,q}\leq c_1$. *Here* $q>2$, $c_1>0$, $a>1$ *and* $\nu>0$ *are some finite constants.*

(c) There exist sequences $\tau_n>0$ *and* $\delta_n>0$ *that converge to zero such that:*

$$\sup_{\boldsymbol{\eta}\in\mathcal{T}_n,\boldsymbol{\psi}\in\mathcal{B}}\|\mathrm{E}\left\{\mathbf{G}(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta})-\mathrm{E}\mathbf{G}(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta}^0)\right\}\|\leq\tau_n\delta_n,\tag{A.1}$$

$$\sup_{\boldsymbol{\eta}\in\mathcal{T}_n,\|\boldsymbol{\psi}-\boldsymbol{\psi}^0\|\leq\tau_n}\left[\mathrm{E}\left\{\|\mathbf{G}(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta})-\mathbf{G}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\|^2\right\}\right]^{1/2}\leq\delta_n, \tag{A.2}$$

$$\sup_{r\in(0,1),\boldsymbol{\eta}\in\mathcal{T}_n,\|\boldsymbol{\psi}-\boldsymbol{\psi}^0\|\leq\tau_n}\|\partial_r^2\mathrm{E}\left[\mathbf{G}\left\{\mathbf{W};\boldsymbol{\psi}^0+r(\boldsymbol{\psi}-\boldsymbol{\psi}^0),\boldsymbol{\eta}^0+r(\boldsymbol{\eta}-\boldsymbol{\eta}^0)\right\}\right]\|\leq\delta_n n^{-1/2}. \tag{A.3}$$

(d) The variance of the score is non-degenerate. All eigenvalues of the matrix

$$\mathrm{E}\left\{\mathbf{G}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\mathbf{G}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)^{\mathrm{T}}\right\}$$

are bounded from below by c_0.

Proof of Theorem 1. Under Assumptions 1, 2, A1, and A2, together with Lemma A1, Theorem 3.3 and Lemma 6.3 of Chernozhukov et al. (2018) show that $\sqrt{n}(\widehat{\boldsymbol{\psi}}-\boldsymbol{\psi}^0)$ is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $\mathbf{V}=\mathbf{J}^{-1}\mathrm{E}\{\mathbf{G}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\mathbf{G}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)^{\mathrm{T}}\}(\mathbf{J}^{-1})^{\mathrm{T}}$, where $\mathbf{J}=\partial_{\boldsymbol{\psi}^{\mathrm{T}}}\mathrm{E}\{\mathbf{G}(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta}^0)\}\mid_{\boldsymbol{\psi}=\boldsymbol{\psi}^0}$. Specifically,

$$\begin{aligned}\mathbf{J} &= \begin{bmatrix}\partial_{\boldsymbol{\theta}^{\mathrm{T}}}\mathrm{E}\{\mathbf{S}_{\text{eff}}(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta}^0)\} & \mathbf{0}\\ \partial_{\boldsymbol{\theta}^{\mathrm{T}}}\mathrm{E}\{\phi(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta}^0)\} & \partial_{\Delta}\mathrm{E}\{\phi(\mathbf{W};\boldsymbol{\psi},\boldsymbol{\eta}^0)\}\end{bmatrix}\Bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0}\\ &= \begin{bmatrix}\mathbf{J}_{11} & \mathbf{0}\\ \mathbf{J}_{21}^{\mathrm{T}} & J_{22}\end{bmatrix},\end{aligned}$$

where, $\mathbf{J}_{11}=-\mathbf{M}$, $\mathbf{J}_{21}=p^{-1}(\mathbf{D}-\mathbf{Q})$ and

$$J_{22} = \mathrm{E}\left\{-\frac{1}{p}T+\frac{1}{p}\frac{T-\pi(\boldsymbol{\theta})}{1-\pi(\boldsymbol{\theta})}\frac{w(\mathbf{X})}{B(\mathbf{X})}\right\}\Bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}^0}=-1.$$

Further,

$$\mathbf{J}^{-1} = \begin{pmatrix}\mathbf{J}_{11}^{-1} & \mathbf{0}\\ \mathbf{0}^{\mathrm{T}} & -1\end{pmatrix}\begin{pmatrix}\mathbf{I} & \mathbf{0}\\ -\mathbf{J}_{21}^{\mathrm{T}}\mathbf{J}_{11}^{-1} & 1\end{pmatrix}=\begin{pmatrix}\mathbf{J}_{11}^{-1} & \mathbf{0}\\ \mathbf{J}_{21}^{\mathrm{T}}\mathbf{J}_{11}^{-1} & -1\end{pmatrix}=\begin{pmatrix}-\mathbf{M}^{-1} & \mathbf{0}\\ -\mathbf{J}_{21}^{\mathrm{T}}\mathbf{M}^{-1} & -1\end{pmatrix}$$

and for simplicity, we can write

$$\begin{aligned}&\mathrm{E}\left\{\mathbf{G}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\mathbf{G}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)^{\mathrm{T}}\right\}\\ =\;&\begin{bmatrix}\mathrm{E}\{\mathbf{S}_{\text{eff}}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\mathbf{S}_{\text{eff}}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)^{\mathrm{T}}\} & \mathrm{E}\{\mathbf{S}_{\text{eff}}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\phi(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\}\\ \mathrm{E}\{\phi(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\mathbf{S}_{\text{eff}}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)^{\mathrm{T}}\} & \mathrm{E}\{\phi(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)^2\}\end{bmatrix}\\ =\;&\begin{pmatrix}\mathbf{M} & \mathbf{A}_{12}\\ \mathbf{A}_{12}^{\mathrm{T}} & A_{22}\end{pmatrix}=\begin{pmatrix}\mathbf{M} & \mathbf{0}\\ \mathbf{0}^{\mathrm{T}} & A_{22}\end{pmatrix}.\end{aligned}$$

Here $\mathbf{A}_{12}=\mathbf{0}$ as $\mathbf{S}_{\text{eff}}(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\in\Lambda^{\perp}$ and $\phi(\mathbf{W};\boldsymbol{\psi}^0,\boldsymbol{\eta}^0)\in\Lambda$. Thus, we get

$$\mathbf{V} = \begin{pmatrix}\mathbf{M}^{-1} & \mathbf{M}^{-1}\mathbf{J}_{21}\\ \mathbf{J}_{21}^{\mathrm{T}}\mathbf{M}^{-1} & \mathbf{J}_{21}^{\mathrm{T}}\mathbf{M}^{-1}\mathbf{J}_{21}+A_{22}\end{pmatrix}.$$

This completes the proof.

□

Appendix E. Nonparametric Estimation

As shown in Appendix A, under assumptions (4) and (5), $\pi(y_0, \mathbf{u})$ is nonparametrically identified by the integral equation (3). In this section, we propose a nonparametric approach to estimating $\pi(\cdot)$, and subsequently Δ, to mitigate the risk of model misspecification. For a square-integrable function $g(y_0, \mathbf{u})$, we define $m(\mathbf{x}; g) = \mathrm{E}\{(1-T)g(Y_0, \mathbf{U}) \mid \mathbf{x}\}$. It follows from equation (3) that $\pi(\cdot)$ satisfies $m\{\mathbf{X}; (1-\pi)^{-1}\} = 1$ almost surely. Consequently, $\pi(\cdot)$ minimizes

$$\inf_{\tilde{\pi} \in \Pi} \mathrm{E}\left[m\{\mathbf{X}; (1-\tilde{\pi})^{-1}\} - 1\right]^2, \tag{A.4}$$

where Π denotes a function space of $(y_0, \mathbf{u})$. This motivates estimating π by solving the empirical analogue of (A.4).

We first estimate m using a series estimator (Newey, 1997). Specifically, let $\{p_j(\cdot)\}_{j=1}^{\infty}$ denote a sequence of known basis functions (such as power series, splines, Fourier series, etc.), with the property that its linear combination can approximate any square-integrable real-valued function of $\mathbf{x}$ well. For a detailed discussion on the construction of basis functions, we refer readers to Chen (2007); Belloni et al. (2015). Denote $\mathbf{p}^{k_n}(\mathbf{x}) = \big(p_1(\mathbf{x}), \ldots, p_{k_n}(\mathbf{x})\big)^{\mathrm{T}}$, and $\mathbf{P} = \big(\mathbf{p}^{k_n}(\mathbf{x}_1), \ldots, \mathbf{p}^{k_n}(\mathbf{x}_n)\big)^{\mathrm{T}} \in \mathbb{R}^{n \times k_n}$, where k_n could increase with the sample size n. We then estimate m by

$$\widehat{m}(\mathbf{x}; g) = \mathbf{p}^{k_n}(\mathbf{x})^{\mathrm{T}} \big(\mathbf{P}^{\mathrm{T}}\mathbf{P}\big)^{-1} \sum_{j=1}^{n} (1 - t_j) g(y_{0j}, \mathbf{u}_j) \mathbf{p}^{k_n}(\mathbf{x}_j).$$

Even after substituting $\widehat{m}$, directly solving (A.4) remains infeasible because Π is infinite-dimensional without further structural constraints on π. To address this, we approximate Π using a sieve space Π_{l_n}, which is finite-dimensional for each n and becomes dense in Π as n grows. Specifically, let $\mathbf{q}^{l_n}(y_0, \mathbf{u})$ be an l_n-vector of basis functions analogous to $\mathbf{p}^{k_n}$. We construct

$$\Pi_{l_n} = \left\{\pi(y_0, \mathbf{u}) = \mathrm{expit}\{\mathbf{q}^{l_n}(y_0, \mathbf{u})^{\mathrm{T}} \boldsymbol{\beta}\} : \boldsymbol{\beta} \in \mathbb{R}^{l_n}\right\}.$$

We then estimate π by solving

$$\widehat{\pi}_{\mathrm{np}} = \arg\min_{\tilde{\pi} \in \Pi_{l_n}} \frac{1}{n} \sum_{i=1}^{n} \left[\widehat{m}\{\mathbf{x}_i; (1-\tilde{\pi})^{-1}\} - 1\right]^2. \tag{A.5}$$

As a result, an estimator of Δ is given by

$$\widehat{\Delta}_{\mathrm{np}} = \sum_{i=1}^{n} \left\{ t_i y_{1i} - (1 - t_i) \frac{\widehat{\pi}_{\mathrm{np}}(y_{0i}, \mathbf{u}_i)}{1 - \widehat{\pi}_{\mathrm{np}}(y_{0i}, \mathbf{u}_i)} y_{0i} \right\} \Big/ \sum_{i=1}^{n} t_i.$$

We now comment on some potential issues with the nonparametric strategy. First, nonparametric estimation is well known to suffer from the curse of dimensionality. This issue mainly arises because we approximate m and π using expanding sieves $\mathbf{p}^{k_n}$ and $\mathbf{q}^{l_n}$. In general, approximation error decreases with the smoothness of the estimand and increases

with the dimension of the arguments. Hence, when $\mathbf{x}$ is of moderate or high dimension, both approximation quality and finite-sample performance can deteriorate. Second, the identification equation (3) is a Fredholm integral equation of the first kind and is therefore ill-posed. As a result, the estimator $\widehat{\pi}_{\text{np}}$ in (A.5) may converge very slowly (Ai and Chen, 2003; Newey and Powell, 2003). In practice, it is often necessary to regularize the sieve space Π_{l_n}, e.g., by imposing boundedness restrictions on the parameters indexing Π_{l_n}, to ensure well-behaved asymptotic properties. Third, inference based on $\widehat{\Delta}_{\text{np}}$ is challenging because a convenient linear representation is difficult to obtain using standard techniques such as Taylor expansion. This difficulty stems from the ill-posed nature of $\widehat{\pi}_{\text{np}}$ so establishing asymptotic normality typically requires additional conditions and more delicate arguments. Accordingly, we omit a theoretical analysis of the nonparametric estimators $\widehat{\pi}_{\text{np}}$ and $\widehat{\Delta}_{\text{np}}$, as it is beyond the scope of this paper. We refer readers to Newey and Powell (2003); Chen and Pouzo (2009); Darolles et al. (2011) for discussions of regularization and the asymptotic behavior of ill-posed estimators, and to Ai and Chen (2007); Shan and Yan (2024); Shan et al. (2025, 2026) for asymptotic distribution theory and nonparametric inference in ill-posed settings.

References

Alberto Abadie and Guido Imbens. Simple and bias-corrected matching estimators for average treatment effects. Nber technical working paper, National Bureau of Economic Research, Cambridge, MA, 2002.

Alberto Abadie and Guido W Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006.

Alberto Abadie and Guido W Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011.

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Chunrong Ai and Xiaohong Chen. Estimation of possibly misspecified semiparametric conditional moment restriction models with different conditioning variables. *Journal of Econometrics*, 141(1):5–43, 2007.

Joshua D Angrist and Alan B Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434): 444–455, 1996.

Gustavo Viani Arruda, Marina Lourencao, Jorge H. Caldeira de Oliveira, Julia S. C. Galendi, and Alexandre A. Jacinto. Cost-effectiveness of stereotactic body radiotherapy

versus conventional radiotherapy for the treatment of surgically ineligible stage I non-small cell lung cancer in the Brazilian public health system. *The Lancet Regional Health – Americas*, 14:100329, 2022.

Michael Baiocchi, Jing Cheng, and Dylan S. Small. Instrumental variable methods for causal inference. *Statistics in Medicine*, 33(13):2297–2340, 2014.

Alexandre Belloni, Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Some new asymptotic theory for least squares series: Pointwise and uniform results. *Journal of Econometrics*, 186(2):345–366, 2015.

Peter J Bickel. On adaptive estimation. *The Annals of Statistics*, 10(3):647–671, 1982.

Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore, 1993.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

Xiaohong Chen. Large sample sieve estimation of semi-nonparametric models. In James J. Heckman and Edward E. Leamer, editors, *Handbook of Econometrics*, volume 6, pages 5549–5632. Elsevier, 2007.

Xiaohong Chen and Demian Pouzo. Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *Journal of Econometrics*, 152(1): 46–60, 2009.

Yong Chen and Becky A. Briesacher. Use of instrumental variable in prescription drug research with observational data: a systematic review. *Journal of Clinical Epidemiology*, 64:687–700, 2011.

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

Paul S Clarke and Frank Windmeijer. Instrumental variable estimators for binary outcomes. *Journal of the American Statistical Association*, 107(500):1638–1652, 2012.

William G Cochran and Donald B Rubin. Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 35(4):417–446, 1973.

S. Darolles, Y. Fan, J. P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

Rajeev H Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

Xavier d'Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471, 2011.

Max H. Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021.

Markus Frölich and Blaise Melly. Identification of treatment effects on the treated with one-sided non-compliance. *Econometric Reviews*, 32(3):384–414, 2013.

Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66(2):315–331, 1998.

Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

Erin Hartman, Richard Grieve, Roland Ramsahai, and Jasjeet S Sekhon. From sample average treatment effect to population average treatment effect on the treated: combining experimental with observational studies to estimate population treatment effects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 178(3):757–778, 2015.

James J Heckman and Jeffrey A Smith. Assessing the case for social experiments. *Journal of Economic Perspectives*, 9(2):85–110, 1995.

James J Heckman and Edward Vytlacil. Policy-relevant treatment effects. *American Economic Review*, 91(2):107–111, 2001.

James J Heckman, Hidehiko Ichimura, and Petra E Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64(4):605–654, 1997.

James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65(2):261–294, 1998.

Miguel A Hernán and James M Robins. Instruments for causal inference: an epidemiologist's dream? *Epidemiology*, 17(4):360–372, 2006.

Miguel A Hernán and James M Robins. *Causal Inference: What If*. Taylor & Francis, 2022.

Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

Yingyao Hu and Ji-Liang Shiu. Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory*, 34(3):659–693, 2018.

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

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, 86(1):4–29, 2004.

Guido W Imbens and Donald B Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press, 2015.

Guido W Imbens, Whitney K Newey, and Geert Ridder. Mean-square-error calculations for average treatment effects. Iepr working paper, Institute of Economic Policy Research, University of Southern California, 2005.

Zhezhen Jin, Zhiliang Ying, and LJ Wei. A simple resampling method by perturbing the minimand. *Biometrika*, 88(2):381–390, 2001.

Rainer Kress. *Linear Integral Equations*, volume 82 of *Applied Mathematical Sciences*. Springer New York, New York, 1989.

Michael Lechner. Earnings and employment effects of continuous off-the-job training in East Germany after unification. *Journal of Business & Economic Statistics*, 17(1):74–90, 1999.

Wei Li, Wang Miao, and Eric Tchetgen Tchetgen. Non-parametric inference about mean functionals of non-ignorable non-response data without identifying the joint distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):913–935, 2023.

Lan Liu, Wang Miao, Baoluo Sun, James Robins, and Eric Tchetgen Tchetgen. Identification and inference for marginal average treatment effect on the treated with an instrumental variable. *Statistica Sinica*, 30(3):1517–1541, 2020.

Edwin P Martens, Wiebe R Pestman, Anthonius de Boer, Svetlana V Belitser, and Olaf H Klungel. Instrumental variables: application and limitations. *Epidemiology*, 17(3):260–267, 2006.

Daniel F McCaffrey, Greg Ridgeway, and Andrew R Morral. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychological Methods*, 9(4):403–425, 2004.

Walter R Mebane and Paul Poast. Causal inference without ignorability: Identification with nonrandom assignment and missing treatment data. *Political Analysis*, 21(2):233–251, 2013.

Wang Miao, Peng Ding, and Zhi Geng. Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111 (516):1673–1683, 2016.

Wang Miao, Zhi Geng, and Eric Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Wang Miao, Lan Liu, Yilin Li, Eric J. Tchetgen Tchetgen, and Zhi Geng. Identification and semiparametric efficiency theory of nonignorable missing data with a shadow variable. *ACM / IMS Journal of Data Science*, 1(2):1–23, 2024.

Erica EM Moodie, Olli Saarela, and David A Stephens. A doubly robust weighting estimator of the average treatment effect on the treated. *Stat*, 7(1):e205, 2018.

Stephen L Morgan and Christopher Winship. *Counterfactuals and Causal Inference*. Cambridge University Press, 2nd edition, 2014.

Kosuke Morikawa and Jae Kwang Kim. Semiparametric optimal estimation with nonignorable nonresponse data. *The Annals of Statistics*, 49(5):2991–3014, 2021.

Whitney K. Newey. Convergence rates and asymptotic normality for series estimators. *Journal of Econometrics*, 79(1):147–168, 1997.

Whitney K. Newey and Daniel McFadden. Large sample estimation and hypothesis testing. In *Handbook of Econometrics*, volume 4, pages 2111–2245. Elsevier, 1994.

Whitney K Newey and James L Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003.

Joseph P Newhouse and Mark McClellan. Econometrics in outcomes research: the use of instrumental variables. *Annual Review of Public Health*, 19(1):17–34, 1998.

Jerzy Neyman. On the application of probability theory to agricultural experiments: Essay on principles. *Roczniki Nauk Rolniczych*, 10:1–51, 1923. English translation of excerpts by D. M. Dabrowska and T. P. Speed, *Statistical Science*, 5(4):465–472, 1990.

Jerzy Neyman. Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander, editor, *Probability and Statistics: The Harald Cramér Volume*, pages 213–234. Wiley, New York, NY, 1959.

Jerzy Neyman. C(α) tests and their use. *Sankhyā: The Indian Journal of Statistics, Series A*, 41(1/2):1–21, 1979.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2009.

R Pirracchio, M Carone, M Resche Rigon, E Caruana, A Mebazaa, and S Chevret. Propensity score estimators for the average treatment effect and the average treatment effect on the treated may yield very different estimates. *Statistical Methods in Medical Research*, 25(5):1938–1954, 2016.

Romain Pirracchio and Marco Carone. The balance super learner: A robust adaptation of the super learner to improve estimation of the average treatment effect in the treated based on propensity score matching. *Statistical Methods in Medical Research*, 27(8): 2504–2518, 2018.

James M Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics – Theory and Methods*, 23(8):2379–2412, 1994.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Donald B Rubin. Matching to remove bias in observational studies. *Biometrics*, 29(1): 159–183, 1973a.

Donald B Rubin. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29(1):185–203, 1973b.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Donald B Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.

Anton Schick. On asymptotically efficient estimation in semiparametric models. *The Annals of Statistics*, 14(3):1139–1151, 1986.

Anand Shah, Stephen M. Hahn, Robert L. Stetson, Joseph S. Friedberg, Taine T. V. Pechet, and David J. Sher. Cost-effectiveness of stereotactic body radiation therapy versus surgical resection for stage I non–small cell lung cancer. *Cancer*, 119(17):3123–3132, 2013.

Jiawei Shan and Xinyu Yan. Nonparametric causal inference with confounders missing not at random. *Statistica Neerlandica*, 78(4):573–583, 2024.

Jiawei Shan, Ting Wang, Wei Li, and Chunrong Ai. Nonparametric estimation of path-specific effects in the presence of nonignorable missing covariates. *Scandinavian Journal of Statistics*, 52(4):1556–1593, 2025.

Jiawei Shan, Wei Li, and Chunrong Ai. Efficient nonparametric inference for mediation analysis with nonignorable missing confounders. *Journal of the American Statistical Association*, pages 1–12, 2026.

Jun Shao and Lei Wang. Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 103(1):175–187, 2016.

Samidha Shetty, Yanyuan Ma, and Jiwei Zhao. The pursuit of efficiency versus robustness: A learning experience from analyzing a semiparametric nonignorable propensity score model. *Observational Studies*, 9(1):97–104, 2023.

Samidha Shetty, Yanyuan Ma, and Jiwei Zhao. Robust estimation under a semiparametric propensity model for nonignorable missing data. *Electronic Journal of Statistics*, 19(1): 956–981, 2025.

Eric J. Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal inference. *Statistical Science*, 39(3):375–390, 2024.

Qinglong Tian, Donglin Zeng, and Jiwei Zhao. Identification and efficient estimation in regression analysis with response missing not at random. *Statistica Sinica*, 2026. In press.

Anastasios A Tsiatis. *Semiparametric Theory and Missing Data*. New York: Springer, 2006.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Aolin Wang, Roch A Nianogo, and Onyebuchi A Arah. G-computation of average treatment effects on the treated and the untreated. *BMC Medical Research Methodology*, 17:3, 2017.

Linbo Wang and Eric Tchetgen Tchetgen. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(3):531–550, 2018.

Sheng Wang, Jun Shao, and Jae Kwang Kim. An instrumental variable approach for identification and estimation with nonignorable nonresponse. *Statistica Sinica*, 24(3): 1097–1116, 2014.

Philip G Wright. *Tariff on animal and vegetable oils*. Macmillan Company, New York, 1928.

Jiwei Zhao and Yanyuan Ma. A versatile estimation procedure without estimating the nonignorable missingness mechanism. *Journal of the American Statistical Association*, 117(540):1916–1930, 2022.

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

Qingyuan Zhao and Daniel Percival. Entropy balancing is doubly robust. *Journal of Causal Inference*, 5(1):20160010, 2017.