Efficient and Doubly Robust Estimation of Average Treatment Effects with Confounders Missing Not at Random

Dingyi Li*

November 1, 2025

Link for the latest version

Abstract

Causal inference in observational studies relies on the conditional independence of outcomes and treatment given confounding covariates. When confounders are missing not at random (MNAR), this condition fails, and causal parameters are no longer identified. We consider scenarios where missingness of the confounder depend on the confounder but is conditionally independent of the outcome given treatment and confounders (OIM). We propose an efficient and doubly robust estimator of the average treatment effect (ATE); it is based on well-known sample averages of observed outcomes and estimated conditional mean outcomes but with novel propensities and weights that adjust for confounder missingness. To estimate these weights we invert integral equations that relate observed distributions to: (i) the joint propensity score, defined as the probability of receiving treatment and having all confounders observed, and (ii) the distribution of confounders conditional on missingness, that are *unobserved* due to missingness. The inversion relies on OIM and a completeness of the full-data distribution in the outcome. To extend the analysis to prevalent empirical settings with insufficient variation in the outcome, e.g., binary outcome or multiple continuous confounders with missing values, we propose a low-rank assumption on the missingness mechanism that regularizes an ill-posed integral equation and leads to efficiency gains when the inversion is well-posed. Furthermore, we derive the semiparametric efficiency bound for the ATE in OIM setting, show that our estimator achieves the bound, and enjoys novel and standard robustness properties of double machine learning estimators. We benchmark our estimator with simulations and three empirical applications: the impact of the Job Corps program on employment, the effect of smoking on blood lead levels, and the influence of education on general health satisfaction.

Keywords: Outcome-independent missingness; nonparametric identification; integral equation; ill-posed problem; doubly robust estimation; semiparametric inference

*Dyson School of Applied Economics and Management, Cornell University. Email: dl922@cornell.edu.
I am especially grateful to my advisors José Luis Montiel Olea, Peng Ding, Francesca Molinari for their generous support and advice. I also thank Brian Dillon, Kirill Brosuyak, Jacob Dorn, Junlong Feng, Mengsi Gao, Lihua Lei, Cynthia Lin Lawell, Yaroslov Mukin, Peizan Sheng, Jörg Stoye, Amilcar Velez, and seminar participants for their valuable comments and suggestions. All errors are my own.

1 Introduction

Estimating the average treatment effect, denoted ATE or τ, in observational studies crucially relies on the unconfoundedness assumption that potential outcomes and treatment are conditionally independent given confounding covariates (Rosenbaum and Rubin, 1984; Imbens and Rubin, 2015). When observations of confounders are missing in a way that depends on the values of outcomes or covariates, referred to as missing not at random (MNAR), unconfoundedness fails. Consequently, the mean conditional outcomes and propensity score cannot be identified, and standard inferences based on complete observations are inconsistent due to selection bias. This issue is prevalent in practice. For instance, in a survey, low education and high income values are commonly undisclosed by respondents in order to protect their privacy (Bollinger et al., 2019).

Existing approaches to missing confounders typically rely on strong assumptions on the missingness mechanism. If data are missing completely at random (MCAR), meaning the event of missingness is independent of every variable in the study, including those with missing values, then complete-case analysis is valid, albeit inefficient. More often, analysts make the missing at random (MAR) assumption. This holds when the event of missingness depends only on completely observed variables and not on variables with missing values. Under MAR the full-data distribution is identified and multiple imputation or inverse-probability weighting methods can be used. The MAR assumption requires the event that a confounder value is missing to be conditionally independent of that confounder given treatment, outcome and completely observed confounders. However, if the missingness mechanism is not independent of the missing values, e.g., withholding education when it is low or income when it is high, these methods fail and yield inconsistent estimates. Accounting for missing confounders in the MNAR setting requires qualitatively different techniques and motivates our work.

We consider estimation of the ATE in scenarios where the event of confounder missingness may depend on the confounder value but is independent of the outcome (OIM). More precisely, following the framework of Yang et al. (2019), we assume that the missingness mechanism is conditionally independent of the outcome given treatment and confounders. This assumption can be understood by analogy with the unconfoundedness of treatment: under this assumption, if the potential outcomes are also conditionally independent of both treatment and missingness given the confounders, then the conditional average treatment effect (CATE) is identified via conditional mean outcomes in the treated and control strata defined by the fully observed confounders. What is not immediate is how one aggregates these conditional effects into the unconditional ATE, since the marginal distribution of

the confounders is not observed due to missingness. Furthermore, we investigate the most statistically efficient way to achieve this.

We propose an efficient and doubly robust estimator of the ATE. Our estimator is based on two familiar weighting strategies: (i) weighting the observed outcomes by the reciprocal of the probability of confounder missingness and treatment; (ii) weighting the conditional mean treated/control outcomes by the marginal distribution of confounders conditional on missingness. The econometric challenge here is that both nuisance weight functions are not directly observed in the data due to missing confounder values. Indeed, we require the conditional probability of treatment and missingness given confounders, and the probability density of confounders conditional on missingness, which are not fully observed. Instead, each weight function is identified via an integral equation that relates a observed marginal outcome density to a *mixture* of an unobserved weight function (solved for) and a observed conditional distribution of the outcome given confounder and missingness.

Integral equations are continuous/infinite analogues of finite matrix equations; when full-data distribution is discrete this analogy is exact, while many intuitions remains valid in the continuous case. Recall that inverting a linear system requires the matrix column rank be greater than the number of equations. The counterpart condition for integral equations is completeness of the integrand signature function. In our OIM setting, completeness requires that there is quantitatively more variation in the outcome than in the confounders with missing values; in the discrete case, this is full column rank of a contingency table. When completeness holds, conditional distribution of confounders given the missingness event and conditional probability of missingness given confounders are identified.

The completeness assumption is too restrictive for common empirical settings, e.g., when outcome is binary and confounder is not, or when there are multiple continuous confounders with missing observations. We propose a low-rank assumption on the missingness mechanism to accommodate OIM analysis in these empirical settings. The low-rank assumption shifts the completeness condition on the variation of the confounder, which is typically not restrictive in practice.

With the key weight functions identified, we follow well-known strategies for identifying the ATE. We construct a doubly robust estimators by inverse joint propensity weighting (IJPW) and averaging the estimated CATE obtained from the complete-data. We derive the semiparametric efficiency bound for the ATE by characterising the tangent space. We then show that the influence function of our estimator lies within that space, which ensures that our estimator achieves the smallest asymptotic variance among regular estimators. By construction, our estimators remain consistent as long as at least one of the two identification weighting strategies is consistently implemented.

We consider two estimation approaches: (i) plug-in of nuisance functions estimated with full data, (ii) double machine learning (DML) with sample-splitting. Our plug-in estimator is consistent if the estimated nuisance functions are consistent in quadratic mean and does not require Donsker-class conditions. In empirical settings where completeness of outcome distribution is plausible, our low-rank assumption can be used to construct a doubly robust estimator that is consistent as long as either the completeness or the rank assumptions hold. Solving integral equations specified with *estimated* input functions is an ill-posed inverse problem (Newey and Powell, 2003), meaning that small estimation errors in the inputs lead to discontinuous changes in the solution. We propose ridge series estimators for nuisance functions estimated by inverting integral equations to regularize the ill-conditioned inversion.

We benchmark our estimators with synthetic designs and three real datasets. Our toy design includes a single confounder subject to missingness generated by a nonlinear mechanism; our realistic design employs six confounders and an aggressive missingness pattern with 49% of incomplete data. In simulations, our efficient estimators exhibit negligible bias, satisfactory coverage, and steadily decreasing variance, by contrast the nonparametric alternative (Yang et al., 2019) has poorer confidence-interval coverage. The DML estimator consistently outperforms the plug-in. We then illustrate our methodology in three empirical applications: the causal effect of a job training program on employment (binary outcome), the effect of smoking on blood-lead levels, and the effect of education on general health satisfaction. In all applications, our estimators yield point estimates of greater magnitude and tighter confidence intervals compared with the nonparametric alternative (Yang et al., 2019); we conclude that accounting for semiparametric structure of the problem meaningfully improves both finite-sample precision and asymptotic power.

We make three contributions to the literature. First, we extend the OIM framework (Bartlett et al., 2014; Miao et al., 2018; Yang et al., 2019; Miao et al., 2023; Lu and Ashmead, 2018; Zuo et al., 2024) to economic applications, where missing data are prevalent and often exhibit selection patterns while variation in outcomes is typically low and the completeness condition does not hold. For example, in estimating the ATE of job-training programs on employment, the LaLonde dataset (LaLonde, 1986; Dehejia and Wahba, 1999; Imai and Ratkovic, 2014; Armstrong and Kolesár, 2021; Breunig et al., 2025) contains roughly 10% of observations with missing baseline income that do not arise from unemployment status, while the dataset analyzed by Lee (2009) exhibits over 30% missingness in baseline income and parental education. Standard approaches for missing confounders often impose MAR or MCAR assumptions for tractability (Rubin, 1976; Agarwal and Singh, 2021), yet these assumptions are frequently violated in applied work due to nonresponse or

administrative censoring that depends on these unobserved values (Bollinger et al., 2019). The OIM assumption allows missingness to depend on the missing confounder, but rules out mechanisms that depend on the outcome (Heckman, 1979; Newey et al., 1990; Lee, 2009; Honoré and Hu, 2024), and is plausible in empirical settings where outcomes are realized after confounders are collected. We propose a new identification strategy for OIM scenarios with limited variation in outcomes, including binary outcomes, which are prevalent in economics and cannot be studied with any existing methods. Whereas existing methods recover the distribution of missing confounders from outcomes, we instead recover the missingness mechanism from completely-observed confounders. Both approaches achieve point identification of the ATE without relying on instruments or bounding strategies, providing tractable alternatives to partial-identification methods (Horowitz and Manski, 2000; Molinari, 2010).

Second, we contribute to the OIM literature a characterization of the semiparametric efficiency bound (Robinson, 1988; Chamberlain, 1992; Bickel et al., 1993; Newey, 1994; Ai and Chen, 2003; Hirano et al., 2003; Tsiatis, 2006; Hahn and Ridder, 2013; Chernozhukov et al., 2018; Hirshberg and Wager, 2021; Chen et al., 2025; Borusyak and Hull, 2025) of the ATE functional. (Hahn, 1998). Under OIM the tangent space is not the entire Hilbert space (Kennedy, 2016). This restriction arises because OIM models impose testable constraints on the observed-data distribution (Sjölander and Hägg, 2025). Causal parameters typically admit both a regression-based and a propensity-based representation, each contributing to the semiparametric efficiency bound. Our estimator operationalizes two separate integral equations in order to leverage full statistical power of the data via both the regression and propensity score nuisance parameters.

Third, we bridge the OIM literature with the DML literature. We extend the classical doubly robust estimator of the ATE to the OIM setting. Compared to the classical estimator, ours requires additional steps: (a) solving integral equations for unobserved propensities and densities, (b) modifying the classical doubly-robust estimator to account for missing data using the weights from step (a). Working within the classical doubly robust estimator allows practitioners to leverage knowledge they already have. Furthermore, in addition to the standard robustness properties, the doubly robust estimator in our setting enables robustness to violations of the completeness assumption of the OIM setting whenever both can be used. Following the DML literature, we enable flexible high-dimensional estimation of nuisance functions and solution of the integral equation via machine learning (Chen and Newey, 2020; Newey and Powell, 2003; Singh and Zhang, 2019; Xu and Zhu, 2020; Chen and Zhan, 2023; Fonseca and Xu, 2024). Our approach builds on ideas from nonparametric instrumental variables and completeness-based identification (Newey and

Powell, 2003; Hu, 2008; Hu and Schennach, 2008; D'Haultfœuille, 2011; Chesher, 2003; Hall and Horowitz, 2005; Chen et al., 2011; Berry and Haile, 2014; Schennach, 2016; Hu and Shiu, 2018; D'Haultfœuille and Février, 2015; Berry and Haile, 2024), linking them to recent developments in doubly robust estimation under novel data structures (Sant'Anna and Zhao, 2020b; Arkhangelsky et al., 2021; Ji et al., 2023; Cui et al., 2024; Sun et al., 2025).

The remainder of the paper is structured as follows. In section 2, we outline our theoretical framework and basic assumptions. In Section 3, we focus on identification where we propose identification strategies based on solving two novel integral equations derived from the observed data. In Section 4 and 5, we introduce a doubly robust estimator, and establish its theoretical guarantees (e.g., $\sqrt{n}$-consistency, asymptotic linearity, and local semiparametric efficiency). Sections 6 and 7 present numerical simulations and an empirical application, respectively, to validate the estimator's finite-sample performance and practical use. Proofs, technical derivations, and additional classical identification strategies are provided in the Supplementary material.

2 Setup

2.1 Notation

Let W denote a generic random variable and $(W_1, \ldots, W_n)$ denote an observed independent and identically distributed sample from the distribution of W. Convergence in distribution is denoted by $\rightsquigarrow$, and convergence in probability is indicated as $\xrightarrow{P}$. We say $W_n = O_P(r_n)$ if W_n/r_n is bounded in probability, and $W_n = o_P(r_n)$ if $W_n/r_n \xrightarrow{P} 0$. The empirical measure is denoted by $\mathbb{P}_n$, and for any function θ, the sample average is written as $\mathbb{P}_n(\theta) \coloneqq \mathbb{P}_n\{\theta(W)\} = \frac{1}{n}\sum_{i=1}^n \theta(W_i)$. Similarly, for a (possibly random) functional $\hat{\theta}$, we define $\mathbb{P}(\hat{\theta}) \coloneqq \int \hat{\theta}(w) dP(w)$, and the squared $L^2(P)$ norm is $\|\hat{\theta}\|_2^2 = \int \hat{\theta}(w)^2 dP(w)$. Here, P in the integral denotes the true probability measure of W. Define the indicator function for the event $W = w_0$ as $\mathbb{I}\{W = w_0\}$, which equals 1 if $W = w_0$ and 0 otherwise. Let $[n] \coloneqq \{1, \cdots, n\}$.

2.2 Framework and assumptions

We adopt a potential outcomes framework. Let $A \in \{0, 1\}$ denote binary treatment (0=control, 1=treatment). For treatment level a, let $Y(a)$ be the potential outcome under treatment a. The observed outcome is $Y = AY(1) + (1 - A)Y(0)$. Let $X = (X_1, \ldots, X_p)$ be a vector of p-dimensional pre-treatment confounders, with an independent and identically

distributed sample of size n from $\{A, X, Y(0), Y(1)\}$. We assume that $Y(1)$ and $Y(0)$ both have finite variance. Define the conditional average treatment effect (CATE) as $\tau(X) \coloneqq \mathbb{E}[Y(1) - Y(0)|X]$, the average treatment effect (ATE) as $\tau \coloneqq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\tau(X)]$, and the conditional mean potential outcome $\mu_a(X) = \mathbb{E}[Y(a) \mid X]$ for $a = 0, 1$.[1] The following are standard assumptions in causal inference with observational studies (Rosenbaum and Rubin, 1983).

Assumption 1. *(Unconfoundedness)* $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid X$.

Assumption 2. *(Overlap) There exist constants* c_1 *and* c_2 *such that* $0 < c_1 \leq e(X) \leq c_2 < 1$ *almost surely, where* $e(X) \coloneqq P(A = 1 \mid X)$ *is the propensity score.*

Under these assumptions, the average treatment effect $\tau = \mathbb{E}[\mathbb{E}(Y|A = 1, X) - \mathbb{E}(Y|A = 0, X)]$ becomes well-defined and identifiable from the observed data distribution. Rosenbaum and Rubin (1983) establish that $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid e(X)$, demonstrating that propensity score adjustment suffices for confounding removal for an ideal case. The effect τ can then be estimated by propensity score matching, subclassification, or weighting methods.

We focus on a setup in which confounders X contain missing values, precluding direct identification of the propensity score. Assume we have n independent and identically distributed draws from $\{A, X, Y(1), Y(0), R\}$, where $R = (R_1, \ldots, R_p)$ is the random missingness indicator vector and $R_j = 1$ if X_j is observed and 0 otherwise. Let $\mathcal{R}$ denote all possible missingness patterns. Define 1_p and 0_p as the p-vectors of all ones and zeros, respectively. Following Rubin (1976), confounders partition into observed X_r and missing $X_{\overline{r}}$ components for patterns $R = r \in \mathcal{R}$. For example, when $R_1 = 1$ and $R_j = 0$ for $j \geq 2$, then $X_R = X_1$ and $X_{\overline{R}} = (X_2, \ldots, X_p)$. We make the following assumption about the missing data patterns formalized by Yang et al. (2019).

Assumption 3. *(Outcome-independent missingness)* $\{Y(0), Y(1)\} \perp\!\!\!\perp R \mid (A, X)$.

We summarize other missing mechanisms in the literature and their limitations into the Supplementary material S. Figure 1 encodes our framework for Assumption 1 and 3 by causal diagrams (Pearl, 1995):

(a) A and Y share no common causes other than X.

(b) R and Y share no common causes other than A and X.

[1] The parameter of interest in this paper is τ but one can extend our results to average treatment effect on the treated.

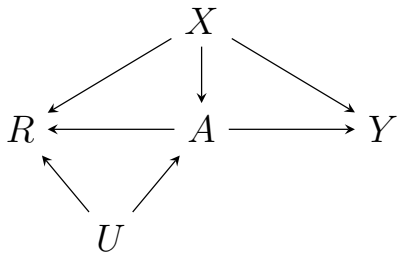

Figure 1: A simple causal diagram consistent with Assumption 1 and 3 (unconfoundedness and outcome-independent missingness). The The treatment variable A, outcome Y, and missing indicator R are observed. The variable U is unobserved, and the confounders X contain missing values.

Our framework is built to accommodate the challenges in missing data studies. Specifically, we permit the possibility that the missingness indicator R and the treatment A share unmeasured common causes U as in Figure 1. That is, the process that leads data to be missing may itself be influenced by factors which also affect treatment assignment, and these factors may not all lie in the observed covariate set. We also allow the missingness mechanism to depend on confounders that themselves may be missing ($X_{\bar{R}}$), meaning that whether an observation is missing can depend on covariate values that are unobserved in that unit. Moreover, we accommodate the situation where missingness occurs after treatment, so that R may depend on A rather than being restricted to pre-treatment covariates alone.

For example, in a study of a job training program where individuals self select into training and we measure their earnings one year later, suppose we observe baseline covariates such as age and education but one key confounder, baseline income, is missing for some participants. The missingness of baseline income is itself tied to the income level. At the same time, baseline income influences the likelihood of enrolling in the training program. Thus the missingness indicator and the treatment decision share a common cause (baseline income). The unobserved factor U might represent the quality and efficiency of the local job training agency's administration, which affects both the likelihood that a participant enrolls in the training program and the likelihood that their baseline income is accurately recorded.

Throughout identification and estimation I rely on Assumption 3. One potential concern is that in the illustrative example the randomness pertains only to a single variable (the missingness of baseline income depends on the income itself). Under what conditions does conditioning on all confounders guarantee independence between the missingness mecha-

nism and the outcome? The following remark addresses those conditions.

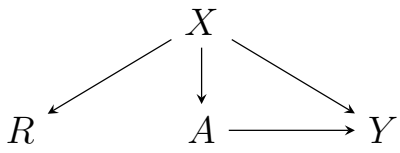

Figure 2: A simple causal diagram consistent with Remark 1 (sufficient condition for outcome-independent missingness). The treatment variable A, outcome Y, and missing indicator R are observed. The confounders X contain missing values.

Remark 1. *A useful sufficient condition for Assumption 3 (outcome-independent missingness) is the following. If*

$$(A, Y) \perp\!\!\!\perp R \mid X$$

hold, Assumption 3 is satisfied. This is a direct result of the contraction variant (Dawid, 1979). In other words, our Assumption 3 also covers the scenario in which missingness R is independent of Y once we condition on X, provided that R is independent with (A, Y) conditional on X. Intuitively, the interpretation is that R is influenced only by X and not by either A or Y.

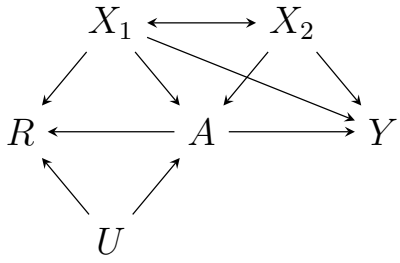

Figure 3: A simple causal diagram consistent with Remark 2 (sufficient condition for outcome-independent missingness). The treatment variable A, outcome Y, and missing indicator R are observed. The variable U is unobserved, and the confounders X_1 contain missing values.

Remark 2. *There is also another similar useful sufficient condition by the contraction variant (Dawid, 1979) for our setting. We write $X^\top = (X_1^\top, X_2^\top)$. Under Assumption 1, if*

$$(X_2, Y) \perp\!\!\!\perp R \mid (A, X_1)$$

hold, Assumption 3 is satisfied. In other words, the framework allows for the missing-data mechanism depends on (A, X_1, X_2) yet only requires that the independence condition hold with respect to (A, X_1) and X_2 carries no additional information about the missingness mechanism R.

3 Identification

In this section, we first present a toy example in which both X and Y are continuous and X is one-dimensional, and we outline our identification strategy in subsection 3.1 without rigorous proof. Then, in subsection 3.2, we extend and formalize our arguments to cover the case where X and Y may be discrete and where X and R are multidimensional by a generic measure.

3.1 Identification under a simple scenario

In this motivating subsection, we use $(\star)$ to highlight a key equation. We assume X is one-dimensional, and that both X and Y are continuous with support $\mathcal{X}$ and $\mathcal{Y}$. We work under the following standard setup: let (A, X, Y, R) admit a joint density f, and suppose we observe only

$$f\Big(A = a,\ X = x,\ Y = y,\ R = 1\Big) \quad \text{and} \quad f\Big(A = a,\ Y = y,\ R = 0\Big).$$

Denote the observed data by $\mathcal{O} = \{A,\ X_R,\ Y,\ R\}$. We build upon the unconfoundedness and OIM assumptions (Assumptions 1 and 3), as illustrated in Figure 1.

Causal identification in Yang et al. (2019) proceeds by exploiting a conditional independence structure to derive an integral equation that recovers the full-data density $f(A = a,\ X = x,\ Y = y)$. For simplicity, we assume that all relevant conditional densities are strictly positive; for example, $f\Big(R = 1 \ \Big|\ A = a,\ X = x,\ Y = y\Big) > 0$. The standard methodology relies on solving for a unknown function $g(x)$ via an integral equation of the form

$$\int g(x) f(x, y) \mathrm{d}x = h(y).$$

from known $f(x, y)$ and $h(y)$.

In order to solve this integral equation, one need to assume that the dimension of the support of X exceeds that of the support of Y; i.e.$|\mathcal{X}| > |\mathcal{Y}|$. The formal condition guaranteeing a unique solution to the integral equation when the support is discrete (or

mixed) is given in Definition 1. We solve three integral equations below which work as the first step for our possible second step identification for the ATE. The first one is from Yang et al. (2019). The second and third equations are new contributions. The third in particular incorporates a low-rank condition on the missing-data pattern to make $|\mathcal{X}| > |\mathcal{Y}|$ possible.

(a) Integral equations. The integral equations serve to recover the latent distributions and act as intermediate steps in identifying the ATE.

(a.1) Integral equation 1. (Yang et al., 2019) The first integral equation is solving

$$(\star) \quad f(A=a, Y=y, R=0) = \int \frac{P(R=0 \mid A=a, X=x)}{P(R=1 \mid A=a, X=x)} f(A=a, X=x, Y=y, R=1)dx$$

In this equation we link the observed joint density $f(A=a, Y=y, R=0)$ with the complete-case joint density $f(A=a, X=x, Y=y, R=1)$ via a weight that depends only on x, for fixed $a \in \{0,1\}$. Solving this equation needs $|\mathcal{X}| \leq |\mathcal{Y}|$. The equation holds since

$$\begin{aligned} f(A=a, Y=y, R=0) &= \int f(A=a, X=x, Y=y, R=0)dx \\ &= \int \frac{f(A=a, X=x, Y=y, R=0)}{f(A=a, X=x, Y=y, R=1)} f(A=a, X=x, Y=y, R=1)dx \\ &= \int \frac{P(R=0 \mid A=a, X=x, Y=y)}{P(R=1 \mid A=a, X=x, Y=y)} f(A=a, X=x, Y=y, R=1)dx \\ &= \int \frac{P(R=0 \mid A=a, X=x)}{P(R=1 \mid A=a, X=x)} f(A=a, X=x, Y=y, R=1)dx. \end{aligned}$$

The first equality reflects the computation of a marginal density. The second equality arises by dividing the numerator and the denominator by the same number. The third equality similarly follows by dividing the numerator and the denominator by $f(A=a, X=x, Y=y)$. The fourth equality holds under the OIM assumption. The formal result and discussion are under Lemma S1.

(a.2) Integral equation 2. The second integral equation is solving

$$(\star) \quad f(Y=y \mid A=a, R=0) = \int f(X=x \mid A=a, R=0) f(Y=y \mid A=a, X=x, R=1)dx$$

which can be viewed as a Bayes-rule variation of **integral equation 1**. Solving this

equation needs $|\mathcal{X}| \leq |\mathcal{Y}|$. It links the conditional distributions under $R = 0$ and $R = 1$ through a weighting function that depends only on x. The equation holds since

$$\begin{aligned} f(Y = y \mid A = a, R = 0) &= \int f(X = x, Y = y \mid A = a, R = 0)dx \\ &= \int f(X = x \mid A = a, R = 0) f(Y = y \mid A = a, X = x, R = 0)dx \\ &= \int f(X = x \mid A = a, R = 0) f(Y = y \mid A = a, X = x, R = 1)dx \end{aligned}$$

The first equality reflects the computation of a marginal density over x. The second equality follows from the definition of conditional density. The third equality holds under the OIM assumption. The formal result and discussion are under Lemma 1.

(a.3) Integral equation 3. We have the third integral equation using a low rank condition,

$$\begin{aligned} (\star) \quad & f(A = a, Y = y) \\ & = \int \phi_a(\widetilde{y}) \left[\int f(\widetilde{Y} = \widetilde{y} \mid A = a, X = x, R = 1) f(A = a, X = x, Y = y, R = 1)dx \right] d\widetilde{y}. \end{aligned}$$

where $f(A = a, Y = y)$ is the target full-data joint distribution. The inner integral over x integrates out the confounders based on the observed distributions, while the outer integral over $\widetilde{y}$ combines information across outcome values. In Theorem 3, we show that $\phi_a(y)$ is identified by the key integral equation $(\star)$ if $f(Y = y \mid A = a, X = x, R = 1)$ is complete in X. Solving this equation needs $|\mathcal{X}| \geq |\mathcal{Y}|$. The **integral equation 3** relies on an assumption of a low-rank structure,

$$\int \phi_a(y) f(Y = y \mid A = a, X = x, R = 1)\, dy = \frac{1}{P(R = 1 \mid A = a, X = x)}.$$

so the equation $(\star)$ holds since

$$\begin{aligned} f(A = a, Y = y) &= \int \frac{1}{P(R = 1 \mid A = a, X = x, Y = y)} f(A = a, X = x, Y = y, R = 1)dx \\ &= \int \frac{1}{P(R = 1 \mid A = a, X = x)} f(A = a, X = x, Y = y, R = 1)dx \\ &= \int \left[\int \phi_a(\tilde{y}) f(\widetilde{Y} \mid A = a, X = x, R = 1_p) d\tilde{y} \right] f(A = a, X = x, Y = \tilde{y}, R = 1_p)dx \\ &= \int \phi_a(\widetilde{y}) \left[\int f(\widetilde{Y} = \widetilde{y} \mid A = a, X = x, R = 1) f(A = a, X = x, Y = y, R = 1)dx \right] d\widetilde{y}. \end{aligned}$$

The first equality reflects the computation of a marginal density over x. The second equality follows from the OIM assumption. The third equality relies on the low-rank condition, and the fourth equality applies Fubini's theorem. For illustration, we assume a binary outcome, $y = \{0, 1\}$. Then the low rank condition becomes

$$\sum_{y=0}^{1} \phi_a(y) f(Y = y \mid A = a, X = x, R = 1) = \frac{1}{P(R = 1 \mid A = a, X = x)}.$$

which is equivalent to

$$\sum_{y=0}^{1} \phi_a(y) f(Y = y, A = a, X = x \mid R = 1) = 1$$

by multiplying both sides of the equation by $P(R = 1 \mid A = a, X = x)$. With this equivalence, a sufficient condition for this low rank condition can be easily shown as

$$c_1 f(Y = 1, A = a, X = x \mid R = 1) + c_2 = f(Y = 0, A = a, X = x \mid R = 1)$$

for some constant c_1 and c_2 where we choose $\phi_a(1) = -c_1\phi_a(0)$ and $\phi_a(0) = \frac{1}{c_2}$. More intuitively, it is also easy to show that one specific example for this sufficient condition is

$$f(Y = 1,\, X = x \mid A = a, R = 1) = \tfrac{1}{2}\, e^{-|x|}, \quad f(Y = 0,\, X = x \mid A = a, R = 1) = 0.3 + e^{-|x|}.$$

for some L_1 and L_2 such that $x \in [L_1, L_2]$ and the joint densities above integrate to one. The formal result and discussion are centered around Assumption 6.

(b) Final identifications. The final identification results serve to connect the latent distributions to the ATE by two familiar nuisance weighting strategies: (i) weighting the observed outcomes by the reciprocal of the probability of confounder missingness and treatment; (ii) weighting the conditional mean treated/control outcomes by the marginal distribution of confounders conditional on missingness. Our final identification differs from that of Yang et al. (2019) and can be readily extended to a sample analog, thereby enabling the construction of a doubly robust estimator.

(b.1) Final identification 1. Under Assumptions 1 and 3, we can identify the ATE τ by

$$\tau = \mathbb{E}\left[\frac{A\mathbb{I}\{R = 1\}Y}{e_1(X)} - \frac{(1 - A)\mathbb{I}\{R = 1\}Y}{e_0(X)}\right],$$

where

$$(\star) \quad e_a(X) \coloneqq P(A = a, R = 1 \mid X)$$

for $a = 0, 1$. In this final identification, we assume that the nuisance $e_a(X)$ is known. Different with classical inverse propensity score weighting (Rosenbaum and Rubin, 1983), we adjust the weight to account for missingness. See the proof of Theorem 1. With this identification, we are able to use sample analog to estimate ATE,

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{A_i \mathbb{I}\{R_i = 1\} Y_i}{e_1(X_i)} - \frac{(1 - A_i) \mathbb{I}\{R_i = 1\} Y_i}{e_0(X_i)} \right].$$

(b.2) Final identification 2. Under Assumptions 1 and 3, by the law of iterated expectation, the ATE τ is identified by

$$\tau = \mathbb{E}[\tau(A, X_R, R)],$$

where

$$\begin{aligned} (\star) \quad \tau(A, X_R, R) :=& \mathbb{E}[\mathbb{E}(Y|A = 1, X, R = 1) - \mathbb{E}(Y|A = 0, X, R = 1) \mid A, X_R, R] \\ =& \mathbb{E}[\mathbb{E}(Y(1) - Y(0)|X) \mid A, X_R, R] \end{aligned}$$

In this final identification, we assume that the nuisance $\tau(A, X_R, R)$ is known. With this identification, we are able to use sample analog to estimate ATE,

$$\begin{aligned} \hat{\tau} =& \sum_{i=1}^{n} [\tau(A_i, X_{R_i}, R_i)] \\ =& \frac{1}{n_1} \sum_{i:R_i=1} \tau(A_i, X_i, R_i = 1) \; + \; \frac{1}{n_0} \sum_{i:R_i=0} \tau(A_i, R_i = 0), \end{aligned}$$

where

$$n_1 = \sum_{i=1}^{n} \mathbb{I}\{R_i = 1\} \quad \text{and} \quad n_0 = \sum_{i=1}^{n} \mathbb{I}\{R_i = 0\}.$$

(c) Identification of nuisances. Finally, we connect the two final identifications to the three integral equations. We come up with three identification methods.

(c.1) Identification 1. Let the nuisance $e_a(X)$ be identified from **integral equation 1**. Therefore, solving this equation needs $|\mathcal{X}| \leq |\mathcal{Y}|$. Specifically, notice that since we are able

to solve $\frac{P(R=0|A=a,X=x)}{P(R=1|A=a,X=x)}$ by the integral equation, we can identify,

$$P(R=1 \mid A=a, X=x) = \frac{1}{1+\frac{P(R=0|A=a,X=x)}{P(R=1|A=a,X=x)}}$$

because the probabilities of all possible values sum to one. Therefore, nuisance $e_a(X)$ is identified by

$$e_a(X) = P(A=a, R=1 \mid X) = \frac{f(A=a, X, R=1)}{\sum_{a'=0}^{1} \frac{f(A=a',X,R=1)}{P(R=1|A=a',X)}}.$$

In the end, we use the identified nuisance function $e_a(X)$ to identify the ATE through **final identification 1**, treating $e_a(X)$ as known.

(c.2) Identification 2. Let the nuisance $\tau(A, X_R, R)$ be identified from **integral equation 2**. Therefore, solving this equation needs $|\mathcal{X}| \leq |\mathcal{Y}|$. Specifically, notice that because we are able to identify $f(X=x \mid A=a, R=0)$ by the integral equation, we can further identify $\tau(A, X_R, R=0) = \tau(A=a, R=0)$ by

$$\begin{aligned}
&\mathbb{E}[\mathbb{E}(Y|A=1,X,R=1) - \mathbb{E}(Y|A=0,X,R=1) \mid A=a, R=0] \\
&= \int [\mathbb{E}(Y|A=1,X=x,R=1) - \mathbb{E}(Y|A=0,X=x,R=1)]\, f(X=x|A=a,R=0)dx.
\end{aligned}$$

With $\tau(A, X_R, R=1) = \tau(A=a, X, R=1) = \mathbb{E}(Y|A=1, X=x, R=1) - \mathbb{E}(Y|A=0, X=x, R=1)$, we are able to identify the $\tau(A, X_R, R)$. In the end, we use the identified nuisance function $\tau(A, X_R, R)$ to identify the ATE through **final identification 2**, treating $\tau(A, X_R, R)$ as known.

(c.3) Identification 3. The nuisance $e_a(X)$ is identified from **integral equation 3**. Let The nuisance $e_a(X)$ be identified from **integral equation 3**. Therefore, solving this equation needs $|\mathcal{X}| \geq |\mathcal{Y}|$. Notice that since we are able to solve $\phi_a(y)$ by integral, we can identify,

$$P(R=1 \mid A=a, X=x) = \frac{1}{\int \phi_a(y) f(Y=y \mid A=a, X=x, R=1)dy}$$

Therefore, nuisance $e_a(X)$ is identified by

$$e_a(X) = P(A=a, R=1 \mid X) = \frac{f(A=a, X, R=1)}{\sum_{a'=0}^{1} \frac{f(A=a',X,R=1)}{P(R=1|A=a',X)}}.$$

In the end, we use the identified nuisance function $e_a(X)$ to identify the ATE through **final identification 1**, treating $e_a(X)$ as known.

(d) Discussion. In general, one can combine the identification results (3.1) and (3.2), or (3.2) and (3.3), to form a doubly-robust estimator.[2] Because in most cases $|\mathcal{X}| \geq |\mathcal{Y}|$ is rare, we focus on the latter combination, which accommodates both dimension-regimes. A detailed discussion of the doubly-robust estimator appears in Section 4.

3.2 Generalized identification

In this subsection, we first revisit formally the identification approach of Yang et al. (2019) in subsection 3.2.1, which underpins our alternative identification formulas and motivates our doubly-robust estimator. We propose three new identification strategies. The first two adhere to the standard completeness regime (i.e., they require a dimension-type condition that the outcome Y is "richer" than the confounders X) and achieve identification of the average treatment effect τ via: (i) weighting the observed outcomes by the reciprocal of the probability of confounder missingness and treatment; (ii) weighting the conditional mean treated/control outcomes by the marginal distribution of confounders conditional on missingness. We refer to these as the first set of identification formulas. The third strategy operates in the converse regime, when X has greater "dimension" than Y, and leads to an integral equation of τ through weighting of the missingness and treatment mechanisms. We refer to this as the second set of identification formulas.[3]

3.2.1 Identification method in the literature

We work under the standard setup: let (A, X, Y, R) admit a joint density f (w.r.t. an appropriate dominating measure ν, which covers both discrete and continuous cases), and suppose we observe only $f(A, X_R, Y, R)$; denote the observed data by $\mathcal{O} = \{A, X_R, Y, R\}$. We build upon the unconfoundedness and OIM assumptions (Assumptions 1 and 3), as illustrated in Figure 1. Causal identification of the model proceeds by using the conditional independence structure to construct an integral equation that recovers the full data density $f(A, X, Y)$. To derive the integral equation, the distribution requires the following additional assumption:

Assumption 4. *(Non-degenerate missingness) There exists some constant c such that $P(R = 1_p \mid A, X, Y) \geq c > 0$ almost surely.*

[2] The low-rank condition for $f(X = x \mid A = a, R = 0)$ is left for future work.

[3] We leave the low-rank condition for the distribution of X conditional on missingness as future work; we nonetheless refer to the resulting expression as the alternative identification formulas for consistency.

Then, the relationship between full data and observed data density can be well defined by:

$$f(A, X, Y, R = 1_p) = f(A, X, Y)P(R = 1_p \mid A, X, Y), \tag{1}$$

and thus we are able to identify $f(A, X, Y)$ by identifying $P(R = 1_p \mid A, X, Y)$.

To achieve this goal, the key functional by utilizing Assumption 3 for equality and Assumption 4 for well-definedness is given by

$$\xi_{ra}(X) \coloneqq \frac{P(R = r \mid A = a, X, Y)}{P(R = 1_p \mid A = a, X, Y)} = \frac{P(R = r \mid A = a, X)}{P(R = 1_p \mid A = a, X)}, \quad (a = 0, 1;\ r \in \mathcal{R}). \tag{2}$$

This functional plays a crucial role because it relates the full data distribution with the complete-case distribution through the following integral equation:

$$f(A = a, X_r, Y, R = r) = \int \xi_{ra}(X)\, f(A = a, X, Y, R = 1_p)\, d\nu(X_{\bar{r}}). \tag{3}$$

Equation (3) holds by virtue of Lemma S1. There are a few remarks for the non-standard notations in these equations (sometimes we refer to cases such as Equation (3) as multiple equations based on values of a and r). We use these non-standard notations for brevity of the formulas and to match Yang et al. (2019).

Remark 3. *Whenever a missingness-indicator vector $r \in \{0, 1\}^p$ satisfies $r = 1_p$ (i.e. all components are observed), we adopt the convention that $X_{\bar{r}} = X_{0_p} = \varnothing$, and accordingly $\int f(X)d\nu(\varnothing) = f(X)$. This is a not a standard result in Lebesgue integration theory; rather we adopt it here as a convenient convention for the degenerate case when the set of integration variables is empty.*

Remark 4. *One should treat X and Y as random variables. If we denote a particular realization by $X = x$ and $Y = y$. Then*

$$f(A = a, X_r = x_r, Y = y, R = r) = \int \xi_{ra}(x) f(A = a,\ X = x_r,\ Y = y,\ R = 1_p) d\nu(x_{\bar{r}}).$$

However, here and in subsequent formulas we suppress the explicit "$= x$" and "$= y$" notations for realizations, in the same spirit as the classic paper Rosenbaum and Rubin (1983), where one writes $e(X) \coloneqq P(A = 1 \mid X)$ rather than $e(x) = P(A = 1 \mid X = x)$ for ATE identification.

Remark 5. *The generic measure $\nu(\cdot)$ is for the reference of continuous or discrete vari-*

ables. For a continuous measure,

$$f(A=a, X_r=x_r, Y, R=r) = \int \xi_{ra}(x)\, f(A=a, X=x, Y, R=1_p)\, dx_{\bar{r}}.$$

And for a discrete measure,

$$f(A=a, X_r=x_r, Y, R=r) = \sum_{\bar{r}} \xi_{ra}(x)\, f(A=a, X=x, Y, R=1_p).$$

Solving this linear operator reduces to tackling a Fredholm integral equation of the first kind (Kress, 1999). Such equation relies on the concept of completeness for general X and Y, which is closely related to the notion of a complete statistic (Lehmann and Scheffé, 1950; Newey and Powell, 2003).

Definition 1. *A function $f(X, Y)$ is complete in Y if, for any square-integrable function $g(X)$, the condition $\int g(X) f(X, Y)\, d\nu(X) = 0$ implies that $g(X) = 0$ almost surely.*

Although the completeness assumption is conceptual, intuitive sufficient conditions are provided by Newey and Powell (2003) and by D'Haultfoeuille (2011) (see Supplementary material). Intuitively, solving the Fredholm integral equation amounts to recognizing that we observe weighted averages of X under different weighting schemes, even though some components of X are unobserved. In the discrete case, this condition requires that the dimension of Y exceeds the dimension of X. Building on their results, we introduce the following assumption to guarantee identification of the functional ξ_{ra}.

Assumption 5. *The joint distribution $f(A=a, X, Y, R=1_p)$ is complete in Y for $a = 0, 1$.*

Remark 6. *Assumption 2 (unconfoundedness) is actually implied by Assumption 5 (completeness), and a demonstration of this appears in the proof of Remark 1 in Yang et al. (2019). In particular, if $e(X) > 0$ were violated so that $e(X) = 0$ on a nonnegligible measurable set, then there would exist a measurable region on which $f(A=a, X, Y, R=1_p) = 0$, which in turn would invalidate completeness. Hence, in all subsequent theorems and lemmas we omit Assumption 2, treating it as redundant under Assumption 5.*

We include the standard identification from the literature to identify ξ_{ra} in the Supplementary material S.2.1. Our first identification strategy in the first set of integration-based derivation include this step (see Section 3.2.2, Theorem 1) employs a similar logical structure. In this equation (3), the observed distribution on the left reflects averages over unobserved values of X, taken under different realizations of Y. Having identified ξ_{ra}, we

can directly recover $\tau(X)$ from the observed data by

$$\tau(X) = \mathbb{E}(Y \mid A = 1, X) - \mathbb{E}(Y \mid A = 0, X) \tag{4}$$
$$= \mathbb{E}(Y \mid A = 1, X, R = 1_p) - \mathbb{E}(Y \mid A = 0, X, R = 1_p), \tag{5}$$

where the first equation is by Assumption 1 and the second equation is by Assumption 3. We can thus obtain τ accordingly,

$$\tau = \sum_{a=0}^{1} \int \tau(X) \frac{f(A = a, X, R = 1_p)}{P(R = 1_p \mid A = a, X)} \, d\nu(X) \tag{6}$$

by averaging $\tau(X)$ over the identified $f(X)$ because the numerator is from the complete-case distribution and the denominator comes from the identified ξ_{ra}.

3.2.2 First set of identification formulas: completeness in Y

We relate the functional ξ_{ra} to the probability of observing the missing data pattern r given treatment $A = a$, confounders X, i.e. $P(R = r \mid A = a, X) = \frac{\xi_{ra}(X)}{\sum_{r' \in \mathcal{R}} \xi_{r'a}(X)}$, which naturally connects to the joint propensity score, for treatment and complete observation, formalized in Theorem 1. Although equation (6) makes use of ξ_{ra}, we opt for a more direct strategy that exploits this connection to the joint propensity score to identify causal effects.

Theorem 1. *Under Assumptions 1, 3-5,*

$$\tau = \mathbb{E}\left[\frac{A\mathbb{I}\{R = 1_p\}Y}{e_1(X)} - \frac{(1 - A)\mathbb{I}\{R = 1_p\}Y}{e_0(X)}\right], \tag{7}$$

where the denominators in the identification equation

$$e_a(X) = P(A = a, R = 1_p \mid X) = \frac{f(A = a, X, R = 1_p)}{\sum_{a'=0}^{1} \frac{f(A=a', X, R=1_p)}{P(R=1_p \mid A=a', X)}}, \quad (a = 0, 1) \tag{8}$$

are identified by (3).

See the proof in the Supplementary material S.3. Theorem 1 employs the functional ξ_{ra} to recover the joint propensity score $e_a(X)$ for complete-case analysis. This approach motivates considering an alternative integral equation. Specifically, the causal effect can be identified using a nonparametric observed-data conditional average treatment effect (OBCATE) estimation based on an alternative set of integral equations. Before discussing

the details of this plug-in approach, define

$$f_{ra}(X) \coloneqq f(X \mid A = a, R = r) \qquad (a = 0, 1;\ r \in \mathcal{R}) \tag{9}$$

as the possible unobserved density of confounders on the treatment and missing indicator and, importantly, serves as the intermediate link between the observed values and the complete-case distribution.

Lemma 1. *Under Assumptions 3 and 4, for any r and a, the following integral equation holds:*

$$f(X_r, Y \mid A = a, R = r) = \int f_{ra}(X) f(Y \mid A = a, X, R = 1_p)\, d\nu(X_{\bar{r}}). \tag{10}$$

See the proof in the Supplementary material S.3. Lemma 1 is the basis of our regression model. $f(X_r, Y \mid A = a, R = r)$ and $f(Y \mid A = a, X, R = 1_p)$ are identifiable from the observed data. We have thus turned the identification of $f_{ra}(X)$ to the problem of solving an alternative integral equation. The uniqueness of the solution again relies on the completeness.

Assumption 5′. *The conditional marginal distribution $f(Y \mid A = a, X, R = 1_p)$ is complete in Y, for $a = 0, 1$.*

The two completeness assumptions 5 and 5′ are in fact equivalent when $f(A = a, X, R = 1_p) > 0$ for any a. This equivalence, by Bayes' rule, is particularly clear in the illustrative case of discrete confounders (see Supplementary material). Moreover, Assumption 5 implies Assumption 2. See its proof by Remark 1 in Yang et al. (2019).

With Assumptions 3 and 5, the following theorem guarantees the uniqueness of the functional $f_{ra}(X)$. Although recovering the full distribution is unnecessary to identify the causal effect, we can still do so since the causal effect is a functional of that distribution.

Lemma 2. *Under Assumptions 3 and 5, for any r and a, there is a unique solution $f_{ra}(X)$ to equation (10) and the distribution of (A, X, Y, R) is identifiable.*

See the proof in the Supplementary material S.3. If the joint distribution of (A, X, Y) is identifiable, a standard argument shows that τ is identifiable under Assumptions 1 and 2. Nevertheless, we provide explicit identification formulas for τ, which serve as the foundation for constructing the nonparametric OBCATE estimator. We define

$$\tau(A, X_R, R) = \int \tau(X) f(X_{\bar{R}} \mid A, X_R, R)\, d\nu(X_{\bar{R}}).$$

We also define $\mu_a(A, X_R, R) \coloneqq \int \mathbb{E}(Y \mid A = a, X) f(X_{\bar{R}} \mid A, X_R, R)\, d\nu(X_{\bar{R}})$ whose identification strategy is identical. Our goal is to identify $\tau(A, X_R, R) = \mathbb{E}[\tau(X) \mid A, X_R, R] = \mathbb{E}[\mu_1(X) - \mu_0(X) \mid A, X_R, R]$ since these quantities depend only on observed variables. The primary objective is to determine the causal effect using OBCATE.

Theorem 2. *Under Assumptions 1, 3, and 5, τ is identified by*

$$\tau = \mathbb{E}[\tau(A, X_R, R)], \tag{11}$$

where

$$\tau(A = a, X_r, R = r) = \int \frac{\tau(X)\, f_{ra}(X)}{f(X_r \mid A = a, R = r)}\, d\nu(X_{\bar{r}}) \tag{12}$$

identified by (10).

See the proof in the Supplementary material S.3.

3.2.3 Alternative identification formula: completeness in X

The previous set of integral equations relies on the completeness assumption in Y. A discrete example illustrates that the dimension of Y should exceed that of X. However, when X is high-dimensional, completeness in Y is likely to fail, raising the question of whether the unknown parameters can still be recovered in such a scenario. More concretely, if completeness holds in X, can we still identify, for example, the probability of a given missing-data pattern r, $\mathbb{P}(R = r \mid A = a, X)$? In this section, we demonstrate that such identification is indeed possible with extra assumptions that the following integral equation is well-identified.

Assumption 6. *For any $a \in \{0, 1\}$, there exists a strictly positive square-integrable function ϕ_a such that, almost surely,*

$$\int \phi_a(Y)\, f(Y \mid A = a, X, R = 1_p)\, d\nu(Y) = \frac{1}{P(R = 1_p \mid A = a, X)}. \tag{13}$$

Two key observations arise from equation (13). First, the conditional distribution $f(Y \mid A = a, X, R = 1_p)$ is observed, whereas the inverse probability of the missingness pattern, $P(R = 1_p \mid A = a, X)^{-1}$, is unobserved. Second, we propose a function $\phi_a(Y)$ that characterizes the relationship between these two quantities at different values of the support of X; these functions are not necessarily densities and therefore may not exist. Identification of $\phi_a(Y)$ enables recovery of the unobserved density of missing data. We focus on the case

when $r = 1_p$ as $\mathbb{P}(R = 1_p \mid A = a, X)$ is sufficient for identifying the causal parameter of interest. Notice,

$$
\begin{aligned}
f(A = a, Y) &= \int \frac{1}{P(R = 1_p \mid A = a, X)} f(A = a, X, Y, R = 1_p)\, d\nu(X) \\
&= \int \left[\int \phi_a(\widetilde{Y})\, f(\widetilde{Y} \mid A = a, X, R = 1_p)\, d\nu(\widetilde{Y}) \right] f(A = a, X, Y, R = 1_p)\, d\nu(X) \\
&= \int \phi_a(\widetilde{Y}) \left[\int f(\widetilde{Y} \mid A = a, X, R = 1_p)\, f(A = a, X, Y, R = 1_p)\, d\nu(X) \right] d\nu(\widetilde{Y}).
\end{aligned} \tag{14}
$$

The first equality holds by definition and the outcome-independent missingness assumption. The second equality follows from equation (13). The third equality arises by interchanging the order of integration. We make the following assumption to identify ϕ_a.

Assumption 7. *For each $a \in \{0, 1\}$, the marginal distribution $f(Y \mid A = a, X, R = 1_p)$ is complete in X.*

First, note Assumption 7 implies that the joint distribution $f(A = a,\ X,\ Y,\ R = 1_p)$ is complete in X (see the Supplementary material for the proof). Second, this assumption on the low-dimensional structure allow us to identify the conditional probability $P(R = 1_p \mid A = a, X)$ by implying that $\int f(\widetilde{Y} \mid A = a, X, R = 1_p)\, f(A = a, X, Y, R = 1_p)\, d\nu(X)$ is complete in Y. One implicit part (see the proof in the Supplementary material S.8.3) about this identification is that we need to show for any $a \in \{0, 1\}$ we have the property $\mathrm{Im}(\Theta_{a1}) \cap \ker(\Theta_{a2}) = \{0\}$, where $\Theta_{a1} : L^1(Y) \to L^1(X)$ and $\Theta_{a2} : L^1(X) \to L^1(Y)$

$$
\begin{aligned}
(\Theta_{a1} g)(X) &:= \int g(Y)\, f(Y \mid A = a, X, R = 1_p)\, d\nu(Y), \\
(\Theta_{a2} u)(Y) &:= \int u(X)\, f(A = a, X, Y, R = 1_p)\, d\nu(X).
\end{aligned}
$$

Consequently, applying equation (7) identifies the ATE. The following theorem formalizes this argument.

Theorem 3. *Under Assumptions 6 and 7, the missing data mechanism $P(R = 1_p \mid A = a, X)$ is identified.*

See the proof in the Supplementary material S.5. The assumptions and identification formulas are fairly abstract, so to build intuition we present discrete examples that illustrate how different identification strategies depend on the relative dimensions of Y and X. For

the continuous case of Assumption 6, we provide sufficient conditions showing that a series estimator for ϕ_a is adequate when $P(R = r \mid A = a, X)$ follows a logistic specification under mild regularity conditions. The central idea is that the model admits a low-dimensional structure: variation in the lower-dimensional component Y is enough to capture all of the relevant variation required for identification.

3.2.4 Sufficient conditions for Assumption 6

We propose a sufficient condition for the existence of a low-dimensional structure in data, akin to the completeness assumption in statistical theory. Our central premise is that the observed data exhibit sufficient variation to infer the limited variation present in the missing data pattern.

Proposition 1. *Suppose that for fixed a and r, the conditional distribution of $Y \mid A = a, X, R = r$ is a regular exponential family with density $f(Y \mid A = a, X, R = r) = \exp\{\eta_r(X)^\top g(Y) - \psi\big(\eta_r(X)\big)\}\, h(Y)$, and that the missingness mechanism is multinomial logit $\pi_r(X) := P(R = r \mid A = a, X) = \exp(\beta_r^\top X)/\sum_{r'} \exp(\beta_{r'}^\top X)$, where X can include a constant coordinate. Suppose there exist vectors $\{t_{r'r}\}_{r'}$ and positive constants $\{c_{r'r}\}_{r'}$ such that for all r', $\psi\big(\eta_r(X) - t_{r'r}\big) - \psi\big(\eta_r(X)\big) = (\beta_{r'} - \beta_r)^\top X + \log c_{r'r}$. Then the function $\phi_{ra}(Y) = \sum_{r'} c_{r'r}^{-1} \exp\Big\{ - t_{r'r}^\top g(Y)\Big\}$ satisfies Assumption 6.*

See the proof in the Supplementary material S.5.

Remark 7. *When $R \in \{0, 1\}$ and $\pi_1(X) = 1/[1 + \exp(\beta^\top X)]$, $1/\pi_1(X)$ reduces to $1 + \exp(\beta^\top X)$. Proposition 1 holds with a constant t that generates $\exp(\beta^\top X)$ by $\psi(\eta - t) - \psi(\eta)$. In a gaussian exponential family with $Y \mid X \sim \mathcal{N}(\mu_0 + MX, \Sigma)$, one explicit choice is*

$$\phi_a(Y) = 1 + \exp(-t^\top Y - \tfrac{1}{2} t^\top \Sigma t), \quad \text{with } M^\top t = \beta,$$

which gives $\mathbb{E}[\phi_a(Y) \mid X] = 1 + \exp(-\beta^\top X) = 1/\pi_1(X)$.

4 Estimation

In this section, we present two foundational plug-in estimators and, building on these, introduce two doubly robust approaches: one based on a straightforward plug-in strategy and the other leveraging double machine learning. Details on the estimation of the nuisance parameters are provided in S.10.

4.1 Doubly robust estimator

Our proposed doubly robust approach is not tied to a single identification strategy. It is useful to consider two semiparametric models that impose parametric structure on different parts of the observed data likelihood while leaving the rest unrestricted. In the first model, $\mathcal{M}_Y$, we assume that Assumptions 1, 3, and 5 hold, so that identification is achieved through completeness in Y as in equation (10). In the second model, $\mathcal{M}_X$, we instead assume that Assumptions 1, 3, 6, and 7 hold, so that identification relies on completeness in X as captured by equation (13) and the alternative representation in (14).

The second set of integral equations motivates an estimator of the joint propensity score that builds on completeness in X.[4] Specifically, starting from equation (13), we can estimate the conditional density $f(Y \mid A = a, X, R = 1_p)$ using kernel or series-based methods, and then construct an estimator for $\phi_a(Y)$ via a nonparametric instrumental variables procedure that exploits the moment restriction in (13). Given these building blocks, the joint density $f(A = a, X, Y, R = 1_p)$ can be approximated empirically, and the outer integration over $\widetilde{Y}$ in (14) is carried out numerically. Combining these steps yields an estimator of the joint propensity score $\hat{e}_a(X)$ that justifies the identification Theorem 1.

To derive the estimator of OBCATE by equation (12), precisely, $\mu_a(A, X_R, R) = \int \mathbb{E}(Y \mid A = a, X) f(X_{\bar{R}} \mid A, X_R, R)\, d\nu(X_{\bar{R}})$, we estimate $\mathbb{E}(Y \mid A = a, X) = \mathbb{E}(Y \mid A = a, X, R = 1_p)$ by complete-case regression, estimate $P(A = a \mid R = r)$ empirically and $f(X_r \mid R = r)$ nonparametrically, and apply the two-stage least squares procedure to recover $f_{ra}(X)$ by the alternative integration formula (10). We therefore yield $\hat{\mu}_a(A, X_R, R)$. There are some technical details for the nonparametric two-stage least squares procedures, and we discuss these details of the nonparametric two-stage least squares procedures for estimating ξ_{ra} and f_{ra} in the Supplementary material.

Then, based on the identification strategies in Theorem 2 and Theorem 3, there are two natural weighted estimators for the ATE. The first is the inverse joint propensity weighting (IJPW) estimator:

$$\hat{\tau}_{\text{IJPW}} = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{A_i \, \mathbb{I}\{R_i = 1_p\} \, Y_i}{\hat{e}_1(X_i)} - \frac{(1 - A_i) \, \mathbb{I}\{R_i = 1_p\} \, Y_i}{\hat{e}_0(X_i)} \right]. \tag{15}$$

The second is a marginalized regression (MREG) estimator, which depends on the regres-

[4] Alternatively, one may invoke equation (3), apply a nonparametric two-stage least squares procedure to estimate $P(R = 1_p \mid A = a, X)$ (equivalently $\xi_{ra}(X)$), and then substitute into Bayes' theorem to derive an estimator $\hat{e}_a(X)$.

sion of potential outcomes $\hat{\tau}(X)$ and imputed weight $\hat{f}(X_{\bar{R}} \mid A, X_R, R)$:

$$\hat{\tau}_{\text{MREG}} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(A_i, X_{R_i}, R_i) - \hat{\mu}_0(A_i, X_{R_i}, R_i)\right]. \tag{16}$$

For $\hat{\mu}_a(A, X_R, R)$ and $\hat{e}_a(X)$ for $a = 0, 1$ that are consistent, under weak conditions, both $\hat{\tau}_{\text{IJPW}}$ and $\hat{\tau}_{\text{MREG}}$ are themselves consistent estimators of the ATE.

The convergence rates of the two estimators in (15) and (16) typically depend on the accuracy of the intermediate estimators for the joint propensity score and the marginalized outcome regression. Importantly, they are also influenced by the quality of the estimated solution to the integral identification functions. Given that the average treatment effect can be estimated through both inverse joint propensity weighting and marginalized outcome regression, we propose the augmented inverse joint propensity weighting (AIJPW) estimator:

$$\begin{aligned}\hat{\tau}_{\text{AIJPW}} =& \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\hat{\mu}_1(A_i, X_{R_i}, R_i) + \frac{A_i\,\mathbb{I}\{R_i = 1_p\}\left[Y_i - \hat{\mu}_1(A_i, X_{R_i}, R_i)\right]}{\hat{e}_1(X_i)} \\ &- \hat{\mu}_0(A_i, X_{R_i}, R_i) - \frac{(1 - A_i)\,\mathbb{I}\{R_i = 1_p\}\left[Y_i - \hat{\mu}_0(A_i, X_{R_i}, R_i)\right]}{\hat{e}_0(X_i)}\Bigg\}.\end{aligned} \tag{17}$$

We also propose the double machine learning (DML) estimator using cross-fitting by Chernozhukov et al. (2018):

$$\begin{aligned}\hat{\tau}_{\text{DML}} =& \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\hat{\mu}_1^{[-k(i)]}\Big(A_i, X_{R_i}, R_i\Big) + \frac{A_i\,\mathbb{I}\{R_i = \mathbf{1}_p\}\left[Y_i - \hat{\mu}_1^{[-k(i)]}(A_i, X_{R_i}, R_i)\right]}{\hat{e}_1^{[-k(i)]}(X_i)} \\ &- \hat{\mu}_0^{[-k(i)]}\Big(A_i, X_{R_i}, R_i\Big) - \frac{(1 - A_i)\,\mathbb{I}\{R_i = \mathbf{1}_p\}\left[Y_i - \hat{\mu}_0^{[-k(i)]}(A_i, X_{R_i}, R_i)\right]}{\hat{e}_0^{[-k(i)]}(X_i)}\Bigg\}.\end{aligned} \tag{18}$$

We divide the overall sample $\mathcal{I}$ into $K \geq 2$ disjoint folds, denoted $\mathcal{I}_1, \ldots, \mathcal{I}_K$. For each fold k, we fit the nuisance estimators $\hat{e}_a^{[-k]}(x)$ and $\hat{\mu}_a^{[-k]}(a, x_r, r)$ for $a \in \{0, 1\}$, using only the data outside the kth fold. Let $k(i)$ be the index of the fold to which the observation $(A_i,\ X_{R_i},\ Y_i,\ R_i)$ belongs. That is to say, we use what is commonly called the DML 1 estimator, though an alternative is the DML 2 estimator. Under conditions in which K is small or the nuisance functions are estimated via auxiliary procedures satisfying standard regularity conditions, the two estimators typically exhibit similar performance. However, in more complex settings, DML 2 enjoys superior higher-order properties. Velez (2024) shows

that, under an asymptotic framework where $K \to \infty$ as $n \to \infty$, DML 2 asymptotically dominates DML 1 in terms of bias and mean squared error. To simplify the notation in our discussion of DML, we suppress the fold-index and denote $\hat{\mu}_a(A_i, X_{R_i}, R_i)$ and $\hat{e}_a(X_i)$ instead of $\hat{\mu}_a^{[-k(i)]}(A_i, X_{R_i}, R_i)$ and $\hat{e}_a^{[-k(i)]}(X_i)$, on the understanding that the fold $k(i)$ is implicitly determined by the observation i.

4.2 Consistency

The classical doubly robust property in the literature requires consistency under a somehow different condition that at least one of these two nuisance models is correctly specified by some parametric assumption. Equivalently, one may view the estimation as being conducted over the union model $\mathcal{M}_X \cup \mathcal{M}_Y$. In other words, so long as either the joint propensity score or the outcome regression model is valid, the estimator still converges to the true parameter, even if the other working model is misspecified. This feature is also appealing in practice, since it guards against bias arising from a single model's misspecification (Bang and Robins, 2005; Sant'Anna and Zhao, 2020a; Wager, 2024; Testa et al., 2025). To formalize this for our nonparametric setup, we impose the following condition for the AIJPW estimator.

Assumption 8. *(Consistency of AIJPW) For any $a \in \{0,1\}$, and $i \in [n]$,*
(a) There exists some constant c such that $0 < c \leq \hat{e}_a(X_i) \leq 1$.
(b) Y_i, $\mu_a(A_i, X_{R_i}, R_i)$, and $\hat{\mu}_a(A_i, X_{R_i}, R_i)$ have bounded second moments.
(c) One of the following conditions (c.1) or (c.2) holds:

(c.1) The nuisance parameter estimator $\hat{e}_a(X)$ satisfies $\mathbb{E}[\hat{e}_a(X_i) - e_a(X_i)]^2 = o(1)$ and $\mathbb{E}[\frac{1}{\hat{e}_a(X_i)} - \frac{1}{e_a(X_i)}]^2 = o(1)$. There exists an estimator $\bar{\mu}_a(A, X_R, R)$ independent of $\mathcal{O}_i$ such that $\mathbb{E}[\hat{\mu}_a(A_i, X_{R_i}, R_i) - \bar{\mu}(A_i, X_{R_i}, R_i)]^2 = o(1)$.

(c.2) The nuisance parameter estimator $\hat{\mu}_a(A_i, X_{R_i}, R_i)$ satisfies $\mathbb{E}[\hat{\mu}_a(A_i, X_{R_i}, R_i) - \mu_a(A_i, X_{R_i}, R_i)]^2 = o(1)$. There exists an estimator $\bar{e}_a(X_i)$ independent of $\mathcal{O}_i$ such that $\mathbb{E}[\frac{1}{\hat{e}_a(X_i)} - \frac{1}{\bar{e}_a(X_i)}]^2 = o(1)$.

Existing work generally relies solely on a single identification moment condition to establish the classical doubly robust property (Ding, 2024). In contrast, our paper advances the framework in two ways: we not only establish doubly robustness under a broader set of assumptions (nonparametric), but we also prove the consistency of the estimator without the Donsker. Our key innovation draws on the stronger convergence conditions considered in Huo et al. (2025), where the authors eliminate the need for a Donsker class argument by assuming that no individual observation exerts excessive influence on the estimation. Unlike

their framework, our approach does not require both nuisance estimators to be consistent simultaneously. Instead, it suffices that one of them is well-behaved for our result to hold. We therefore present the following theorem, which formally establishes the classical doubly robustness of consistency under our assumptions. There is an important remark: in the classical doubly robust framework, each nuisance model is assumed to converge to some fixed limit (which may be the true model if correctly specified, or a misspecified target otherwise). $\mathbb{E}[\hat{\mu}(A_i, X_{R_i}, R_i) - \bar{\mu}(A_i, X_{R_i}, R_i)]^2 = o(1)$ and $\mathbb{E}[\frac{1}{\hat{e}_a(X_i)} - \frac{1}{\bar{e}_a(X_i)}]^2 = o(1)$ play the same role here.

Theorem 4. *Under Assumptions 1–4 and 9, the doubly robust estimator $\hat{\tau}_{AIJPW}$ is consistent.*

To formalize classical doubly robustness for the DML estimator, we impose the following condition.

Assumption 9. *(Consistency of DML) For any $a \in \{0, 1\}$, and $i \in [n]$,*
(a) There exists some constant c such that $0 < c \leq \hat{e}_a(X_i) \leq 1$.
(b) Y_i, $\mu_a(A_i, X_{R_i}, R_i)$, and $\hat{\mu}_a(A_i, X_{R_i}, R_i)$ have bounded second moments.
(c) One of the following conditions (c.1) or (c.2) holds:
(c.1) The nuisance parameter estimator $\hat{e}_a(X)$ satisfies $\mathbb{E}\left\{[\hat{e}_a(X_i) - e_a(X_i)]^2 \mid \mathcal{I}_{-k(i)}\right\} = o_p(1)$ and $\mathbb{E}\left\{[\frac{1}{\hat{e}_a(X_i)} - \frac{1}{e_a(X_i)}]^2 \mid \mathcal{I}_{-k(i)}\right\} = o_p(1)$.
(c.2) The nuisance parameter estimator $\hat{\mu}_a(A, X_R, R)$ satisfies $\mathbb{E}\{[\hat{\mu}_a(A_i, X_{R_i}, R_i) - \mu_a(A_i, X_{R_i}, R_i)]^2 \mid \mathcal{I}_{-k(i)}\} = o_p(1)$.
(d) There exists some constant $\kappa > 0$ such that $\frac{\mathcal{I}_k}{n} \geq \kappa$ for any $k \in [K]$.

As in Chernozhukov et al. (2018), Assumptions 8 and 9 can be readily extended to their high-probability versions, thereby allowing for estimators with heavy-tailed distributions. Nevertheless, Our key implication is that, under plug-in or cross-validated estimation, the AIJPW and DML estimators remain consistent provided that at least one of the nuisance components is correctly specified. We now formalize the classical doubly robust property of the DML estimator in the theorem below.

Theorem 5. *Under Assumptions 1–4 and 8, the doubly robust estimator $\hat{\tau}_{DML}$ is consistent.*

5 Inference

5.1 Asymptotic normality

We also establish a rate doubly robust property: the estimator is $\sqrt{n}$ consistent requiring both the joint propensity score and marginalized outcome regression models are consistent.

The AIJPW estimator begins by modeling conditional outcomes to estimate the ATE. It then adjusts for potential biases in these models by applying inverse joint probability weighting to the residuals. This method combines the strengths of marginalized outcome regression and joint propensity score weighting, offering robustness against misspecification in either model. Notably, under mild regularity conditions, and even when nuisance parameters are estimated flexibly using nonparametric or machine learning methods, the AIJPW and DML estimators achieve the $\sqrt{n}$-rate of convergence, as demonstrated by Kennedy (2016); Chernozhukov et al. (2018). This part is well-known in the literature, so we present the assumptions in a concise manner.

Formally, suppose we estimate the nuisance parameters $\mu_a(A, X_R, R)$ and $e_a(X)$ by $\hat{\mu}_a(A, X_R, R)$ and $\hat{e}_a(X)$, respectively, such that the following assumption holds:

Assumption 10. *For each $a \in \{0, 1\}$, there exists some constant c such that $0 < c \leq \hat{e}_a(X_i) \leq 1$. Y_i, $\mu_a(A_i, X_{R_i}, R_i)$, and $\hat{\mu}_a(A_i, X_{R_i}, R_i)$ have bounded second moments. Moreover, we assume that the estimators $\hat{\mu}_a(A, X_R, R)$ and $\hat{e}_a(X)$ satisfy the following rate conditions:*

(a) $\|\hat{\mu}_a(A, X_R, R) - \mu_a(A, X_R, R)\|_2 = o_P(1)$.

(b) $\|\hat{e}_a(X) - e_a(X)\|_2 = o_P(1)$.

(c) $\|\hat{\mu}_a(A, X_R, R) - \mu_a(A, X_R, R)\|_2 \cdot \|\hat{e}_a(X) - e_a(X)\|_2 = o_P(n^{-1/2})$.

This assumption holds if each nuisance estimator converges at a rate of $o_p(n^{-1/4})$ in the $L^2(P)$-norm, where $\|\hat{\theta}\|_2^2 = \int \hat{\theta}(w)^2 \, dP(w)$. We use the notation $o_p(\cdot)$ to indicate convergence in probability, distinguishing it from the deterministic $o(\cdot)$ used in Assumption 8. This distinction is crucial as we do not explicitly control for the randomness inherent in the estimators $\hat{e}$ and $\hat{\mu}$, which would otherwise make this assumption more stringent. (See Lemma in the Supplementary material) In classical semiparametric theory, ensuring that nuisance estimators belong to a Donsker class guarantees that the additional variability from estimating $\hat{e}$ and $\hat{\mu}$ becomes asymptotically negligible. This condition allows for the use of more flexible methods while still achieving consistency under strong conditions on the randomness of the estimated nuisance parameters.

Equivalently, one may view the estimation under Assumption 10 as being conducted over the intersection model $\mathcal{M}_X \cap \mathcal{M}_Y$. The rate condition connects the literature on causal inference with machine learning methods that converge at a slow rate while still producing asymptotically valid confidence intervals. Then, we introduce the following assumption as part of the standard procedure for implementing doubly robust estimators.

Assumption 11. *We assume that either (a) $\hat{\mu}_a(A, X_R, R)$ and $\hat{e}_a(X)$ are obtained by fixed K-fold cross-fitting with $K > 1$, or (b) $\mu_a(A, X_R, R)$, $\hat{\mu}_a(A, X_R, R)$, $e_a(X)$, $\hat{e}_a(X)$ belong to a Donsker class.*

This assumption ensures the necessary rate conditions for the doubly robust estimator to achieve asymptotic normality and efficiency. One can interpret the estimation under $\mathcal{M}_X \cap \mathcal{M}_Y$. Cross-fitting helps mitigate overfitting and relaxes the need for Donsker class conditions, as demonstrated by Kennedy (2016); Chernozhukov et al. (2018). Alternatively, assuming the estimators lie within a Donsker class provides a route to control the empirical process term in the asymptotic analysis, as discussed by Andrews (1994). One weaker assumption for the Donsker class is by assuming the asymptotic equicontinuity (van der Vaart, 2000). Assume that the function class $\mathcal{F} = \Phi_\omega$ satisfies asymptotic equicontinuity with respect to the L^1–norm: for every $\varepsilon > 0$,

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\Bigg(\sup_{\substack{f,g \in \Phi_\omega \\ \mathbb{E}[|f-g|] < \delta}} \mathbb{G}_n(f - g) > \varepsilon \Bigg) = 0,$$

where $\mathbb{G}_n = \sqrt{n}\,(\mathbb{P}_n - \mathbb{P})$. With these assumptions, we now state the following theorem, which establishes the $\sqrt{n}$-consistency and asymptotic linearity of our doubly robust estimators.

Theorem 6. *Under Assumptions 1–4, 10, and 11, the doubly robust estimators $\hat{\tau}_{AIJPW}$ and $\hat{\tau}_{DML}$ are asymptotically linear. Let $\hat{\tau}$ denote either $\hat{\tau}_{AIJPW}$ or $\hat{\tau}_{DML}$, then*

$$\sqrt{n}\,(\hat{\tau} - \tau) \rightsquigarrow \mathcal{N}(0, V) \tag{19}$$

$$V = Var\left[\tau(A, X_R, R)\right] + \mathbb{E}\left[\frac{\sigma_1^2(X)}{e_1(X)}\right] + \mathbb{E}\left[\frac{\sigma_0^2(X)}{e_0(X)}\right] \tag{20}$$

where $\sigma_a^2(X) = Var\left[Y_i(a) \mid X\right]$ for $a = 0, 1$.

See the proof in Supplementary material S.7.1.

Remark 8. *In the Appendix we show that*

$$\begin{aligned}\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \Bigg\{ & \mu_1(A_i, X_{R_i}, R_i) + \frac{A_i\, \mathbb{I}\{R_i = 1_p\}\,[Y_i - \mu_1(A_i, X_{R_i}, R_i)]}{e_1(X_i)} \\ & - \mu_0(A_i, X_{R_i}, R_i) - \frac{(1 - A_i)\, \mathbb{I}\{R_i = 1_p\}\,\,[Y_i - \mu_0(A_i, X_{R_i}, R_i)]}{e_0(X_i)} \Bigg\} + o_p(n^{-1/2}).\end{aligned}$$

Thus, up to an $o_p(n^{-1/2})$ remainder, an AIJPW estimator is first-order equivalent to a sum of independent and identically distributed random influence functions. This asymptotically linear representation immediately justifies constructing confidence intervals and conducting hypothesis tests either by the usual normal theory approximations or by applying the bootstrap.

Under the conditions of Theorem 6, let $\widehat{V}$ denote a consistent estimator of V. For example,

$$\widehat{V} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\tau}(A_i, X_{R_i}, R_i) - \bar{\hat{\tau}}\right]^2 + \frac{1}{n}\sum_{i=1}^{n}\frac{A_i\,\mathbb{I}\{R_i = 1_p\}\hat{\sigma}_1^2(X_i)}{\hat{e}_1^2(X_i)} + \frac{1}{n}\sum_{i=1}^{n}\frac{(1-A_i)\,\mathbb{I}\{R_i = 1_p\}\hat{\sigma}_0^2(X_i)}{\hat{e}_0^2(X_i)}.$$

Then an asymptotically valid two-sided $1-\alpha$ confidence interval for τ is $\left[\hat{\tau} \pm z_{1-\alpha/2}\sqrt{\widehat{V}/n}\right]$. it can be shown (with somewhat stronger conditions, typically higher-order moment assumptions on the parameters and estimators) that this confidence interval has uniform asymptotic validity:

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}}\left|P\left(\tau \in \left[\hat{\tau} \pm z_{1-\alpha/2}\sqrt{\frac{\widehat{V}}{n}}\right]\right) - (1-\alpha)\right| = 0.$$

5.2 Semiparametric efficiency bounds

In this section, we focus on the optimal inference for the OBCATE parameter $\tau^\star := \mathbf{E}\big[\tau(A = a, X_r, R = r)\big]$, and under Assumptions 1, 3, and 5, Theorem 2 implies that $\tau = \tau^\star$. Our aim is to study the functional that achieves the semiparametric efficiency bound and possesses local efficiency. In the above theorems, we showed that the proposed estimator is asymptotically linear. The leading linear term, commonly referred to as the influence function, captures the first-order behavior of the estimator under the stated assumptions. A natural question that arises is whether this influence function is in fact the efficient influence function, in the sense that no other regular estimator is more efficient than the current one.[5]

To address this question, we examine whether the influence function corresponds to the efficient influence function for $\tau^\star$ in a nonparametric model, where the full data density is assumed to be estimable directly from the observed data with structural restrictions. Since

[5]Traditional local asymptotic minimax theory focuses on regular estimators and excludes superefficient procedures. Nonetheless, one can formulate versions of the Local Asymptotic Minimax (LAM) theorem that accommodate superefficient estimators, allowing for a meaningful comparison with regular estimators such as the maximum likelihood estimator Chamberlain (1992); Le Cam and Yang (1992).

our identification relies on the completeness conditions, parameters are bounded to ensure pathwise differentiability, and, it is also worth considering whether the additional structure from MNAR assumption can be used to sharpen the efficiency bound by restricting the tangent space. The next theorem formalizes these results.

Theorem 7. *Under Assumptions 1-7, the efficient influence function of* τ *is*

$$\varphi = \mu_1(A, X_R, R) - \mu_0(A, X_R, R) - \tau + \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_1(A, X_R, R)]}{e_1(X)} - \frac{(1-A)\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_0(A, X_R, R)]}{e_0(X)}$$

and the semiparametric efficiency bound for all regular estimators of the model $\tau^\star$ *is*

$$\mathbb{E}\left[\varphi^2\right] = Var[\tau(A, X_R, R)] + \mathbb{E}\left[\frac{\sigma_1^2(X)}{e_1(X)}\right] + \mathbb{E}\left[\frac{\sigma_0^2(X)}{e_0(X)}\right].$$

See the proof in the Supplementary material S.11.

6 Simulation

Table 1: Simulation: bias ($\times 10^{-2}$) and variance ($\times 10^{-3}$) of the point-estimator of τ, variance estimate ($\times 10^{-3}$), and coverage (%) of 95% confidence intervals.

	$n = 1,000$				$n = 3,000$				$n = 10,000$			
Method	Bias	Var	VE	Cvg	Bias	Var	VE	Cvg	Bias	Var	VE	Cvg
(a) One confounder subject to missingness												
NonPara	28.7	90.4	91.1	15.4 %	28.3	48.2	46.0	25.6 %	27.9	22.6	23.4	38.1 %
AIJPW	–22.4	46.3	48.0	88.2 %	–21.1	23.9	22.2	92.3 %	–20.6	11.0	12.2	94.1 %
DML	18.1	37.8	38.6	90.9 %	17.5	18.2	20.1	93.0 %	17.2	8.9	9.8	95.4 %
(b) Multiple confounders subject to missingness												
NonPara	–116.3	81.5	79.2	<0.1 %	–114.8	39.7	36.1	<0.1 %	–113.5	20.1	19.3	<0.1 %
AIJPW	10.2	182.0	254.5	70.6 %	9.6	89.5	102.1	80.3 %	9.1	42.3	46.0	90.4 %
DML	8.1	176.4	241.0	75.8 %	7.5	92.2	104.3	85.5 %	7.0	46.8	53.2	91.2 %

Notes: NonPara = nonparametric estimator; AIJPW = augmented inverse joint-propensity weighting estimator; DML = double machine learning estimator.

In this section, we compare the results from the nonparametric two-stage regression approach in Yang et al. (2019) with our doubly robust estimators. Our simulation methodology builds upon that of Yang et al. (2019). However, unlike their low-dimensional simulation, our setup introduces a more intricate data generation process for both potential outcomes and the probability of missingness, incorporating nonlinear terms to better capture real-world complexities. For the second simulation, while the data generation process

for the multivariate case remains consistent, we employ nonparametric methods to estimate the average treatment effect, as opposed to the parametric approaches used in their study. In each setting, we choose sample sizes of $n = 1,000$, $3,000$, and $10,000$, generating 2,000 Monte Carlo samples for each sample size. For all estimators, we use the bootstrap with 200 replicates to estimate the variances.

In the first setup, we generate two confounders $X_{1i} \sim \mathcal{N}(1,1)$ and $X_{2i} \sim \text{Bernoulli}(0.5)$. The potential outcomes are defined as

$$\begin{aligned} Y_i(0) &= 0.5 + X_{1i} + X_{2i} + \max\{X_{1i}, X_{2i}\} + \epsilon_{0i}, \\ Y_i(1) &= 1 + 3X_{1i} + 2X_{2i} + 2\max\{X_{1i}, X_{2i}\} + \epsilon_{1i}, \end{aligned}$$

where $\epsilon_{0i}, \epsilon_{1i} \sim \mathcal{N}(0,1)$, resulting in an average causal effect $\tau = \mathbb{E}[Y_i(1) - Y_i(0)] = 4.25$. Treatment assignment follows $A_i \sim \text{Bernoulli}(\pi_i)$, where $\text{logit}(\pi_i) = 1.25 - 0.5X_{1i} - 0.5X_{2i}$, and the missingness indicator for X_{1i}, denoted R_{1i}, is generated by $R_{1i} \sim \text{Bernoulli}(p_i)$, $\text{logit}(p_i) = -2 + 2X_{1i} + A_i\,(1.5 + X_{2i}) + \max\{X_{1i}, X_{2i}\}$, which yields an overall response rate of approximately 77%. We begin to show that the two integral equations are different in the sense that they return different results for the underlying distributions.

In the second setup, let $X_i = (X_{1i}, \ldots, X_{6i})$. We generate X_{1i} and X_{2i} from $\mathcal{N}(1,1)$, X_{3i} and X_{4i} from $2\text{Bernoulli}(0.5) - 1$, $X_{5i} = X_{1i} + X_{2i} + X_{3i} + X_{4i} + \epsilon_{5i}$ with $\epsilon_{5i} \sim \mathcal{N}(0,1)$, and X_{6i} from $\text{Bernoulli}(p_{6i})$ with $\text{logit}(p_{6i}) = -X_{5i}$. The potential outcomes are defined by

$$\begin{aligned} Y_i(0) &= -1.5 + X_{1i} - X_{2i} + X_{3i} - X_{4i} + X_{5i} + X_{6i} + \epsilon_{0i}, \\ Y_i(1) &= 0 - X_{1i} + X_{2i} - X_{3i} + X_{4i} - X_{5i} - X_{6i} + \epsilon_{1i}, \end{aligned}$$

where $\epsilon_{0i}, \epsilon_{1i} \sim \mathcal{N}(0,1)$. The average treatment effect is $\tau = -0.5$. The treatment indicator $A_i \sim \text{Bernoulli}(\pi_i)$, where $\text{logit}(\pi_i) = 1 + 0.5X_{1i} + 0.5X_{2i} + 0.5X_{3i} + 0.5X_{4i} - X_{5i} - X_{6i}$. Among the confounders, X_{5i} and X_{6i} have missing values, while the other confounders do not. The missingness pattern for $(X_{5i},\, X_{6i})$ is denoted by $R_i = (R_{5i},\, R_{6i})$, which takes values in the set $\{(1,1),\ (1,0),\ (0,1),\ (0,0)\}$. The distribution of R_i is multinomial with probabilities $(p_{11,i},\, p_{10,i},\, p_{01,i},\, p_{00,i})$. Specifically,

$$p_{11,i} = \frac{1}{1 + 3e^{L_i}} \qquad \text{and} \qquad p_{10,i} = p_{01,i} = p_{00,i} = \frac{e^{L_i}}{1 + 3e^{L_i}}.$$

where $L_i = -1 + 0.25A_i + 0.25X_{1i} + 0.25X_{2i} + 0.25X_{3i} + 0.25X_{4i} - 0.25X_{5i} - 0.25X_{6i}$. On average, the four missing-ness patterns $(1,1)$, $(1,0)$, $(0,1)$ and $(0,0)$ occur in about 49%, 17%, 17% and 17% of observations, respectively.

We use 10-fold cross-validation to select the tuning parameters J and B. Table 1(a) compares our doubly robust estimators with the nonparametric estimator for the case of a single confounder subject to missingness. The nonparametric method exhibits poor coverage rates for the confidence intervals. In contrast, our proposed estimators show negligible bias, maintain good coverage, and display decreasing variance as the sample size increases. Table 1(b) presents a comparison for the scenario of a single confounder subject to multiple missingness patterns. Again, our estimators demonstrate negligible bias and reliable coverage, outperforming the nonparametric approach. Overall, the DML estimator shows superior performance compared to the AIJPW estimator.

7 Empirical applications

7.1 The causal effect of job training program on employment

Table 2: Point estimate, standard error by the bootstrap and 95% confidence interval

Method	Estimation	Standard Error	95% Confidence Interval
(a) The causal effect of job training program on employment			
TDWC	0.037	0.011	[0.0154, 0.0586]
AIJPW	0.042	0.009	[0.0246, 0.0594]
DML	0.045	0.008	[0.0294, 0.0606]
(b) The causal effect of smoking on blood-lead level			
NonPara	0.207	0.072	[0.0660, 0.3480]
AIJPW	0.221	0.063	[0.0966, 0.3454]
DML	0.246	0.052	[0.1440, 0.3480]
(c) The causal effect of education on general health satisfaction			
NonPara	–0.283	0.094	[–0.4685, –0.0975]
AIJPW	–0.341	0.046	[–0.4321, –0.2499]
DML	–0.356	0.041	[–0.4376, –0.2744]

Notes: TDWC = treatment dummy without control; NonPara = nonparametric estimator; AIJPW = augmented inverse joint-propensity weighting estimator; DML = double machine learning estimator.

We analyze data from the 1994–1995 National Job Corps Study, a randomized job-training program evaluated in Lee (2009). The analytic sample consists of 9,145 individuals, including treated participants ($A = 1$) and controls ($A = 0$). The set of confounders X includes age, an indicator for ever being arrested, gender, race/ethnicity, marital status, number of children, education, mother's education (16.4% missing), father's education (33.5% missing), household income (31.9% missing), and personal income categories; all

other confounders are nearly fully observed. The outcome Y is employment status during the follow-up period. The substantial non-response in parental education and income suggests that missingness may depend on the unobserved values themselves (for example, higher-educated parents or higher-income households may be less likely to report). At the same time, it is plausible that, conditional on treatment assignment and the observed confounders, the missingness mechanism is independent of subsequent employment outcomes, a form of outcome-independent missingness.

7.2 The causal effect of smoking on blood lead level

We analyze data from the 2015–2016 National Health and Nutrition Examination Survey to estimate the causal effect of smoking on blood-lead level. The analytic sample includes 2,949 adults, of whom 1,102 are smokers ($A = 1$) and 1,847 are non-smokers ($A = 0$). All participants are aged 15 or older and reported no tobacco use other than cigarette smoking in the preceding five days. The outcome Y is the blood-lead concentration (ranging from 0.05 to 23.51 $\mu g/dL$). Confounders X include the income-to-poverty ratio, age, and gender; only the income-to-poverty ratio has missing values (14.0 % for smokers and 15.2 % for non-smokers). While the missingness of income may be not at random (for example, higher-income individuals may be less likely to report), it is also plausible that, conditional on smoking status and the observed confounders, the missing-data process is independent of the blood-lead outcome that is consistent with an outcome-independent missingness assumption.

7.3 The causal effect of education on general health satisfaction

We analyze data from the 2015–2016 National Health and Nutrition Examination Survey to estimate the average causal effect of education on general health satisfaction. The analytic sample comprises 4,845 individuals, of whom 76 % completed at least high-school education ($A = 1$) while the remaining 24 % did not ($A = 0$). The outcome Y is a general health-satisfaction score on a 1-5 scale (lower values indicate better satisfaction); in the observed data the mean of Y is 2.88 with standard deviation 0.96. Confounders X include age, gender, race, marital status, the income-to-poverty ratio, and an indicator of ever having pre-diabetes risk. Among these, the income-to-poverty ratio and the pre-diabetes-risk indicator contain missing values; all other confounders are fully observed. We believe the missingness for these two variables may depend on their unobserved values. However, it is plausible that, conditional on the treatment and observed confounders, the missingness is independent of the outcome.

7.4 Results

Table 2 panel (a) presents the estimated effect of the job training program on employment. All three estimators indicate a positive impact: the treatment dummy without control (nonparametric estimate is not applicable with a binary outcome), which is a benchmark result in this randomized experiment, is 0.0370 (SE = 0.0110; 95% CI [0.0154, 0.0586]), the AIJPW estimator is 0.0420 (SE = 0.0090; 95% CI [0.0246, 0.0594]), and the DML estimator yields 0.0450 (SE = 0.0080; 95% CI [0.0294, 0.0606]). In panel (b), the nonparametric estimate is 0.2070 (SE = 0.0720; 95% CI [0.0660, 0.3480]), the AIJPW estimate is 0.2210 (SE = 0.0630; 95% CI [0.0966, 0.3454]), and the DML estimate is 0.2460 (SE = 0.0520; 95% CI [0.1440, 0.3480]). In panel (c), the nonparametric estimate is –0.2830 (SE = 0.0940; 95% CI [–0.4685, –0.0975]), the AIJPW estimate is –0.3410 (SE = 0.0460; 95% CI [–0.4321, –0.2499]) and the DML estimate is –0.3560 (SE = 0.0410; 95% CI [–0.4376, –0.2744]). Across all applications, the more advanced estimators produce larger (in absolute value) effects and yield tighter confidence intervals, suggesting that accounting for high-dimensional confounders and missing confounders can meaningfully improve both precision and inference.

8 Conclusion

We contribute to the causal inference literature by providing a robust and efficient method for settings with MNAR data, thereby enhancing the reliability of treatment effect estimates in the presence of missing confounders. We address the challenges of causal inference in observational studies where confounders are MNAR. We introduced a semiparametric framework that derives the efficiency bound for estimating ATE under the assumption of outcome-independent missingness. Our proposed doubly robust estimator, which solves two Fredholm integral equations—one integrating over confounders and the other over outcomes with an unknown low-rank structure, attains this efficiency bound.

The estimator exhibits robustness in two respects: it remains consistent when either the confounder or the outcome has greater dimensionality, provided other necessary assumptions are satisfied. Additionally, by leveraging fourth-root rate convergence for nuisance parameters, the estimator achieves $\sqrt{n}$-consistency, asymptotic linearity, and local efficiency. Simulation studies demonstrate the estimator's superior finite-sample performance compared to existing methods. An application to the 2015–2016 U.S. National Health and Nutrition Examination Survey illustrates its practical utility in real-world data analysis.

This article also opens several promising avenues for future work. A natural extension is

to derive a Bayesian analog of our doubly robust estimator under the outcome-independent MNAR assumption. One could place priors on the nuisance functions (e.g. the missingness mechanism, conditional densities, regression models) and then incorporate influence-function or bias-correction adjustments in the posterior for the average treatment effect, thereby combining Bayesian uncertainty quantification with semiparametric efficiency and robustness (e.g. as in Breunig et al. (2025)).

Another key direction is to study the finite-sample performance of our estimator beyond the asymptotic regime. Although our theoretical guarantees ensure $\sqrt{n}$-consistency, asymptotic linearity, and local efficiency under fourth-root convergence of nuisance estimates, real-world sample sizes may strain these conditions. Inspired by methods such as those in Armstrong and Kolesár (2021), one could develop bias-aware confidence sets or "honest" inference procedures that maintain nominal coverage even when integral operator inversions approach ill-posedness. Further research is also needed to refine practical estimation of nuisance parameters that involve convolution or integral operators in finite samples.

References

A. Agarwal and R. Singh. Causal inference with corrupted data: Measurement error, missing values, discretization, and differential privacy. *arXiv preprint arXiv:2107.02780*, 2021.

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

D. W. Andrews. Asymptotics for semiparametric econometric models via stochastic equicontinuity. *Econometrica: Journal of the Econometric Society*, pages 43–72, 1994.

D. Arkhangelsky, G. W. Imbens, L. Lei, and X. Luo. Double-robust two-way-fixed-effects regression for panel data. *arXiv preprint arXiv:2107.13737*, 2:12, 2021.

T. B. Armstrong and M. Kolesár. Finite-sample optimal estimation and inference on average treatment effects under unconfoundedness. *Econometrica*, 89(3):1141–1177, 2021.

H. Bang and J. M. Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973, 2005.

J. W. Bartlett, J. R. Carpenter, K. Tilling, and S. Vansteelandt. Improving upon the efficiency of complete case analysis when covariates are mnar. *Biostatistics*, 15(4):719–730, 2014.

S. T. Berry and P. A. Haile. Identification in differentiated products markets using market level data. *Econometrica*, 82(5):1749–1797, 2014.

S. T. Berry and P. A. Haile. Nonparametric identification of differentiated products demand using micro data. *Econometrica*, 92(4):1135–1162, 2024.

P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, 1993.

C. R. Bollinger, B. T. Hirsch, C. M. Hokayem, and J. P. Ziliak. Trouble in the tails? what we know about earnings nonresponse 30 years after lillard, smith, and welch. *Journal of Political Economy*, 127(5):2143–2185, 2019.

K. Borusyak and P. Hull. Optimal formula instruments. Technical report, National Bureau of Economic Research, 2025.

C. Breunig, R. Liu, and Z. Yu. Double robust bayesian inference on average treatment effects. *Econometrica*, 93(2):539–568, 2025.

I. A. Canay, A. Santos, and A. M. Shaikh. On the testability of identification in some nonparametric models with endogeneity. *Econometrica*, 81(6):2535–2559, 2013.

G. Chamberlain. Efficiency bounds for semiparametric regression. *Econometrica*, 60(3): 567–596, 1992.

C. V. Chen, X. and W. K. Newey. Orthogonal machine learning for demand estimation: Theory and application. *Econometrica*, 88(5):1473–1500, 2020.

L. Z. Chen, X. and Y. Zhan. Sieve minimum distance estimation with machine learning generated regressors. *Econometrica*, 91(6):2189–2225, 2023.

X. Chen and T. M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018.

X. Chen and D. Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.

X. Chen and D. Pouzo. Sieve wald and qlr inferences on semi/nonparametric conditional moment models. *Econometrica*, 83(3):1013–1079, 2015.

X. Chen, H. Hong, and D. Nekipelov. Nonlinear models of measurement errors. *Journal of Economic Literature*, 49(4):901–937, 2011.

X. Chen, P. H. Sant'Anna, and H. Xie. Efficient difference-in-differences and event study estimators. *arXiv preprint arXiv:2506.17729*, 2025.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double machine learning for treatment and causal parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

A. Chesher. Identification in nonseparable models with measurement error. *Econometrica*, 71(5):1405–1441, 2003.

Y. Cui, H. Pu, X. Shi, W. Miao, and E. Tchetgen Tchetgen. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024.

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

A. P. Dawid. Conditional independence in statistical theory. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 41(1):1–15, 1979.

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

P. Ding. *A first course in causal inference*. CRC Press, 2024.

R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

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

X. D'Haultfœuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471, 2011.

X. D'Haultfœuille and P. Février. Identification of nonseparable models with measurement errors and instruments. *Quantitative Economics*, 6(2):423–468, 2015.

K. B. Fonseca, J. and H. Xu. Stochastic approximation for generalized deep instrumental variables. *Journal of Econometrics*, 2024. Forthcoming.

A. R. Gallant and D. W. Nychka. Semi-nonparametric maximum likelihood estimation. *Econometrica: Journal of the econometric society*, pages 363–390, 1987.

J. Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

J. Hahn and G. Ridder. Asymptotic variance of semiparametric estimators with generated regressors. *Econometrica*, 81(1):315–340, 2013.

P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics*, 33(6):2904–2929, 2005.

J. J. Heckman. Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, pages 153–161, 1979.

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

D. A. Hirshberg and S. Wager. Title of their work on semiparametric ml inference. *Journal Name*, 2021. (fill in volume, pages).

B. E. Honoré and L. Hu. Sample selection models without exclusion restrictions: Parameter heterogeneity and partial identification. *Journal of Econometrics*, 243(1-2):105360, 2024.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

J. L. Horowitz and C. F. Manski. Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American statistical Association*, 95 (449):77–84, 2000.

Y. Hu. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144(1):27–61, 2008.

Y. Hu and S. M. Schennach. Instrumental variable treatment of nonclassical measurement error models. *Econometrica*, 76(1):195–216, 2008.

Y. Hu and J.-L. Shiu. Nonparametric identification using proxy variables. *Econometrica*, 86(3):759–783, 2018.

Y. Huo, P. Ding, and F. Han. Sensitivity and bias. *Working Paper*, 2025.

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

G. W. Imbens and D. B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

W. Ji, L. Lei, and A. Spector. Model-agnostic covariate-assisted inference on partially identified causal effects. *arXiv preprint arXiv:2310.08115*, 2023.

Z. Jiang, S. Yang, and P. Ding. Multiply robust estimation of causal effects under principal ignorability. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1423–1445, 2022.

E. H. Kennedy. Semiparametric theory and empirical processes in causal inference. *Statistical Science*, 31(4):414–432, 2016.

E. H. Kennedy, S. Balakrishnan, and M. G'sell. Sharp instruments for classifying compliers and generalizing causal effects. 2020.

R. Kress. *Linear integral equations*, volume 82. Springer, 1999.

R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620, 1986.

L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer Series in Statistics. Springer, New York, 2nd edition, 1992.

D. S. Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies*, 76(3):1071–1102, 2009.

E. L. Lehmann and H. Scheffé. Completeness, similar regions, and unbiased estimation: Part i. *Sankhyā: The Indian Journal of Statistics*, 10:305–340, 1950.

B. Lu and R. Ashmead. Propensity score matching analysis for causal effects with mnar covariates. *Statistica Sinica*, 28(4):2005–2025, 2018.

W. Miao, Z. Geng, and E. J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

W. Miao, W. Hu, E. L. Ogburn, and X.-H. Zhou. Identifying effects of multiple treatments in the presence of unmeasured confounding. *Journal of the American Statistical Association*, 118(543):1953–1967, 2023.

F. Molinari. Missing treatments. *Journal of Business & Economic Statistics*, 28(1):82–95, 2010.

W. K. Newey. The asymptotic variance of semiparametric estimators. *Econometrica*, 62 (6):1349–1382, 1994.

W. K. Newey and J. L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003. doi: 10.1111/1468-0262.00459.

W. K. Newey, J. L. Powell, and J. R. Walker. Semiparametric estimation of selection models: some empirical results. *The american economic review*, 80(2):324–328, 1990.

J. Pearl. *Causal diagrams for empirical research*, volume 82. Oxford University Press, 1995.

S. J. Prince. *Understanding deep learning*. MIT press, 2023.

J. M. Robins, A. Rotnitzky, and L. P. Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89 (427):846–866, 1994.

P. M. Robinson. Root-n consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

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

P. R. Rosenbaum and D. B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79 (387):516–524, 1984.

D. B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

P. H. Sant'Anna and J. Zhao. Doubly robust difference-in-differences estimators. *Journal of econometrics*, 219(1):101–122, 2020a.

P. H. C. Sant'Anna and J. B. Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122, 2020b. doi: 10.1016/j.jeconom.2020.06.003.

S. M. Schennach. Recent advances in the measurement error literature. *Annual Review of Economics*, 8:341–377, 2016.

K. S. Singh, R. and T. Zhang. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

A. Sjölander and S. Hägg. Testable implications of outcome-independent missingness not at random in covariates. *Biometrika*, 112(2):asaf009, 2025.

Y. Sun, H. Xie, and Y. Zhang. Difference-in-differences meets synthetic control: Doubly robust identification and estimation. *arXiv preprint arXiv:2503.11375*, 2025.

L. Testa, T. Boschi, F. Chiaromonte, E. H. Kennedy, and M. Reimherr. Doubly-robust functional average treatment effect estimation. *arXiv preprint arXiv:2501.06024*, 2025.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer, 2006.

A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.

A. Velez. On the asymptotic properties of debiased machine learning estimators. *arXiv preprint arXiv:2411.01864*, 2024.

S. Wager. Causal inference: A statistical learning approach, 2024.

D. Williams. *Probability with martingales*. Cambridge university press, 1991.

L. H. Xu, T. and J. Zhu. Deep feature instrumental variable regression. *Journal of Econometrics*, 214(1):35–50, 2020.

S. Yang, L. Wang, and P. Ding. Causal inference with confounders missing not at random. *Biometrika*, 106(4):875–888, 2019.

B. Zhang and E. J. Tchetgen Tchetgen. A semi-parametric approach to model-based sensitivity analysis in observational studies. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185:S668–S691, 2022.

S. Zuo, D. Ghosh, P. Ding, and F. Yang. Mediation analysis with the mediator and outcome missing not at random. *Journal of the American Statistical Association*, pages 1–11, 2024.

S Supplementary material

S.1 Alternative missing data mechanisms

Several assumptions have been proposed to address scenarios involving missing confounders. In this subsection we will discuss these assumptions which do not adequately handle complex missing data patterns, such as those illustrated in Figure 1. For a concrete discussion, see Yang et al. (2019).

To address the missing confounders, Rosenbaum and Rubin (1984) introduced a modified unconfoundedness assumption.

Assumption S1. $\{Y(0), Y(1)\} \perp\!\!\!\perp A \mid (X_r, R)$.

Under Assumption S1, the generalized propensity score $e(X_r, R) = \mathbb{P}(A = 1 \mid X_r, R)$ plays the same role as the classical score $e(X) = P(A = 1 \mid X)$ when confounders are fully observed. Adjusting for $e(X_r, R)$ balances both the observed values and missingness pattern without modeling the missing-data mechanism. However, when $R = \mathbf{0}_p$ indicating that all confounders are missing, it is unlikely that the treatment assignment A is independent of the outcome Y, as no observed confounders are available to adjust for potential biases. In such scenarios, the assumption of no unmeasured confounding may be violated, leading to biased estimates of treatment effects.

By considering missing mechanisms, the causal effects can be identified in different ways. The first mechanism is missing completely at random (Rubin, 1976).

Assumption S2. *(Missing completely at random)* $R \perp\!\!\!\perp (A, X, Y)$.

Assumption S2 requires that the missingness of confounders is independent of all variables (A, X, Y). It implies $\tau = \mathbb{E}[\tau(X) \mid R = 1_p]$ and thus justifies the complete-case analysis that uses only the units with fully observed confounders. However, confounders are rarely missing completely at random and missing at random (Rubin, 1976) deals with that consideration.

Assumption S3. *(Missing at random)* $R \perp\!\!\!\perp X \mid (A, Y)$.

Under Assumption S3, conditioning on the treatment A and outcome Y, the missingness mechanism of the confounders X is independent of the missing values themselves. This assumption implies that the joint distribution factorizes as $f(A, X, Y) = f(A, Y)f(X \mid A, Y, R = 1_p)$. Consequently, the joint distribution and its functionals, including the average treatment effect τ, are identifiable. However, the assumption of missing at random

is untenable when the probability of missingness depends on the unobserved values themselves. In such cases, the data necessitate alternative modeling approaches to account for the non-ignorable missingness mechanism.

S.2 Main proofs for identification

S.2.1 Identification in Yang et al. (2019)

In this subsection, Lemma S1 constructs the integral equation; Lemma S2 recovers the full data distribution; and Theorem 3 estimates the average treatment effect.

Lemma S1. *Under Assumption 3, for any r and a, the following integral equation holds:*

$$f(A=a, X_r, Y, R=r) = \int \xi_{ra}(X) f(A=a, X, Y, R=1_p)\, d\nu(X_{\bar{r}}).$$

Proof. The conclusion follows because the observed data distribution is the complete data distribution averaged over the missing data

$$\begin{aligned} f(A=a, X_r, Y, R=r) &= \int f(A=a, X, Y, R=r) d\nu(X_r) \\ &= \int \frac{f(A=a, X, Y, R=r)}{f(A=a, X, Y, R=1_p)} f(A=a, X, Y, R=1_p) d\nu(X_{\bar{r}}) \\ &= \int \frac{P(R=r \mid A=a, X, Y)}{P(R=1_p \mid A=a, X, Y)} f(A=a, X, Y, R=1_p) d\nu(X_{\bar{r}}) \\ &= \int \xi_{ra}(X) f(A=a, X, Y, R=1_p) d\nu(X_r). \end{aligned}$$

□

Lemma S2. *Under Assumptions 3-5, the distribution of (A, X, Y, R) is identifiable.*

Proof. Suppose that $\xi_{ra}^{(1)}(X)$ and $\xi_{ra}^{(2)}(X)$ are two solutions to the integral equation

$$f(A=a,\, X_r,\, Y,\, R=r) \;=\; \int \xi_{ra}(X)\, f(A=a,\, X,\, Y,\, R=1_p)\, d\nu(X_r) \quad (k=1,2).$$

Then their difference must satisfy

$$\int \left[\xi_{ra}^{(1)}(X) - \xi_{ra}^{(2)}(X)\right] f(A=a,\, X,\, Y,\, R=1_p)\, d\nu(X_r) = 0.$$

Integrating both sides further over X_r, we get

$$\int \left[\xi_{ra}^{(1)}(X) - \xi_{ra}^{(2)}(X)\right] f(A=a,\, X,\, Y,\, R=1_p)\, d\nu(X) = 0.$$

Then by Definition 1 and Assumption 5, we conclude that

$$\xi_{ra}^{(1)}(X) - \xi_{ra}^{(2)}(X) = 0 \quad \text{a.s.}$$

Thus the integral equation (3) admits a unique solution $\xi_{ra}(X)$. Next, from the definition of $\xi_{ra}(X)$, we can identify

$$P(R = r \mid A, X, Y).$$

by $\sum_r P(R = r \mid A, X, Y) = 1$. Finally, we recover

$$f(A, X, Y) = \frac{f(A, X, Y,\ R = 1_p)}{P(R = 1_p \mid A, X, Y)},$$
$$f(A, X, Y, R) = f(R \mid A, X, Y) P(A, X, Y).$$

by equation (1). □

Theorem S1. *Under Assumptions 1, 3-5, the average causal effect τ is identified by*

$$\tau \ = \ \sum_{a=0}^{1} \int \tau(X) \, \frac{f\Big(A = a, X, R = 1_p\Big)}{P\Big(R = 1_p \mid A = a, X\Big)} \, d\nu(X), \tag{S1}$$

Here $\tau(x)$ is identified by (5), $P(A = a,\ R = 1_p)$ and $f(A = a, X, R = 1_p)$ depend only on the observed data, and $P(R = 1_p \mid A = a, X)$ can be identified from Lemma S2 for $a = 0, 1$.

Proof. First, we can identify the conditional distribution of X given $A = a$ by

$$f(X \mid A = a) \ = \ \frac{f\Big(A = a,\ X,\ R = 1_p\Big)}{P\Big(R = 1_p \mid A = a,\ X\Big)}, \quad (a = 0, 1).$$

Averaging $\tau(X)$ over this distribution yields the identification formula (S1). □

S.3 First set of identification formulas

Proof of Theorem 1. We first show that (7) holds. Notice

$$\begin{aligned}
\mathbb{E}\left[\frac{A\mathbb{I}\{R = 1_p\}Y}{P(A = 1, R = 1_p \mid X)}\right] &= \mathbb{E}\left\{\mathbb{E}\left[\frac{A\mathbb{I}\{R = 1_p\}Y}{P(A = 1, R = 1_p \mid X)} \mid A, X\right]\right\} \\
&= \mathbb{E}\left[\frac{A}{P(A = 1, R = 1_p \mid X)} \mathbb{E}\left(\mathbb{I}\{R = 1_p\}Y \mid A, X\right)\right]
\end{aligned}$$

by the law of iterated expectations. The denominator is well defined because of Assumptions 2 and 4. We can further simplify the expectation by

$$\mathbb{E}\left[\frac{A}{P(A=1,R=1_p\mid X)}\mathbb{E}(\mathbb{I}\{R=1_p\}\mid A,X)\,\mathbb{E}\left(Y\mid A,X\right)\right]$$

with the outcome-independent missingness of Assumption 3. This equals to

$$\begin{aligned}\mathbb{E}\left[\frac{AP(R=1_p\mid A,X)}{P(A=1,R=1_p\mid X)}\mathbb{E}\left(Y\mid A,X\right)\right] &= \mathbb{E}\left[\frac{AP(R=1_p\mid A=1,X)}{P(A=1,R=1_p\mid X)}\mathbb{E}\left(Y\mid A=1,X\right)\right]\\ &= \mathbb{E}\left[\frac{A}{P(A=1\mid X)}\mathbb{E}\left(Y\mid A=1,X\right)\right]\\ &= \mathbb{E}[Y(1)]\end{aligned}$$

because the random variable A governs the value of the random variable within the expectation. The third equality follows from Assumption 1, which relies on the classical rationale underlying propensity score weighting. Similarly, we can obtain

$$\mathbb{E}\left[\frac{(1-A)\mathbb{I}\{R=1_p\}Y}{P(A=0,R=1_p\mid X)}\right]=\mathbb{E}\left[Y(0)\right]$$

and thus, we finish the first part of the proof.

Second, we can identify the expectation on the right hand of (7) by

$$\begin{aligned}P(A=a,R=1_p\mid X) &= \frac{f(A=a,X,R=1_p)}{f(X)},\\ &= \frac{f(A=a,X,R=1_p)}{\sum_{a'=0}^{1}\frac{f(A=a',X,R=1_p)}{P(R=1_p\mid A=a',X)}} \quad (a=0,1),\end{aligned}$$

where $f(A=a,X,R=1_p)$ is identifiable from the observed data and $\mathbb{P}(R=1_p\mid A=a,X)$ is identified by the functional ξ_{ra}. Moreover, the indicator function $\mathbb{I}\{R=1_p\}$ ensures the nonnegative part in the expectation with observed density, so the expectation is identified. □

Proof of Lemma 1. The integral equation equals to

$$\begin{aligned} f(X_r, Y \mid A = a, R = r) &= \int f(X, Y \mid A = a, R = r)\, d\nu(X_{\bar{r}}) \\ &= \int f(X \mid A = a, R = r) f(Y \mid A = a, X, R = r)\, d\nu(X_{\bar{r}}) \\ &= \int f(X \mid A = a, R = r) f(Y \mid A = a, X, R = 1_p)\, d\nu(X_{\bar{r}}) \end{aligned}$$

To ensure all of these conditional densities and integrals are well-defined, we require Assumption 4 for:

$$f(\cdot \mid A = a, R = r) = \frac{f(\cdot, R = r \mid A = a)}{P(R = r \mid A = a)}$$

so the denominator is strictly positive. □

Proof of Lemma 2. Suppose that $f_{ra}^{(1)}(X)$ and $f_{ra}^{(2)}(X)$ are two solutions to the integral equation

$$f(X_r, Y \mid A = a, R = r) = \int f_{ra}(X) f(Y \mid A = a, X, R = 1_p)\, d\nu(X_{\bar{r}}),$$

which implies $\int [f_{ra}^{(1)}(X) - f_{ra}^{(2)}(X)] f(Y \mid A = a, X, R = 1_p)\, d\nu(X_{\bar{r}}) = 0$. Integrating both sides with respect to X_r, we obtain

$$\int [f_{ra}^{(1)}(X) - f_{ra}^{(2)}(X)] f(Y \mid A = a, X, R = 1_p)\, d\nu(X) = 0.$$

Under Assumption 5′, completeness implies that $f_{ra}^{(1)}(X) = f_{ra}^{(2)}(X)$ almost surely. Therefore, Equation (10) has a unique solution $f_{ra}(X)$.

For the other part of the theorem, note that the left-hand side of $f(X \mid A = a, R = r) f(A = a, R = r) = f(A = a, X, R = r)$ is identifiable, and so is the right-hand side. Therefore, the conditional distribution $f(R = r \mid X, A = a)$ is identifiable, and consequently, the functional ξ_{ra} is also identified. As a result, the distribution of the full data is identified. □

Proof of Theorem 2. $\tau = \mathbb{E}[\tau(A, X_R, R)] = \mathbb{E}\{\mathbb{E}[\tau(X) \mid A, X_R, R]\}$ is straightforward once we identify $\tau(A, X_R, R)$. We are able to derive the ATE because the random variable $\tau(A, X_R, R)$ is an identified function of observed random variables (A, X_R, R). We focus

on

$$\int \frac{\tau(X)\, f_{ra}(X)}{f(X_r \mid A=a, R=r)}\, d\nu(X_{\bar{r}}) = \int \frac{\tau(X)\, f(X \mid A=a, R=r)}{f(X_r \mid A=a, R=r)}\, d\nu(X_{\bar{r}})$$
$$= \int \tau(X)\, f(X_{\bar{r}} \mid A=a, X_r, R=r)\, d\nu(X_{\bar{r}})$$
$$= \tau(A=a, X_r, R=r).$$

By Assumptions 1 and 3, we have $\tau(X) = \mathbb{E}\big(Y \mid A=1,\, X,\, R=1_p\big) - \mathbb{E}\big(Y \mid A=0,\, X,\, R=1_p\big)$ identified. Moreover, the confounder mask mechanism $f\big(X_r \mid A=a, R=r\big)$ is from the fully observed distribution. By Theorem 1, each $f_{ra}(X)$ is identified, and substituting these into the representation (12) shows that $\tau(A, X_R, R)$ itself is identified. An implicit assumption is that $f(X_r \mid A=a, R=r)$ is well defined. This well-definedness follows immediately from Assumptions 3 and 5. □

S.4 Second set of identification formulas

Proof of Theorem 3. Define the product kernel

$$H_a(Y, \widetilde{Y}) := \int f(\widetilde{Y} \mid A=a, X, R=1_p)\, f(A=a, X, Y, R=1_p)\, d\nu(X).$$

Let g be a bounded measurable function on $\mathcal{Y}$ and suppose $\int g(\widetilde{Y})\, H_a(Y, \widetilde{Y})\, d\nu(\widetilde{Y}) = 0$. Then, by Fubini,

$$0 = \int \Big[\int g(\widetilde{Y})\, f(\widetilde{Y} \mid A=a, X, R=1_p)\, d\nu(\widetilde{Y}) \Big] f(A=a, X, Y, R=1_p)\, d\nu(X) = (\Theta_{a2} u)(Y),$$

where $u(X) := (\Theta_{a1} g)(X) = \int g(\widetilde{Y})\, f(\widetilde{Y} \mid A=a, X, R=1_p)\, d\nu(\widetilde{Y})$ for $X \in S_k$.

Thus $\Theta_{a2} u = 0$ with $u \in \mathrm{Im}(\Theta_{a1})$. By $\mathrm{Im}(\Theta_{a1}) \cap \ker(\Theta_{a2}) = 0$, $u = 0$ almost surely. Hence,

$$\int g(\widetilde{Y})\, f(\widetilde{Y} \mid A=a, X, R=1_p)\, d\nu(\widetilde{Y}) = 0.$$

By Assumption 7 (completeness in X of $f(\widetilde{Y} \mid A=a, X, R=1_p)$), this implies $g = 0$ almost surely. Therefore H_a is complete in Y.

Consequently, the link function $\phi_a(Y)$ solving (13) is unique, and hence $\mathbb{P}(R=1_p \mid A=a, X)$ is identified. □

S.5 Sufficient condition for the low-dimensional structure assumption

Proof of Proposition 1. By the density assumption on $f(Y \mid A=a, X, R=r)$, we observe

$$\begin{aligned}
\mathbb{E}\left[\exp\{-t^\top g(Y)\}\middle| A=a, X, R=r\right] &= \int \exp\{-t^\top g(Y)\} f(Y \mid A=a, X, R=r) d\nu(Y) \\
&= \int \exp\{-t^\top g(Y)\} \exp\{\eta_r(X)^\top g(Y) - \psi\Big(\eta_r(X)\Big)\} h(Y) d\nu(Y) \\
&= \exp\{-\psi\Big(\eta_r(X)\Big)\} \int \exp\{(\eta_r(X)-t)^\top g(Y)\} h(Y) d\nu(Y) \\
&= \exp\{\psi\Big(\eta_r(X)-t\Big) - \psi\Big(\eta_r(X)\Big)\}.
\end{aligned}$$

The fourth equality follows from the fact that the integral of the density equals 1. By assumption, for each r', $\psi\Big(\eta_r(X)-t_{r'r}\Big) - \psi\Big(\eta_r(X)\Big) = (\beta_{r'}-\beta_r)^\top X + \log c_{r'r}$. Hence

$$\mathbb{E}\Big[\exp\{-t_{r'r}^\top g(Y)\} \mid A=a, X, R=r\Big] = c_{r'r}\, \exp\{(\beta_{r'}-\beta_r)^\top X\}.$$

Then

$$\begin{aligned}
\mathbb{E}[\phi_{ra}(Y) \mid A=a, X, R=r] &= \sum_{r'} c_{r'r}^{-1} \mathbb{E}\Big[\exp\{-t_{r'r}^\top g(Y)\} \mid A=a, X, R=r\Big] \\
&= \sum_{r'} \exp\{(\beta_{r'}-\beta_r)^\top X\} \;=\; \frac{1}{\pi_r(X)}.
\end{aligned}$$

□

S.6 Main proofs for the doubly robustness

S.6.1 Proof of Theorem 4

Proof. Conditions (a) and (b) in Assumption 9 ensure the applicability of weak law of large numbers. We omit the formal proof of this result in the main argument. Our goal is the classical doubly robustness of

$$\begin{aligned}
\hat{\tau}_{\text{AIJPW}} = \frac{1}{n}\sum_{i=1}^{n} \Bigg\{ & \hat{\mu}_1(A_i, X_{R_i}, R_i) - \hat{\mu}_0(A_i, X_{R_i}, R_i) \\
& + \frac{A_i\, \mathbb{I}\{R_i = 1_p\}\, [Y_i - \hat{\mu}_1(A_i, X_{R_i}, R_i)]}{\hat{e}_1(X_i)} - \frac{(1-A_i)\, \mathbb{I}\{R_i = 1_p\}\; [Y_i - \hat{\mu}_0(A_i, X_{R_i} R_i)]}{\hat{e}_0(X_i)} \Bigg\}
\end{aligned}$$

(1) We want to show that $\hat{\tau}_{\text{AIJPW}}$ is consistent if $\mathbb{E}\left[|\hat{\mu}_a(A, X_R, R) - \mu_a(A, X_R, R)|^2\right] = o(1)$ holds.

We focus on

$$
\begin{aligned}
S_1 \coloneqq & \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_1(A_i, X_{R_i}, R_i) + \frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \hat{\mu}_1(A_i, X_{R_i}, R_i)]}{\hat{e}_1(X_i)}\right\} - \mu_1 \\
= & \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_1(A_i, X_{R_i}, R_i) - \mathbb{E}\left[\mu(A_i, X_{R_i}, R_i)\right] + \frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \hat{\mu}_1(A_i, X_{R_i}, R_i)]}{\hat{e}_1(X_i)}\right\} \\
= & \frac{1}{n}\sum_{i=1}^{n}\left\{\mu(A_i, X_{R_i}, R_i) - \mathbb{E}\left[\mu(A_i, X_{R_i}, R_i)\right]\right\} + \frac{1}{n}\sum_{i=1}^{n}\frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \mu_1(A_i, X_{R_i}, R_i)]}{\hat{e}_1(X_i)} \\
& + \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(A_i, X_{R_i}, R_i) - \mu(A_i, X_{R_i}, R_i)][1 - \frac{A_i\,\mathbb{I}\{R_i = 1_p\}}{\hat{e}_1(X_i)}] \\
= & I_1 + I_2 + I_3.
\end{aligned}
$$

where

$$
\begin{aligned}
I_1 &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\left\{\mu(A_i, X_{R_i}, R_i) - \mathbb{E}\left[\mu(A_i, X_{R_i}, R_i)\right]\right\}, \\
I_2 &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \mu_1(A_i, X_{R_i}, R_i)]}{\hat{e}_1(X_i)}, \\
I_3 &\coloneqq \frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(A_i, X_{R_i}, R_i) - \mu(A_i, X_{R_i}, R_i)][1 - \frac{A_i\,\mathbb{I}\{R_i = 1_p\}}{\hat{e}_1(X_i)}].
\end{aligned}
$$

And the second equality is by definition and the law of iterated expectation that $\mathbb{E}\left[\mu(A_i = a, X_{R_i}, R_i)\right] = \mathbb{E}[Y(a)]$. Therefore, by triangle inequality, we find that

$$
|S_1| \leq |I_1| + |I_2| + |I_3|.
$$

Since the second moment is bounded, by the law of large numbers we have $|I_1| \xrightarrow{p} 0$. We also find that

$$
\begin{aligned}
|I_2| \leq & \left|\frac{1}{n}\sum_{i=1}^{n}\frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \mu_1(A_i, X_{R_i}, R_i)]}{e_1(X_i)}\right| \\
& + \left|\frac{1}{n}\sum_{i=1}^{n}A_i\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \mu_1(A_i, X_{R_i}, R_i)][\frac{1}{\hat{e}_1(X_i)} - \frac{1}{e_1(X_i)}]\right| \\
\xrightarrow{p} & \left|\frac{1}{n}\sum_{i=1}^{n}A_i\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \mu_1(A_i, X_{R_i}, R_i)][\frac{1}{\hat{e}_1(X_i)} - \frac{1}{e_1(X_i)}]\right|
\end{aligned}
$$

where the first inequality follows from the triangle inequality. The second equality follows

from the law of large numbers and Lemma S11. Now, let's examine

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} A_i \mathbb{I}\{R_i = 1_p\}\left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]\left[\frac{1}{\hat{e}_1(X_i)} - \frac{1}{e_1(X_i)}\right]\right| \\
=&\left|\frac{1}{n}\sum_{i=1}^{n} A_i \mathbb{I}\{R_i = 1_p\}\left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]\left[\frac{1}{\hat{e}_1(X_i)} - \frac{1}{\bar{e}_1(X_i)} + \frac{1}{\bar{e}_1(X_i)} - \frac{1}{e_1(X_i)}\right]\right| \\
\leq&\left|\frac{1}{n}\sum_{i=1}^{n} A_i \mathbb{I}\{R_i = 1_p\}\left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]\left[\frac{1}{\hat{e}_1(X_i)} - \frac{1}{\bar{e}_1(X_i)}\right]\right| \\
&+\left|\frac{1}{n}\sum_{i=1}^{n} A_i \mathbb{I}\{R_i = 1_p\}\left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]\left[\frac{1}{\bar{e}_1(X_i)} - \frac{1}{e_1(X_i)}\right]\right| \\
\xrightarrow{p}&\left|\frac{1}{n}\sum_{i=1}^{n} A_i \mathbb{I}\{R_i = 1_p\}\left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]\left[\frac{1}{\hat{e}_1(X_i)} - \frac{1}{\bar{e}_1(X_i)}\right]\right|
\end{aligned}
$$

where the second inequality follows from the triangle inequality and the last convergence follows from the law of large numbers and Lemma S11. For

$$
\begin{aligned}
&\mathbb{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n} A_i \mathbb{I}\{R_i = 1_p\}\left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]\left[\frac{1}{\hat{e}_1(X_i)} - \frac{1}{\bar{e}_1(X_i)}\right]\right|\right) \\
\leq&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\left|A_i \mathbb{I}\{R_i = 1_p\}\left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]\left[\frac{1}{\hat{e}_1(X_i)} - \frac{1}{\bar{e}_1(X_i)}\right]\right|\right) \\
\leq&\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}\left\{A_i \mathbb{I}\{R_i = 1_p\}\left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]\right\}^2\right)^{\frac{1}{2}}\left\{\mathbb{E}\left[\frac{1}{\hat{e}_1(X_i)} - \frac{1}{\bar{e}_1(X_i)}\right]^2\right\}^{\frac{1}{2}} \\
=&o(1).
\end{aligned}
$$

Therefore, $|I_2| = o_p(1)$ by Markov's inequality. Finally, for I_3, by the Cauchy-Schwarz inequality we obtain

$$
\begin{aligned}
\mathbb{E}(|I_3|) &= \mathbb{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(A_i, X_{R_i}, R_i) - \mu(A_i, X_{R_i}, R_i)]\left[1 - \frac{A_i \mathbb{I}\{R_i = 1_p\}}{\hat{e}_1(X_i)}\right]\right|\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\left|[\hat{\mu}(A_i, X_{R_i}, R_i) - \mu(A_i, X_{R_i}, R_i)]\left[1 - \frac{A_i \mathbb{I}\{R_i = 1_p\}}{\hat{e}_1(X_1)}\right]\right|\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[\hat{\mu}(A_i, X_{R_i}, R_i) - \mu(A_i, X_{R_i}, R_i)]^2\right)^{\frac{1}{2}}\left\{\mathbb{E}\left[1 - \frac{A_i \mathbb{I}\{R_i = 1_p\}}{\hat{e}_1(X_i)}\right]^2\right\}^{\frac{1}{2}} \\
&= o(1).
\end{aligned}
$$

Therefore, we have $|I_3| = o_p(1)$ by Markov's inequality. Put $|I_1|$, $|I_2|$, and $|I_3|$ together. $|S_1| \xrightarrow{p} 0$, which in turn implies that $S_1 \xrightarrow{p} 0$ by the definition of convergence in probability.

Similarly, we can show that

$$S_0 \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{\mu}_0(A_i, X_{R_i}, R_i) + \frac{(1-A_i)\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \hat{\mu}_0(A_i, X_{R_i}, R_i)]}{\hat{e}_0(X_i)}\right\} - \mu_0$$
$$\xrightarrow{p} 0.$$

With $S_1 \xrightarrow{p} 0$ and $S_0 \xrightarrow{p} 0$, we obtain $\hat{\tau}_{\text{AIJPW}} \xrightarrow{p} \tau$.

(2) We want to show that $\hat{\tau}_{\text{AIJPW}}$ is consistent if $\mathbb{E}[|\hat{e}_a(X) - e_a(X)|^2] = o(1)$ and $\mathbb{E}[|\frac{1}{\hat{e}_a(X)} - \frac{1}{e_a(X)}|^2] = o(1)$ holds. Hence, the classical doubly robust property is satisfied.

We write the $\hat{\tau}_{\text{AIJPW}}$ in another format compared to (1):

$$\hat{\tau}_{\text{AIJPW}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,Y_i}{\hat{e}_1(X_i)} - \frac{(1-A_i)\,\mathbb{I}\{R_i = 1_p\}\,Y_i}{\hat{e}_0(X_i)}\right.$$
$$\left. + \hat{\mu}_1(A_i, X_{R_i}, R_i)\left(1 - \frac{A_i\,\mathbb{I}\{R_i = 1_p\}}{\hat{e}_1(X_i)}\right) - \hat{\mu}_1(A_i, X_{R_i}, R_i)\left[1 - \frac{(1-A_i)\,\mathbb{I}\{R_i = 1_p\}}{\hat{e}_0(X_i)}\right]\right\}.$$

We notice

$$\begin{aligned}
S_1' \coloneqq & \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,Y_i}{\hat{e}_1(X_i)} + \hat{\mu}_1(A_i, X_{R_i}, R_i)\left(1 - \frac{A_i\,\mathbb{I}\{R_i = 1_p\}}{\hat{e}_1(X_i)}\right)\right\} - \mu_1 \\
= & \frac{1}{n}\sum_{i=1}^{n}[\frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,Y_i}{e_1(X_i)} - \mu_1] + \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(A_i, X_{R_i}, R_i)\left(1 - \frac{A_i\,\mathbb{I}\{R_i = 1_p\}}{e_1(X_i)}\right) \\
& + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\hat{e}(X_i)} - \frac{1}{e(X_i)}\right)A_i\,\mathbb{I}\{R_i = 1_p\}[Y_i - \hat{\mu}_1(A_i, X_{R_i}, R_i)] \\
= & I_1' + I_2' + I_3'
\end{aligned}$$

where

$$\begin{aligned}
I_1' &\coloneqq \frac{1}{n}\sum_{i=1}^{n}[\frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,Y_i}{e_1(X_i)} - \mu_1], \\
I_2' &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_1(A_i, X_{R_i}, R_i)\left(1 - \frac{A_i\,\mathbb{I}\{R_i = 1_p\}}{e_1(X_i)}\right), \\
I_3' &\coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{\hat{e}(X_i)} - \frac{1}{e(X_i)}\right)A_i\,\mathbb{I}\{R_i = 1_p\}[Y_i - \hat{\mu}_1(A_i, X_{R_i}, R_i)].
\end{aligned}$$

Similarly, we can show that $I_1' = o_p(1)$, $I_2' = o_p(1)$, and $I_3' = o_p(1)$. Moreover, by a similar

way we can show that $S_0' \xrightarrow{p} 0$ where

$$S_0' := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{(1-A_i)\,\mathbb{I}\{R_i=1_p\}\,Y_i}{\hat{e}_0(X_i)} + \hat{\mu}_0(A_i, X_{R_i}, R_i)\left[1 - \frac{(1-A_i)\,\mathbb{I}\{R_i=1_p\}}{\hat{e}_0(X_i)}\right]\right\} - \mu_0.$$

S.6.2 Proof of Theorem 5

Then we show the doubly robustness of

$$\begin{aligned}\hat{\tau}_{\mathrm{DML}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\hat{\mu}_1^{[-k(i)]}\Big(A_i, X_{R_i}, R_i\Big) - \hat{\mu}_0^{[-k(i)]}\Big(A_i, X_{R_i}, R_i\Big) \\ &+ \frac{A_i\,\mathbb{I}\{R_i=\mathbf{1}_p\}\left[Y_i - \hat{\mu}_1^{[-k(i)]}(A_i, X_{R_i}, R_i)\right]}{\hat{e}_1^{[-k(i)]}(X_i)} - \frac{(1-A_i)\,\mathbb{I}\{R_i=\mathbf{1}_p\}\left[Y_i - \hat{\mu}_0^{[-k(i)]}(A_i, X_{R_i}, R_i)\right]}{\hat{e}_0^{[-k(i)]}(X_i)}\Bigg\}. \\ = \frac{1}{n}\sum_{k=1}^{K}|\mathcal{I}_k|\hat{\tau}_{\mathrm{AIJPW}}^{[-k]}.\end{aligned}$$

where $\hat{\tau}_{\mathrm{AIJPW}}^{[-k]}$ is the AIJPW estimator with data $\mathcal{I}_{-k}$. The proof proceeds by treating each fold individually. We present the detailed argument under condition (c.2) only, since the result under (c.1) follows directly by combining the technique used for (c.2) with the proof of Theorem 4. Specifically,

$$\begin{aligned}\hat{\tau}_{\mathrm{AIJPW}}^{[-k]} = \frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k}\Bigg\{&\hat{\mu}_1^{[-k]}\Big(A_i, X_{R_i}, R_i\Big) - \hat{\mu}_0^{[-k]}\Big(A_i, X_{R_i}, R_i\Big) \\ &+ \frac{A_i\,\mathbb{I}\{R_i=\mathbf{1}_p\}\left[Y_i - \hat{\mu}_1^{[-k]}(A_i, X_{R_i}, R_i)\right]}{\hat{e}_1^{[-k]}(X_i)} - \frac{(1-A_i)\,\mathbb{I}\{R_i=\mathbf{1}_p\}\left[Y_i - \hat{\mu}_0^{[-k]}(A_i, X_{R_i}, R_i)\right]}{\hat{e}_0^{[-k]}(X_i)}\Bigg\}.\end{aligned}$$

It is sufficient to show that $\hat{\tau}^{[-k]}_{\text{AIJPW}} \xrightarrow{p} 0$ which is implied by $\hat{\tau}^{[-k]}_{\text{AIJPW}} \xrightarrow{p} 0 \mid \mathcal{I}_{-k}$. We focus on

$$
\begin{aligned}
S_1^{[-k]} :=& \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left\{ \hat{\mu}_1^{[-k]}(A_i, X_{R_i}, R_i) + \frac{A_i \mathbb{I}\{R_i = 1_p\} \left[Y_i - \hat{\mu}_1^{[-k]}(A_i, X_{R_i}, R_i)\right]}{\hat{e}_1^{[-k]}(X_i)} \right\} - \mu_1 \\
=& \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left\{ \hat{\mu}_1^{[-k]}(A_i, X_{R_i}, R_i) - \mathbb{E}\left[\mu(A_i, X_{R_i}, R_i)\right] + \frac{A_i \mathbb{I}\{R_i = 1_p\} \left[Y_i - \hat{\mu}_1^{[-k]}(A_i, X_{R_i}, R_i)\right]}{\hat{e}_1^{[-k]}(X_i)} \right\} \\
=& \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left\{ \mu(A_i, X_{R_i}, R_i) - \mathbb{E}\left[\mu(A_i, X_{R_i}, R_i)\right] \right\} + \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \frac{A_i \mathbb{I}\{R_i = 1_p\} \left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]}{\hat{e}_1^{[-k]}(X_i)} \\
& + \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} [\hat{\mu}^{[-k]}(A_i, X_{R_i}, R_i) - \mu(A_i, X_{R_i}, R_i)][1 - \frac{A_i \mathbb{I}\{R_i = 1_p\}}{\hat{e}_1^{[-k]}(X_i)}] \\
=& I_1^{[-k]} + I_2^{[-k]} + I_3^{[-k]}
\end{aligned}
$$

where

$$
\begin{aligned}
I_1^{[-k]} &:= \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \left\{ \mu(A_i, X_{R_i}, R_i) - \mathbb{E}\left[\mu(A_i, X_{R_i}, R_i)\right] \right\}, \\
I_2^{[-k]} &:= \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} \frac{A_i \mathbb{I}\{R_i = 1_p\} \left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]}{\hat{e}_1^{[-k]}(X_i)}, \\
I_3^{[-k]} &:= \frac{1}{|\mathcal{I}_k|} \sum_{i \in \mathcal{I}_k} [\hat{\mu}^{[-k]}(A_i, X_{R_i}, R_i) - \mu(A_i, X_{R_i}, R_i)][1 - \frac{A_i \mathbb{I}\{R_i = 1_p\}}{\hat{e}_1^{[-k]}(X_i)}].
\end{aligned}
$$

By Assumption 9, we have $\text{Var}(I_1^{[-k]})$ and $\text{Var}(I_2^{[-k]})$ converges to 0. In concreteness,

$$
\begin{aligned}
\text{Var}(I_1^{[-k]}) &= \mathbb{E}[\text{Var}(I_1^{[-k]} \mid \mathcal{I}_{-k})] + \text{Var}(\mathbb{E}[I_1^{[-k]} \mid \mathcal{I}_{-k})] \\
&= \frac{1}{|\mathcal{I}_k|} \text{Var}[\mu(A_i, X_{R_i}, R_i)] = o(1) \\
\text{Var}(I_2^{[-k]}) &= \mathbb{E}[\text{Var}(I_2^{[-k]} \mid \mathcal{I}_{-k})] + \text{Var}(\mathbb{E}[I_2^{[-k]} \mid \mathcal{I}_{-k})] \\
&= \frac{1}{|\mathcal{I}_k|} \text{Var}\left\{ \frac{A_i \mathbb{I}\{R_i = 1_p\} \left[Y_i - \mu_1(A_i, X_{R_i}, R_i)\right]}{\hat{e}_1^{[-k]}(X_i)} \right\} = o(1)
\end{aligned}
$$

since $|\mathcal{I}_k|$ is determined by $\mathcal{I}_{-k}$, Lemma S11, and the random variables are either almost surely bounded or have uniformly bounded second moments. Hence, by Chebyshev's in-

equality and $\mathcal{I}_{-k} \xrightarrow{p} 0$. Finally, for I_3, by the Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}
\mathbb{E}(|I_3| \mid \mathcal{I}_{-k}) &= \mathbb{E}\left(\left|\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}(A_i, X_{R_i}, R_i) - \mu(A_i, X_{R_i}, R_i)][1 - \frac{A_i\,\mathbb{I}\{R_i = 1_p\}}{\hat{e}_1(X_i)}]\right| \mid \mathcal{I}_{-k}\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\left|[\hat{\mu}(A_i, X_{R_i}, R_i) - \mu(A_i, X_{R_i}, R_i)][1 - \frac{A_i\,\mathbb{I}\{R_i = 1_p\}}{\hat{e}_1(X_1)}]\right| \mid \mathcal{I}_{-k}\right) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{E}[\hat{\mu}(A_i, X_{R_i}, R_i) - \mu(A_i, X_{R_i}, R_i)]^2 \mid \mathcal{I}_{-k}\right)^{\frac{1}{2}}\left\{\mathbb{E}\left[1 - \frac{A_i\,\mathbb{I}\{R_i = 1_p\}}{\hat{e}_1(X_i)}\right]^2 \mid \mathcal{I}_{-k}\right\}^{\frac{1}{2}} \\
&= o_p(1).
\end{aligned}$$

Since we have bounded second moments, by Lemma S14, we have $\mathbb{E}(|I_3| \mid \mathcal{I}_{-k})$ uniform integrable

$$\lim_{B\to\infty}\sup_{n}\mathbb{E}\Big[\mathbb{E}(|I_3| \mid \mathcal{I}_{-k})\mathbb{I}\{\mathbb{E}(|I_3| \mid \mathcal{I}_{-k}) > B\}\Big] = 0,$$

where together with $\mathbb{E}(|I_3| \mid \mathcal{I}_{-k}) = o_p(1)$ and Chapter 13 in Williams (1991), we find that $\mathbb{E}(|I_3|) \to o(1)$. Therefore, we have $|I_3| = o_p(1)$ by Markov's inequality. By the same argument as Theorem 4, we obtain $S_1^{[-k]} \xrightarrow{p} 0$. We can also construct $S_0^{[-k]} \xrightarrow{p} 0$, and thus $\hat{\tau}_{\text{DML}} \xrightarrow{p} \tau$. □

S.7 Theoretical foundations of inference

S.7.1 Proof of Theorem 6

Denote $\mathbf{Z} = (A, X, Y, R)$. Recall

$$\begin{aligned}
\hat{\tau}_{\text{AIJPW}} = \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\hat{\mu}_1(A_i, X_{R_i}, R_i) - \hat{\mu}_0(A_i, X_{R_i}, R_i) \\
&+ \frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \hat{\mu}_1(A_i, X_{R_i}, R_i)]}{\hat{e}_1(X_i)} - \frac{(1 - A_i)\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \hat{\mu}_0(A_i, X_{R_i}, R_i)]}{\hat{e}_0(X_i)}\Bigg\},
\end{aligned}$$

and define

$$\begin{aligned}
\tau_{\text{AIJPW}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Bigg\{&\mu_1(A_i, X_{R_i}, R_i) - \mu_0(A_i, X_{R_i}, R_i) \\
&+ \frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \mu_1(A_i, X_{R_i}, R_i)]}{e_1(X_i)} - \frac{(1 - A_i)\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \mu_0(A_i, X_{R_i}, R_i)]}{e_0(X_i)}\Bigg\}.
\end{aligned}$$

For the simplicity of the proof, we also define following notations for decomposing the ATE

$$\tau(\mathbf{Z}) = \mu_1(A, X_R, R) - \mu_0(A, X_R, R), \tag{S2}$$

$$\psi(\mathbf{Z}) = \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_1(A, X_R, R)]}{e_1(X)} - \frac{(1-A)\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_0(A, X_R, R)]}{e_0(X)}, \tag{S3}$$

$$\hat{\tau}(\mathbf{Z}) = \hat{\mu}_1(A, X_R, R) - \hat{\mu}_0(A, X_R, R), \tag{S4}$$

$$\hat{\psi}(\mathbf{Z}) = \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \hat{\mu}_1(A, X_R, R)]}{\hat{e}_1(X)} - \frac{(1-A)\,\mathbb{I}\{R = 1_p\}\,[Y - \hat{\mu}_0(A, X_R, R)]}{\hat{e}_0(X)}. \tag{S5}$$

Lemma S5 establishes that $\hat{\tau}_{\text{AIJPW}}$ is asymptotic linear with influence function

$$\begin{aligned}&\mu_1(A, X_R, R) - \mu_0(A, X_R, R)\\ &+ \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_1(A, X_R, R)]}{e_1(X)} - \frac{(1-A)\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_0(A, X_R, R)]}{e_0(X)}.\end{aligned}$$

Its proof, by the first two following Lemmas, splits the estimation error into two parts. Shortly, we want to prove $\tau_{\text{AIJPW}} - \hat{\tau}_{\text{AIJPW}} = \mathbb{P}_n\left[(\tau - \hat{\tau}) + (\psi - \hat{\psi})\right] = o_p(n^{-1/2})$. Finally, Lemma S6 computes the variance using the corresponding influence function.

Lemma S3. *Under Assumptions 10 and 11,* $(\mathbb{P}_n - \mathbb{P})\left[(\tau - \hat{\tau}) + (\psi - \hat{\psi})\right] = o_p(n^{-1/2})$.

Proof. This empirical process term requires that both $\hat{\mu}_a$ and $\hat{e}_a$ are consistent. Or it might diverge. We can derive that

$$\begin{aligned}(\mathbb{P}_n - \mathbb{P})\left[(\tau - \hat{\tau}) + (\psi - \hat{\psi})\right] &= (\mathbb{P}_n - \mathbb{P})(\tau - \hat{\tau}) + (\mathbb{P}_n - \mathbb{P})(\psi - \hat{\psi})\\ &= (\mathbb{P}_n - \mathbb{P})(\psi - \hat{\psi}) + o_p(n^{-1/2})\end{aligned}$$

For the second equality, the rate holds by Assumption 11(a) and Lemma 2 in Kennedy et al. (2020) or Assumption 11(b), Lemma 19.24 in Van der Vaart (2000). We also notice

$$\begin{aligned}(\mathbb{P}_n - \mathbb{P})(\psi - \hat{\psi}) = (\mathbb{P}_n - \mathbb{P})&\left\{\frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_1(A, X_R, R)]}{e_1(X)} - \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \hat{\mu}_1(A, X_R, R)]}{\hat{e}_1(X)}\right\}\\ &+ (\mathbb{P}_n - \mathbb{P})\left\{\frac{(1-A)\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_1(A, X_R, R)]}{e_1(X)}\right.\\ &\left. - \frac{(1-A)\,\mathbb{I}\{R = 1_p\}\,[Y - \hat{\mu}_1(A, X_R, R)]}{\hat{e}_1(X)}\right\}.\end{aligned}$$

Further, we observe

$$
\begin{aligned}
&(\mathbb{P}_n - \mathbb{P}) \left\{ \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_1(A, X_R, R)]}{e_1(X)} - \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \hat{\mu}_1(A, X_R, R)]}{\hat{e}_1(X)} \right\} \\
&= (\mathbb{P}_n - \mathbb{P}) \left\{ \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_1(A, X_R, R)]}{e_1(X)} - \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_1(A, X_R, R)]}{\hat{e}_1(X)} \right\} \\
&+ (\mathbb{P}_n - \mathbb{P}) \left\{ \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_1(A, X_R, R)]}{\hat{e}_1(X)} - \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \hat{\mu}_1(A, X_R, R)]}{\hat{e}_1(X)} \right\} \\
&= o_P(n^{-1/2}).
\end{aligned}
$$

The first equality is by add and subtract and the consistency of $e_1(X)$. The second equality is by Assumption 11(a) and Lemma 2 in Kennedy et al. (2020) or Assumption 11(b), Lemma 19.24 in Van der Vaart (2000). □

Lemma S4. *Under Assumptions 1–4, 10, and 11,* $\mathbb{P}\left[(\tau - \hat{\tau}) + (\psi - \hat{\psi})\right] = o_p(n^{-1/2})$.

Proof. We have $\mathbb{P}(\psi) = 0$ by Lemma S12, therefore we focus on $\mathbb{P}\left[\tau - \hat{\tau} - \hat{\psi}\right]$. It is equal to

$$
\begin{aligned}
&\mathbb{P}\Big\{ \mu_1(A, X_R, R) - \hat{\mu}_1(A, X_R, R) - \mu_0(A, X_R, R) + \hat{\mu}_0(A, X_R, R) \\
&\quad - \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \hat{\mu}_1(A, X_R, R)]}{\hat{e}_1(X)} + \frac{(1 - A)\,\mathbb{I}\{R = 1_p\}\,[Y - \hat{\mu}_0(A, X_R, R)]}{\hat{e}_0(X)} \Big\}.
\end{aligned}
$$

We concentrate on

$$
\begin{aligned}
&\mathbb{P}\Big\{ \mu_1(A, X_R, R) - \hat{\mu}_1(A, X_R, R) - \frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \hat{\mu}_1(A, X_R, R)]}{\hat{e}_1(X)} \Big\} \\
&= \mathbb{P}\Big\{ \mu_1(A, X_R, R) - \hat{\mu}_1(A, X_R, R) - \frac{A\,\mathbb{I}\{R = 1_p\}\,[\mu_1(A, X_R, R) - \hat{\mu}_1(A, X_R, R)]}{\hat{e}_1(X)} \Big\} \\
&= \mathbb{P}\Big\{ \mu_1(A, X_R, R = 1_p) - \hat{\mu}_1(A, X_R, R = 1_p) - \frac{e_1(X)\,[\mu_1(A, X_R, R = 1_p) - \hat{\mu}_1(A, X_R, R = 1_p)]}{\hat{e}_1(X)} \Big\}
\end{aligned}
$$

The first equality follows from Lemma S11. The second is by conditioning on X and use the law of iterated expectaion by noticing that $\mu_1(A, X_R, R = 1_p)$ and $\hat{\mu}_1(A, X_R, R = 1_p)$ are functions of X.

$$
\begin{aligned}
&\mathbb{P}\left(\frac{[\mu_1(A, X_R, R = 1_p) - \hat{\mu}_1(A, X_R, R = 1_p)][e_1(X) - \hat{e}_1(X)]}{e_1(X)\hat{e}_1(X)} \right), \\
&\lesssim \|\hat{\mu}_a(A, X_R, R = 1_p) - \mu_a(A, X_R, R = 1_p)\|_2 \cdot \|\hat{e}_1(X) - e_1(X)\|_2 = o_P(n^{-1/2})
\end{aligned}
$$

where the inequality is followed by the Cauchy–Schwarz inequality, and the last equality is

by Assumption 10. Specifically, we can show that

$$\|\hat{\mu}_a(A, X_R, R = \mathbf{1}_p) - \mu_a(A, X_R, R = \mathbf{1}_p)\|_2 \;\leq\; \frac{\|\hat{\mu}_a(A, X_R, R) - \mu_a(A, X_R, R)\|_2}{\sqrt{\mathbb{P}(R = \mathbf{1}_p)}}.$$

Also,

$$\mathbb{P}\Big\{\mu_0(A, X_R, R) - \hat{\mu}_0(A, X_R, R) - \frac{(1-A)\,\mathbb{I}\{R = 1_p\}\,[Y - \hat{\mu}_0(A, X_R, R)]}{\hat{e}_0(X)}\Big\} = o_p(n^{-1/2})$$

by a similar argument. Therefore, $\mathbb{P}\left[(\tau - \hat{\tau}) + (\psi - \hat{\psi})\right] = o_p(n^{-1/2})$. □

Lemma S5. *Under Assumptions 1–4, 10, and 11,* $\hat{\tau}_{AIJPW} = \tau_{AIJPW} + o_p(n^{-1/2})$.

Proof. The Lemma holds because

$$\begin{aligned}\tau_{\text{AIJPW}} - \hat{\tau}_{\text{AIJPW}} &= (\mathbb{P}_n - \mathbb{P})\left[(\tau - \hat{\tau}) + (\psi - \hat{\psi})\right] + \mathbb{P}\left[(\tau - \hat{\tau}) + (\psi - \hat{\psi})\right] \\ &= o_p(n^{-1/2})\end{aligned}$$

by Lemma S3 and S4. □

Lemma S6. *Under Assumptions 1–4, 10, and 11,* $Var[\tau(\mathbf{Z}) + \psi(\mathbf{Z})] = Var[\tau(A, X_R, R)] + \mathbb{E}\left[\frac{\sigma_1^2(X)}{e_1(X)}\right] + \mathbb{E}\left[\frac{\sigma_0^2(X)}{e_0(X)}\right]$.

Proof. We simplify the formula $V = \text{Var}[\tau(\mathbf{Z})] + \text{Var}[\psi(\mathbf{Z})]$ and we notice

$$\begin{aligned}\text{Var}[\psi(\mathbf{Z})] = \mathbb{E}&\left(\frac{A\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_1(A, X_R, R)]}{e_1(X)}\right)^2 \\ &+ \mathbb{E}\left(\frac{(1-A)\,\mathbb{I}\{R = 1_p\}\,[Y - \mu_0(A, X_R, R)]}{e_0(X)}\right)^2 \qquad \text{(S6)}\end{aligned}$$

which includes two parts, and for the first one

$$
\begin{aligned}
&\mathbb{E}\left(\frac{A\,\mathbb{I}\{R=1_p\}\,[Y-\mathbb{E}(Y\mid A=1,X,R=1_p)]}{e_1(X)}\right)^2\\
&=\mathbb{E}\left(\frac{A\,\mathbb{I}\{R=1_p\}\,[Y-\mathbb{E}(Y\mid A=1,X)]^2}{e_1(X)^2}\right)\\
&=\mathbb{E}\left\{\mathbb{E}\left(\frac{\mathbb{I}\{R=1_p\}\,[Y-\mathbb{E}(Y\mid A=1,X)]^2}{e_1(X)^2}\mid A=1,X\right)\mathbb{P}(A=1\mid X)\right\}\\
&=\mathbb{E}\left\{\frac{\mathbb{P}(R=1_p\mid A=1,X)}{e_1^2(X)}\mathbb{E}\left([Y-\mathbb{E}(Y\mid A=1,X)]^2\mid A=1,X\right)\mathbb{P}(A=1\mid X)\right\}\\
&=\mathbb{E}\left\{\frac{\mathbb{E}\left([Y-\mathbb{E}(Y\mid A=1,X)]^2\mid A=1,X\right)}{e_1(X)}\right\}\\
&=\mathbb{E}\left\{\frac{\mathrm{Var}\left[Y(1)\mid X\right]}{e_1(X)}\right\}
\end{aligned}
$$

where the first equality holds by definition and Assumption 3. The second equality uses law of iterated expectation, and the third equality is from Assumption 3. We can also yield that the second term in (S6) equals $\mathbb{E}\left\{\frac{\mathrm{Var}[Y(0)|X]}{e_0(X)}\right\}$.

Therefore the doubly robust estimator $\hat{\tau}_{\mathrm{AIJPW}}$ is asymptotically linear and by the central limited theorem,

$$
\begin{aligned}
&\sqrt{n}\left(\hat{\tau}_{\mathrm{AIJPW}}-\tau\right)\rightsquigarrow\mathcal{N}(0,V)\\
&V=\mathrm{Var}\left[\tau(A,X_R,R)\right]+\mathbb{E}\left[\frac{\sigma_1^2(X)}{e_1(X)}\right]+\mathbb{E}\left[\frac{\sigma_0^2(X)}{e_0(X)}\right]
\end{aligned}
$$

where $\sigma_a^2(X)=\mathrm{Var}\left[Y_i(a)\mid X\right]$ for $a=0,1$. □

S.8 Discussions on completeness

S.8.1 Sufficient conditions for completeness

Our work extends the sufficient conditions to both integral equations, whereas Yang et al. (2019) only address the first. Remember that a function $f(X,Y)$ is said to be complete in Y if, for any square-integrable function $g(X)$, the condition $\int g(X)f(X,Y)\,d\nu(X)=0$ implies that $g(X)=0$ almost surely. We provide two sufficient conditions for completeness. The first condition applies to discrete support and requires a full-rank condition on the observed transition matrix. The second condition applies to continuous support within the exponential family. We then connect the conditions to the integral equation $\xi(X)$ and the integral equation $f(X)$, which together yield the identification formulas.

Lemma S7. *Suppose that X and Y are discrete, $X_j \in \{x_{j1}, \ldots, x_{jJ_j}\}$ and $Y \in \{y_1, \ldots, y_K\}$. Let $q = J_1 \times \cdots \times J_p$, and let Θ by a $K \times q$ matrix with k-th row $\Theta_k = f(X, y_k)$ evaluated at all possible values of X. Then $g(X) = 0$ if Rank$(\Theta) = q$.*

Proof. Denote $\text{vec}(X) = (x_1, \ldots, x_q)^\top$ where the column vector listing all possible combinations of the entries of X. Then it is easy to rewrite everything into matrix formats. The full column-rank to trivial-nullspace result follows directly from the Rank–Nullity Theorem (Horn and Johnson, 2012). □

Lemma S8. *Suppose $f(X, Y) = \alpha(X)h(Y)\exp\{\lambda(Y)^\top \eta(X)\}$, then it is complete in Y if (i) $\alpha(X) > 0$, (ii) $\lambda(Y) > 0$ for every $Y \in \mathcal{B}$ where $\mathcal{B}$ is an open set, and (iii) the mapping $\eta(X)$ is one-to-one.*

Proof. See Yang et al. (2019). □

Remark 9. *In the Gaussian model*

$$f(X, Y) = f(Y \mid X)\, f(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big[-\tfrac{(Y-\beta_0-\beta_1^\top X)^2}{2\sigma^2}\Big]\, f(X), \quad \beta_1 \in \mathbb{R}^p,\ X \in \mathbb{R}^p,$$

is complete in Y. Identify,

$$\alpha(X) = \frac{1}{\sqrt{2\pi\sigma^2}}\, f(X), \quad \lambda(Y) = \sigma^{-2} Y \beta_1, \quad \eta(X) = X.$$

Clearly $\alpha(X) > 0$, $\lambda(Y) > 0$ on any open set $\mathcal{B}$. All conditions of Lemma S8 hold, and $f(X, Y)$ is complete in Y.

Remark 10. *We consider a sufficient completeness condition for our identifications. For the joint distribution $f(A = a, X, Y, R = 1_p)$ in the identification formula, we assume completeness in Y for $a = 0, 1$. That is, we can define $f_a(X, Y) = f(A = a, X, Y, R = 1_p)$. Suppose that X and Y are discrete, with $X_j \in \{x_{j1}, \ldots, x_{jJ_j}\}$ and $Y \in \{y_1, \ldots, y_K\}$. Let $q = J_1 \times \cdots \times J_p$, and let Θ_a be a $K \times q$ matrix where the k-th row is $\Theta_{ak} = f_a(X, y_k)$ evaluated at all possible values of X. Then, by Lemma S7, Rank$(\Theta_a) = q$. A similar argument holds for the conditional marginal distribution $f(Y \mid A = a, X, R = 1_p)$, which is complete in X or Y for $a = 0, 1$.*

S.8.2 Equivalence of Assumptions 5 and 5′

The two completeness assumptions are equivalent if $f(A = a, X, R = 1_p) > 0$ for any a and this equivalence is particularly clear in the illustrative case of discrete confounders. For

example, when $p = 1$, equation (3) becomes

$$f\{A = a, Y, R = 0\} = \sum_{i=1}^{J} \frac{\mathbb{P}\{R = 0 \mid X_i, A = a\}}{\mathbb{P}\{R = 1 \mid X_i, A = a\}} \times f\{A = a, X_i, Y, R = 1\}, \quad a \in \{0, 1\}.$$

with the matrix format

$$\begin{pmatrix} f\{A = a, Y = y_1, R = 0\} \\ \vdots \\ f\{A = a, Y = y_K, R = 0\} \end{pmatrix}_{K \times 1} = \Theta_a \begin{pmatrix} \xi_{0,a}(x_1) \\ \vdots \\ \xi_{0,a}(x_J) \end{pmatrix}_{J \times 1}.$$

where

$$\Theta_a = \begin{pmatrix} f\{A = a, X_1, Y = y_1, R = 1\} & \cdots & f\{A = a, X_J, Y = y_1, R = 1\} \\ \vdots & & \vdots \\ f\{A = a, X_1, Y = y_K, R = 1\} & \cdots & f\{A = a, X_J, Y = y_K, R = 1\} \end{pmatrix}_{K \times J}.$$

In the alternative representation, equation (10) becomes

$$f\{Y \mid A = a, R = 0\} = \sum_{i=1}^{J} f\{X_i \mid A = a, R = 0\} \times f\{Y \mid A = a, X = x_i, R = 1\}, \quad a \in \{0, 1\}.$$

with the matrix format

$$\begin{pmatrix} f\{Y = y_1 \mid A = a, R = 0\} \\ \vdots \\ f\{Y = y_K \mid A = a, R = 0)\} \end{pmatrix}_{K \times 1} = \Gamma_a \begin{pmatrix} f_{0,a}(X_1) \\ \vdots \\ f_{0,a}(X_J) \end{pmatrix}_{J \times 1}.$$

where

$$\Gamma_a = \begin{pmatrix} f\{y_1 \mid A = a, X_1, R = 1\} & \cdots & f\{y_1 \mid A = a, X_J, R = 1\} \\ \vdots & & \vdots \\ f\{y_K \mid A = a, X_1, R = 1\} & \cdots & f\{y_K \mid A = a, X_J, R = 1)\} \end{pmatrix}_{K \times J}.$$

The full column ranks of Θ_a and Γ_a (the discrete analogue of the completeness assump-

tion) coincide. Indeed, writing $\Theta_a = D_a\,\Gamma_a$, where

$$D_a = \begin{pmatrix} f\{A=a, X_1, R=1\} & 0 & \cdots & 0 \\ 0 & f\{A=a, X_2, R=1\} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f\{A=a, X_J, R=1\} \end{pmatrix}_{J\times J},$$

the positivity condition $f\{A=a, x_j, R=1\} > 0$ for all j implies that D_a is invertible. Therefore $\operatorname{rank}(\Theta_a) = \operatorname{rank}(D_a\Gamma_a) = \operatorname{rank}(\Gamma_a)$, so Θ_a and Γ_a have identical full column rank.

For the general proof, specifically, we first show that the completeness of $f(A = a, X, Y, R = 1_p)$ in Y implies the completeness of $f(Y \mid A = a, X, R = 1_p)$ in Y. For any square-integrable function $g(X)$, such that $\int g(X) f(Y \mid A = a, X, R = 1_p)\, d\nu(X) = 0$, we have $\int \frac{g(X)}{f(A=a,X,R=1_p)} f(Y, A = a, X, R = 1_p)\, d\nu(X) = 0$. Because $f(Y, A = a, X, R = 1_p)$ is complete in Y, we then get $\frac{g(X)}{f(A=a,X,R=1_p)} = 0$. And thus $g(X) = 0$ by $f(A = a, X, R = 1_p) > 0$. Therefore, $f(Y \mid A = a, X, R = 1_p)$ is complete in Y.

Conversely, we assume the completeness of $f(Y \mid A = a, X, R = 1_p)$ in Y. From $\int g(X) f(A = a, X, Y, R = 1_p) d\nu(X) = 0$, we observe that $\int g(X) f(A = a, X, R = 1_p) \frac{f(A=a,X,Y,R=1_p)}{f(A=a,X,R=1_p)} d\nu(X) = \int g(X) f(A = a, X, R = 1_p) f(Y \mid A = a, X, R = 1_p) d\nu(X) = 0$. Since $f(Y \mid A = a, X, R = 1_p)$ is complete in Y, $g(X) f(A = a, X, R = 1_p) = 0$. Hence, $f(A = a, X, Y, R = 1_p)$ is complete in Y.

S.8.3 Trivial intersection of $\operatorname{Im}(\Theta_{a1})$ and $\ker(\Theta_{a2})$

For any $a \in \{0, 1\}$ we have the property $\operatorname{Im}(\Theta_{a1}) \cap \ker(\Theta_{a2}) = \{0\}$, where recall

$$(\Theta_{a1} g)(X) = \int g(Y)\, f(Y \mid A = a, X, R = 1_p)\, d\nu(Y),$$
$$(\Theta_{a2} u)(Y) = \int u(X)\, f(A = a, X, Y, R = 1_p)\, d\nu(X).$$

Proof. This result follows from the Rank–Nullity Theorem since the kernel of Θ_{a2} coincides with that of the adjoint estimator of Θ_{a1}. However, we opt to give a direct proof here rather than appealing to the theorem's general statement.
(1) We want to show that $\operatorname{Im}(\tilde{\Theta}_{a1}) \cap \ker(\Theta_{a2}) = \{0\}$ where

$$(\tilde{\Theta}_{a1} g)(X) \coloneqq \int g(Y) f(A = a, X, Y, R = 1_p) d\nu(Y).$$

We define the inner product $\langle \gamma, \eta \rangle_X \coloneqq \int \gamma(X)\eta(X)d\nu(X)$. Then, we find that

$$\begin{aligned}\langle \tilde{\Theta}_{a1} g, u \rangle_X &= \int \left[\int g(Y) f(A=a, X, Y, R=1_p) d\nu(Y) \right] u(X) d\nu(X) \\ &= \int \left[\int u(X) f(A=a, X, Y, R=1_p) d\nu(X) \right] g(Y) d\nu(Y) \\ &= \langle g, \Theta_{a2} u \rangle_Y\end{aligned}$$

by Fubini's theorem. Assume that there exists $\check{u}(X)$ and $\check{g}(Y)$ such that $\check{u}(X) = (\tilde{\Theta}_{a1}\check{g})(X)$ and $(\Theta_{a2}\check{u})(Y) = 0$. We notice

$$\langle \tilde{\Theta}_{a1}\check{g}, \tilde{\Theta}_{a1}\check{g} \rangle_X = \langle \tilde{\Theta}_{a1}\check{g}, \check{u} \rangle_X = \langle \check{g}, \Theta_{a2}\check{u} \rangle_Y = 0$$

From $\langle \tilde{\Theta}_{a1}\check{g}, \tilde{\Theta}_{a1}\check{g} \rangle_X = 0$, we get $\check{u}(X) = (\tilde{\Theta}_{a1}\check{g})(X) = 0$ by the basic property in the functional analysis. Thus, $\mathrm{Im}(\Theta_{a1}) \cap \ker(\Theta_{a2}) = \{0\}$.

(2) We want to show that $\mathrm{Im}(\Theta_{a1}) \cap \ker(\Theta_{a2}) = \{0\}$. For any $\check{u}(X)$ and $\check{g}(Y)$ such that

$$\check{u}(X) = (\Theta_{a1}\check{g})(X) \quad \text{and} \quad (\Theta_{a2}\check{u})(Y) = 0,$$

we have $\check{u}(X) = (\tilde{\Theta}_{a1}\tilde{g})(X)$, where $\tilde{g}(Y)$ is defined by

$$\tilde{g}(Y) f(A=a, X, R=1_p) = \check{g}(Y).$$

Since $\mathrm{Im}(\tilde{\Theta}_{a1}) \cap \ker(\Theta_{a2}) = \{0\}$ from (1), it follows that $\check{u}(X) = 0$. Therefore, $\mathrm{Im}(\Theta_{a1}) \cap \ker(\Theta_{a2}) = \{0\}$. □

S.9 Illustrative examples for two sets of identification formulas

We present a few examples to illustrate that the two sets of identification formulas apply in different scenarios. The first example demonstrates that the initial set of formulas is valid when the dimension of Y exceeds that of X. By contrast, the second example emphasizes the case in which the dimension of Y is lower than that of X. We observe (A, X_R, R, Y) and focus on the treated arm $A = 1$. All probabilities and expectations below are conditional on $A = 1$. Also for simplicity, we assume that R is a dummy variable.

S.9.1 Completeness in Y

The completeness in Y is quite straight forward. As an illustration, let Y have three categories and X have two categories. The top panel reports the (full-data) joint cell probabilities of (Y, X) within each R stratum, conditional on $A = 1$. The bottom panel shows the observed-data distribution. When $R = 0$, only one column of X is observed, whereas for $R = 1$ both columns are observed.

$R = 0$

$Y \backslash X$	0	1
0	1/8	0
1	1/8	1/8
2	0	1/8

$R = 1$

$Y \backslash X$	0	1
0	1/12	0
1	1/12	1/12
2	1/12	1/6

Notes. Cell probabilities are conditional on $A = 1$. Each block sums to $1/2$, so $P(R = 0 \mid A = 1) = P(R = 1 \mid A = 1) = 1/2$.

$R = 0$

$Y \backslash X$	N.A.
0	1/8
1	1/4
2	1/8

$R = 1$

$Y \backslash X$	0	1
0	1/12	0
1	1/12	1/12
2	1/12	1/6

Notes. We observe only the cells revealed under each R pattern: for $R = 0$ a single X column is observed; for $R = 1$ both X columns are observed. Totals again equal $1/2$ within each R block.

By the identification procedure in Section 3.2.2 and the observed data, Bayes' rule yields $\mathbb{P}(Y \mid A = 1, R = 0) = (1/4,\ 1/2,\ 1/4)$. Let

$$\Theta = \begin{pmatrix} 1/3 & 0 \\ 1/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix},$$

where Θ denotes $\mathbb{P}(Y \mid A = 1, X, R = \mathbf{1}_p)$ (see S.8.1). Under outcome-independent MNAR $Y \perp R \mid A, X$, equation (10) implies

$$\mathbb{P}(Y \mid A = 1, R = r) = \Theta\, \mathbb{P}(X \mid A = 1, R = r), \qquad r \in \{0, 1\}.$$

Solving this system gives $\mathbb{P}(X \mid A = 1, R = r) = (1/2,\ 1/2)^\top$ for $r = 0, 1$.

S.9.2 Completeness in X

The completeness condition in X is rather abstract. To provide intuition, we present two examples that illustrate the low-dimensional structure embodied in (13). We now consider the case where Y takes two values and X takes three values. For the first example. Take $j \in \{0, 1, 2\}$ with $\mathbb{P}(X = j) = 1/3$, let

$$\mathbb{P}(Y \mid X) = \begin{pmatrix} 1/4 & 1/2 & 3/4 \\ 3/4 & 1/2 & 1/4 \end{pmatrix} \quad \text{(rows } Y = 0, 1\text{; columns } X = 0, 1, 2\text{)},$$

and set $\mathbb{P}(R = 1 \mid A = 1, X) = (1/2, 1/2, 1/2)$. This construction enforces $Y \perp R \mid A, X$.

$R = 0$

$Y \backslash X$	0	1	2
0	1/24	1/12	1/8
1	1/8	1/12	1/24

$R = 1$

$Y \backslash X$	0	1	2
0	1/24	1/12	1/8
1	1/8	1/12	1/24

Notes. Each block sums to 1/2 (conditional on $A = 1$). Because $P(R = 1 \mid A = 1, X)$ is constant, $P(Y \mid X, R = 0) = P(Y \mid X, R = 1) = P(Y \mid X)$, so the two panels are identical column-wise.

$R = 0$

$Y \backslash X$	N.A.
0	1/4
1	1/4

$R = 1$

$Y \backslash X$	0	1	2
0	1/24	1/12	1/8
1	1/8	1/12	1/24

Notes. For $R = 0$, X is unobserved but $P(Y, A = 1, R = 0) = (1/4, 1/4)$ is observed. For $R = 1$, X is fully observed. Totals again equal 1/2 within each R block.

From the observed sample with $R = 1$, we summarize the conditional and joint distributions as follows. The first matrix,

$$\Theta_1 = \mathbb{P}(Y \mid A = 1, X, R = 1) = \begin{pmatrix} 1/4 & 1/2 & 3/4 \\ 3/4 & 1/2 & 1/4 \end{pmatrix},$$

captures the conditional probabilities of Y given treatment $A = 1$, confounders X, and being observed ($R = 1$). Each row corresponds to a value of X, and each column represents a value of Y.

The second matrix,

$$\Theta_2 = \mathbb{P}(X, Y, R = 1) = \begin{pmatrix} 1/24 & 1/12 & 1/8 \\ 1/8 & 1/12 & 1/24 \end{pmatrix},$$

represents the joint distribution of (X, Y) among the observed units. This matrix encodes both the marginal variation in X and the dependence structure between X and Y within the $R = 1$ subpopulation.

Meanwhile, the unobserved value $\frac{1}{\mathbb{P}(R=1|A=1,X)} = (2, 2, 2)$, and we have $\phi = (2, 2)$ such that

$$\phi \begin{pmatrix} 1/4 & 1/2 & 3/4 \\ 3/4 & 1/2 & 1/4 \end{pmatrix} = \begin{pmatrix} 2 & 2 & 2 \end{pmatrix}.$$

Thus, Assumption 6 holds. To identify the ATE, note that based on observed values we have

$$\phi\Theta_1\Theta_2^\top = \phi \begin{pmatrix} 7/48 & 5/48 \\ 5/48 & 7/48 \end{pmatrix} = \mathbb{P}(Y) = \begin{pmatrix} 1/2 & 1/2 \end{pmatrix}.$$

Hence, ϕ can be uniquely determined. For the second example, we take $X \in \{0, 1, 2\}$ with $\mathbb{P}(X = j) = 1/3$. Let

$$\mathbb{P}(Y \mid A = 1, X) = \begin{pmatrix} 1/4 & 1/4 & 3/4 \\ 3/4 & 3/4 & 1/4 \end{pmatrix} \quad \text{(rows } Y = 0, 1\text{; columns } X = 0, 1, 2\text{)},$$

and set $\mathbb{P}(R = 1 \mid A = 1, X) = (1/3, 1/3, 1/2)$. Similarly, we can get $\phi = (7/12, 1/4)$. In the same manner, we are also able to identify the underlying distribution. These two examples highlight two key intuitions. Across strata, either the variation in the missing patterns is identical, or the missing patterns coincide with the outcome distribution.

S.10 Nonparametric (machine learning) two-stage least squares

S.10.1 Extra notations

Let $\mathcal{H}_J^p = \{ h_j(W) = \exp(-W^\top W) W^{\lambda_j} : j = 1, \ldots, J \}$ denote the class of Hermite-type basis functions, where $W^{\lambda_j} = W_1^{\lambda j_1} \cdots W_p^{\lambda_{j_p}}$, $\lambda_j = (\lambda_{j_1}, \ldots, \lambda_{j_p})$ is a multi-index, and $|\lambda_j| = \sum_{l=1}^p \lambda_{jl}$, with $|\lambda_j|$ non-decreasing in j. Define the standardized variable $\bar{W} = \Sigma^{-1/2}(W - \mu)$, where μ and Σ are fixed mean and covariance.

S.10.2 Compactness and computational complexity

Although equations (10) and (14) are point-identified under the proper assumption, estimating $f_{ra}(X)$ and $\phi_a(Y)$ is difficult because it entails solving a Fredholm integral equation of the first kind which is an ill-posed inverse problem.[6] Meanwhile, the associated linear operator is compact and its inverse is unbounded, so even small sampling error in the estimated densities $f(A = a, X, Y, R = 1_p)$ and $f(Y \mid A = a, X, R = 1_p)$ can be greatly amplified, producing substantial bias in the plug-in estimators $\hat{f}_{ra}(X)$ and $\hat{\phi}_a(Y)$. See Newey and Powell (2003) for a detailed exposition of this ill-posedness in nonparametric IV models and Kress (1999) for perturbation theory.

Before discussing compactness, for computational simplicity, we assume the following Hermite expansions,

$$f_{ra}(X) \approx \sum_{j=1}^{J_X} \theta_{ra}^j h_j(\bar{X}) \quad \text{and} \quad \phi_a(Y) \approx \sum_{j=1}^{J_Y} \gamma_a^j h_j(\bar{Y}),$$

where the standardization stabilizes the coefficients. Therefore,

$$\begin{aligned} f(X_r, Y \mid A = a, R = r) &= \int f_{ra}(X) f(Y \mid A = a, X, R = 1_p)\, d\nu(X_{\bar{r}}) \\ &= \sum_{j=1}^{J_X} \theta_{ra}^j \int h_j(\bar{X}) f(Y \mid A = a, X, R = 1_p)\, d\nu(X_{\bar{r}}) \\ &= \sum_{j=1}^{J_X} \theta_{ra}^j H_{ra}^j(X_r, Y) \end{aligned} \tag{S7}$$

by the second identification equation (10) where the conditional expectation $H_{ra}^j(X_r, Y) \coloneqq \int h_j(\bar{X}) f(Y \mid A = a, X, R = 1_p)\, d\nu(X_{\bar{r}})$. And

$$\begin{aligned} f(A = a, Y) &= \int \phi_a(\widetilde{Y}) \left[\int f(\widetilde{Y} \mid A = a, X, R = 1_p)\, f(A = a, X, Y, R = 1_p)\, d\nu(X) \right] d\nu(\widetilde{Y}) \\ &= \sum_{j=1}^{J_Y} \gamma_a^j \int h_j(\bar{\widetilde{Y}}) \left[\int f(\widetilde{Y} \mid A = a, X, R = 1_p) f(A = a, X, Y, R = 1_p) d\nu(X) \right] d\nu(\widetilde{Y}) \\ &= \sum_{j=1}^{J_Y} \gamma_a^j H_a^j(Y) \end{aligned} \tag{S8}$$

where the conditional expectation $H_a^j(Y) \coloneqq \int f(\widetilde{Y} \mid A = a, X, R = 1_p) f(A = a, X, Y, R = 1_p) d\nu(Y)$.

[6] See the estimation of (10) in Yang et al. (2019).

First of all, let $D^\lambda g(W) = \frac{\partial^\lambda g(W)}{\partial W_1^{\lambda_1} \cdots \partial W_p^{\lambda_p}}$ where $\lambda = (\lambda_1, \ldots, \lambda_p)$. In particular, $D^{\mathbf{0}} g(W) = g(W)$. For a vector valued function $\mathbf{H}(W) = \{h_1(W), \ldots, h_J(W)\}^\top$, also define $D^\lambda \mathbf{H}(W) = \left\{D^\lambda h_1(W), \ldots, D^\lambda h_J(W)\right\}^\top$. For parameters $m > 0$, m_0, $\delta_0 > p/2$, and $\delta \in (p/2, \delta_0)$, consider the function space

$$\mathcal{G}_{m,m_0,\delta_0,B} = \left\{ g(W) : \sum_{|\lambda| \le m+m_0} \int \left\{D^\lambda g(\bar{W})\right\}^2 (1 + \bar{W}^\top \bar{W})^{\delta_0} dx \le B \right\}$$

and $\bar{W}$ is the standardized version of W. Define the norm $\|g\|_{\mathcal{G}} = \max_{|\lambda| \le m} \sup_W \left| D^\lambda g(\bar{W}) \right| (1 + \bar{W}^\top \bar{W})^\delta$. Gallant and Nychka (1987) show that the closure of $\mathcal{G}_{m,m_0,\delta_0,B}$ with respect to the norm $\|g\|_{\mathcal{G}}$ is compact.

Assumption S4. *Assume that the functions $f_{ra}(X)$ as well as their estimators $f_{ra}(X)$ all belong to the function class $\mathcal{G}_{m_X,m_{X0},\delta_{X0},B_X}$ for any r and a.*

Assumption S5. *Assume that the functions $\phi_a(Y)$ as well as their estimators $\hat{\phi}_a(Y)$ all belong to the function class $\mathcal{G}_{m_Y,m_{Y0},\delta_{Y0},B_Y}$ for any a.*

Given the Hermite approximations, the regularization in Assumption S4 translates to

$$\theta_{ra}^\top \left[\sum_{|\lambda| \le m_X + m_{X0}} \int \left\{D^\lambda \mathbf{H}(\bar{X})\right\} \left\{D^\lambda \mathbf{H}(\bar{X})\right\}^\top (1 + \bar{X}^\top \bar{X})^{\delta_{X0}} d\nu(X) \right] \theta_{ra} \le B_X,$$

where $\theta_{ra} = (\theta_{ra}^1, \ldots, \theta_{ra}^{J_X})^\top$. The regularization in Assumption S5 translates to

$$\gamma_a^\top \left[\sum_{|\lambda| \le m_Y + m_{Y0}} \int \left\{D^\lambda \mathbf{H}(\bar{Y})\right\} \left\{D^\lambda \mathbf{H}(\bar{Y})\right\}^\top (1 + \bar{Y}^2)^{\delta_{Y0}} d\nu(Y) \right] \gamma_a \le B_Y,$$

where $\gamma_a = (\gamma_a^1, \ldots, \gamma_a^{J_Y})^\top$. Therefore, we choose the positive definite matrices $\mathbf{\Lambda}_X$ and $\mathbf{\Lambda}_Y$ in the constraint for regularization as

$$\begin{aligned}
\mathbf{\Lambda}_X &= \sum_{|\lambda| \le m_X + m_{X0}} \int \left\{D^\lambda \mathbf{H}(\bar{X})\right\} \left\{D^\lambda \mathbf{H}(\bar{X})\right\}^\top (1 + \bar{X}^\top \bar{X})^{\delta_{X0}} d\nu(X), \\
\mathbf{\Lambda}_Y &= \sum_{|\lambda| \le m_Y + m_{Y0}} \int \left\{D^\lambda \mathbf{H}(\bar{Y})\right\} \left\{D^\lambda \mathbf{H}(\bar{Y})\right\}^\top (1 + \bar{Y}^2)^{\delta_{Y0}} d\nu(Y).
\end{aligned}$$

Therefore, the proposed estimators of $\xi_{ra}(X)$ and $f_{ra}(X)$ are

$$\hat{f}_{ra}(X) = \sum_{j=1}^{J_X} \hat{\theta}_{ra}^j h_j(\bar{X}) \quad \text{and} \quad \hat{\phi}_a(Y) = \sum_{j=1}^{J_Y} \hat{\gamma}_a^j h_j(\bar{Y})$$

where $\hat{\theta}_{ra}$ and $\hat{\gamma}_{ra}$ minimizes the objective function subject to the constraint that

$$\gamma_{ra}^{\top}\mathbf{\Lambda}_X\,\gamma_{ra} \;\leq B_X \quad \text{and} \quad \theta_a^{\top}\mathbf{\Lambda}_Y\,\theta_a \;\leq B_Y.$$

The regularization is not restrictive in two senses. By definition, functions in $\mathcal{G}_{m,m_0,\delta_0,B}$ satisfy two natural conditions that first, the bound B guarantees sufficient differentiability. Second, it constrains function behavior outside the main region, avoiding explosive growth at extreme values. In practice, we only care about the functions $f_{ra}(X)$ and $\phi_a(Y)$ within a compact region covering the observed data, where they are expected to be smooth. Thus, this regularization merely formalizes standard assumptions and does not impose restrictions.

S.10.3 Detailed computation algorithms

We provide neat conclusions of our implementation algorithm in this section. Our computational algorithms build on the machine learning frameworks, for example, in Prince (2023).[7] In particular, we use a neural network for nonparametric regression, learning the conditional mean function $m(x) = \mathbb{E}[y \mid x]$ without assuming any specific parametric form. Meanwhile, to model both the marginal density $f(x)$ and the full conditional density $f(y \mid x) = \frac{f(x,y)}{f(x)}$, we employ normalizing flows that is the chains of invertible transformations T that map a simple base distribution $p_0(z)$, such as a Gaussian, to a complex target distribution. The training objective is the log likelihood by the change of variables,

$$\log p(y \mid x) = \log p_0\Big(T^{-1}(x,y)\Big) + \log\left|\det \frac{\partial T^{-1}(x,y)}{\partial y}\right|,$$

which ensures both exact density evaluation and efficient sampling. Due to their invertibility and tractable Jacobian determinants, these flows can capture multimodal structures and heteroskedasticity behavior, providing richer uncertainty quantification compared to standard Gaussian density. Normalizing flows have been widely used for expressive density and conditional modeling due to these properties.

We provide three algorithms to compute τ below. The three algorithms are based on three identification equations that we proposed in Theorem 1 equation (7) and Theorem 2 equation (11). We therefore have computation algorithm. We want to know how to calculate different values and plug them into different estimators. We can estimate by minimizing the residual sum of squares. Although common techniques, such as Tikhonov

[7] Our implementation draws from the UVADL course notebooks (https://uvadlc-notebooks.readthedocs.io/en/latest/).

regularization (Darolles et al., 2011) or penalized sieve minimum distance (Chen and Pouzo, 2015) can also stabilize ill-posed inverse problems, we adhere to the approach of (Newey and Powell, 2003) and (Yang et al., 2019), applying compactness constraints directly to the approximating function spaces. Concretely, we assume $f_{ra}, \hat{f}_{ra} \in \mathcal{G}_{m_X, m_{X0}, \delta_{X0}, B_X}$ and $\phi_a, \hat{\phi}_a \in \mathcal{G}_{m_Y, m_{Y0}, \delta_{Y0}, B_Y}$ a Sobolev-type space whose closure is compact under a norm enforcing bounded derivatives and tails.

In practice, when we approximate $f_{ra}(X)$ and $\phi_a(X)$ by basis expansions $\hat{f}_{ra}(X) = \sum_{j=1}^{J_X} \hat{\theta}_{ra}^j h_j(\bar{X})$ and $\hat{\phi}_a(Y) = \sum_{j=1}^{J_Y} \hat{\gamma}_a^j h_j(\bar{Y})$, because the sample version of the approximation is linear, we can estimate the γ_{ra}^j and θ_{ra}^j by minimizing the residual sum of squares

$$\sum_{i=1}^{n} \left[\hat{f}(X_{r,i}, Y_i \mid A_i = a, R_i = r) - \sum_{j=1}^{J_X} \theta_{ra}^j \hat{H}_{ra}^j(X_{r,i}, Y_i) \right]^2 . \tag{S9}$$

and for the alternative integral equation

$$\sum_{i=1}^{n} \left[\hat{f}(A_i = a, Y_i) - \sum_{j=1}^{J_Y} \gamma_a^j \hat{H}_a^j(Y_i) \right]^2 . \tag{S10}$$

where to address the ill-conditioned nature of our estimation, we restrict the infinite dimensional functions $f_{ra}(X)$, $\phi_a(Y)$ as well as their sieve-estimators $\hat{f}_{ra}(X)$, $\hat{\phi}_a(Y)$ to lie in a compact function space. This restriction serves as a regularization, rendering the underlying inverse problem well-posed. The compactness condition becomes a constraint on the coefficient vector $\gamma_{ra}^\top \mathbf{\Lambda}_X \gamma_{ra} \leq B_X$ and $\theta_a^\top \mathbf{\Lambda}_Y \theta_a \leq B_Y$. where $\mathbf{\Lambda}_X$ and $\mathbf{\Lambda}_Y$ are positive definite matrices derived from integrals of basis derivatives. After getting the estimators of $\hat{f}_{ra}(X)$ and $\hat{\phi}_a(Y)$, we can derive the average treatment effect by three proposed ways.

Algorithm 1 (OBCATE)

Step 1 Obtain machine learning estimators of $f(X \mid A = a, R = 1_p)$, $f(X_r, Y \mid A = a, R = 1_p)$, for all r and a. Denote a neural network estimator of $\mathbb{E}(Y \mid A = a, R = 1_p)$ by $\hat{\mathbb{E}}(Y \mid A = a, X, R = 1_p)$ for $a = 0, 1$. Let $\hat{f}(X_r, Y \mid A = a, R = r)$ denote the normalizing flows estimators of $f(X_r, Y \mid A = a, R = r)$, respectively.

Step 2 Obtain a series estimator of $f_{ra}(X)$ using Hermite polynomials,

$$\hat{f}_{ra}(X) = \sum_{j=1}^{J} \hat{\theta}_{ra}^j h_j(\bar{X})$$

where $\left(\hat{\theta}_{ra}^1, \ldots, \hat{\theta}_{ra}^J\right)^\top$ minimizes equation (S9) subject to the constraint $\hat{\theta}_{ra}^\top \Lambda_X \hat{\theta}_{ra} \leq B_X$.

Step 3 For any $a = 0, 1$

$$\hat{\mu}_a(A, X_r, R = r) = \int \frac{\hat{\mathbb{E}}(Y \mid A = 1, X, R = 1_p)\hat{f}_{ra}(X)}{\hat{f}(X_r \mid A = a, R = r)} d\nu(X_{\bar{r}}).$$

Step 4 Based on Theorem 2 and subsection 4.1, estimate τ by using a numerical approximation

$$\hat{\tau}_{\text{MREG}} = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(A_i, X_{R_i}, R_i) - \hat{\mu}_0(A_i, X_{R_i}, R_i)\right].$$

Algorithm 2 (IJPW)

Step 1 Obtain machine learning estimators of $\tau(X)$, $f(X \mid A = a, R = 1_p)$, $f(X_r, Y \mid A = a, R = 1_p)$, for all r and a. Let

$$\hat{\tau}(X) = \hat{\mathbb{E}}(Y \mid A = 1, X, R = 1_p) - \hat{\mathbb{E}}(Y \mid A = 0, X, R = 1_p), \tag{S11}$$

where $\hat{\mathbb{E}}(Y \mid A = a, X, R = 1_p)$ denotes a neural network estimator of $\mathbb{E}(Y \mid A = a, R = 1_p)$, for $a = 0, 1$. Let $\hat{f}(X \mid A = a, R = 1_p)$ and $\hat{f}(X_r, Y \mid A = a, R = r)$ denote the normalizing flows estimators of $f(X \mid A = a, R = 1_p)$ and $f(X_r, Y \mid A = a, R = r)$, respectively.

Step 2 Obtain a series estimator of $\phi_a(X)$ using Hermite polynomials,

$$\hat{\phi}_a(Y) \approx \sum_{j=1}^{J} \hat{\gamma}_a^j \, h_j(\bar{Y}),$$

where $\left(\hat{\gamma}_a^1, \ldots, \hat{\gamma}_a^J\right)^\top$ minimizes equation (S10) subject to the constraint $\hat{\boldsymbol{\gamma}}_a^\top \Lambda_Y \hat{\boldsymbol{\gamma}}_a \leq B_Y$.

Step 3 Estimate $\int \hat{\phi}_a(Y)\hat{f}(Y \mid A = a, X, R = 1_p)\, d\nu(Y)$ and then calculate the joint propensity score

$$\hat{e}_a(X) = \frac{\hat{f}(A = a, X, R = 1_p)}{\hat{P}(R = 1_p \mid A = a, X)}.$$

for $a = 0, 1$.

Step 4 Calculate the inverse joint propensity weighting (IJPW) estimator:

$$\hat{\tau}_{\text{IJPW}} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{A_i \, \mathbb{I}\{R_i = 1_p\}\, Y_i}{\hat{e}_1(X_i)} - \frac{(1 - A_i)\, \mathbb{I}\{R_i = 1_p\}\, Y_i}{\hat{e}_0(X_i)}\right]. \tag{S12}$$

Algorithm 3 (Doubly Robust)

Step 1 Follow the steps 1-3 in Algorithm 1-2.

Step 2 Based on the doubly robust estimator given in equation (17) or (18),

$$\begin{aligned}\hat{\tau}_{\text{AIJPW}} =& \frac{1}{n}\sum_{i=1}^{n}\Bigg\{\hat{\mu}_1(A_i, X_{R_i}, R_i) - \hat{\mu}_0(A_i, X_{R_i}, R_i) \\ &+ \frac{A_i\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \hat{\mu}_1(A_i, X_{R_i}, R_i)]}{\hat{e}_1(X_i)} \\ &- \frac{(1-A_i)\,\mathbb{I}\{R_i = 1_p\}\,[Y_i - \hat{\mu}_0(A_i, X_{R_i}, R_i)]}{\hat{e}_0(X_i)}\Bigg\}. \end{aligned} \tag{S13}$$

S.10.4 Choice of tuning parameters

When implementing the estimator, we need to carefully balance several tuning parameters: the number of Hermite polynomial terms J, the regularization bound B. Choosing larger values for J and B improves the estimator's ability to closely approximate the true function. On the other hand, if J and B become too large, we introduce excessive variance into our estimates. Prior work by Chen and Pouzo (2012) and Chen and Christensen (2018) provides theoretical guidance on how these parameters should scale with sample size in the context of penalized sieve minimum distance estimation. In practice, we recommend selecting tuning parameters using data-driven techniques, such as cross-validation, and complementing this approach with sensitivity analyses to assess the robustness of the results under different parameter choices.

S.11 Proof of Theorem 7

In this proof, we verify two essential properties of the influence function. First, we show that it satisfies the geometric characterization both for the parameter of interest and the nuisance parameter. Second, we demonstrate that the influence function lies in the tangent space. Together, these two parts establish that the influence function is the efficient influence function, achieving the optimal behavior with respect to both components of the parameter space.

S.11.1 Setting

Let the observed data be denoted by $\mathcal{O} = (A,\, X_R,\, Y,\, R)$ where Y is always observed, and R is the missingness indicator vector specifying which components of the confounder vector X are observed. The notation X_R refers to the subvector of X corresponding to those components for which R indicates observation. We define the full data as $Z = (A, X, Y, R)$,

which is not fully observed when some entries of X are missing. We assume outcome-independent missingness, $R \perp Y \mid A, X$. We also assume a bridge restriction, there exists a measurable function $\phi_{ra}(Y)$ such that for each missingness pattern r and treatment level a, $\int \phi_{ra}(Y)\, f(Y \mid A = a, X, R = r)\, d\nu(Y) = 1/\pi_r(a, X)$, where $\pi_r(a, X) := P(R = r \mid A = a, X)$. Additionally, we impose extra conditions on certain parts of the distribution, as specified in Assumptions 5 and 7.

Let $f_\theta(Z)$ satisfying the assumptions described above be a regular parametric submodel, with the true density given by $f = f_0$ at $\theta = 0$. A common choice of submodel for nonparametric P is, for some mean-zero function $h : \mathcal{Z} \to \mathbb{R}$,

$$f_\theta(Z) \;=\; f(Z)\{1 + \delta\, h(Z)\},$$

with respect to the dominating measure ν, where $\|\delta\|_\infty \leq M < \infty$ and $|\theta| < 1/M$ so that $f_\theta(Z) \geq 0$ ν-almost surely. By the outcome-independent missing not at random assumption, the full-data score decomposes as

$$\begin{aligned} s_\theta(Z) &= s_\theta(A, X) + s_\theta(Y \mid A, X) + s_\theta(R \mid A, X, Y), \\ &= s_\theta(A, X) + s_\theta(Y \mid A, X) + s_\theta(R \mid A, X), \end{aligned}$$

where each term is the score for the corresponding component of the full-data likelihood. We adopt a slightly different decomposition from Jiang et al. (2022) because the score $S(A, X)$ is enough for characterizing the efficient influence function. Specifically,

$$\begin{aligned} s_\theta(A, X) &= \frac{\partial}{\partial\theta} \log f_\theta(A, X), & s_\theta(R \mid A, X, Y) &= \frac{\partial}{\partial\theta} \log P_\theta(R \mid A, X, Y), \\ s_\theta(Y \mid A, X) &= \frac{\partial}{\partial\theta} \log f_\theta(Y \mid A, X), & s_\theta(R \mid A, X) &= \frac{\partial}{\partial\theta} \log P_\theta(R \mid A, X). \end{aligned}$$

These score components satisfy the zero-mean property. For additional notation, we use a dot to denote the partial derivative with respect to θ. For example, $\dot{\tau}_\theta(A, X_R, R) = \frac{\partial}{\partial\theta}\, \tau_\theta(A, X_R, R)$, where the derivative is taken along the parametric submodel.

We rely on the complete-case pattern 1_p, which is the mask that reveals all required entries of X. For any observed data point $\mathcal{O}$, we write

$$\varphi(\mathcal{O}) = h(A, X_R, R) + \mathbb{I}\{R = 1_p\}\rho(A, X, Y),$$

where

$$
\begin{aligned}
h(A, X_R, R) &\coloneqq \mu_1(A, X_R, R) - \mu_0(A, X_R, R) - \tau, \\
\rho(Y, A, X) &\coloneqq \frac{A}{e_1(X)}[Y - m_1(X)] - \frac{1-A}{e_0(X)}[Y - m_0(X)] \\
&= \frac{A}{e_1(X)}[Y - \mu_1(A, X_R, R)] - \frac{1-A}{e_0(X)}[Y - \mu_0(A, X_R, R)].
\end{aligned}
$$

The extra equality in the second equation holds because $R = 1_p$ determines the value of φ.

S.11.2 Semiparametric efficiency bound

As shown in Tsiatis (2006), every regular asymptotically linear estimator admits the Riesz representation, $\frac{d}{d\theta}\Big|_{\theta=0} \tau_\theta = \mathbb{E}\left[\varphi(\mathcal{O}) s_\theta(\mathcal{O})\right]$. To establish efficiency, one must further show that this influence function lies in the model's tangent space. The extra constraint requires that for each a, there exists a function ϕ_a such that

$$
\int \phi_a(Y)\, f(Y \mid A = a, X, R = 1_p)\, d\nu(Y) = \frac{1}{\pi(a, X)}, \qquad \pi(a, X) = \mathbb{P}(R = 1_p \mid A = a, X),
$$

and so does the submodel,

$$
\int \phi_{a,\theta}(Y)\, f_\theta(Y \mid A = a, X, R = 1_p)\, d\nu(Y) = \frac{1}{\pi_\theta(a, X)}, \qquad \pi_\theta(a, X) = \mathbb{P}_\theta(R = 1_p \mid A = a, X).
$$

Differentiating this constraint with respect to θ on both sides gives the linear constraint on the scores

$$
\pi_\theta(a, X) = \Lambda_{s_\theta}(a, X) \tag{S14}
$$

where $\Lambda_{s_\theta}(a, X) \coloneqq -\pi_\theta(a, X)\, \mathbb{E}\left[\dot{\phi}_{a,\theta}(Y) + \phi_{a,\theta}(Y)\, s_\theta(Y \mid A = a, X) \mid A = a, X\right]$. Note we can have $\dot{\phi}_a(Y)$ be the unique solution in $s_\theta(R = 1_p \mid A = a, X) \mid_{\theta=0} = \Lambda_{s_\theta} \mid_{\theta=0}$ because $f(Y \mid A, X, R = 1)$ is complete in Y and $Y \perp\!\!\!\perp R \mid A, X$.

The observed data tangent space can be defined as the set of linear operators acting on the full data scores (Bickel et al., 1993; Robins et al., 1994; Zhang and Tchetgen Tchetgen, 2022). For instance, Theorem 7.1 in Tsiatis (2006) provides a useful reference, where the full data is instead defined as (A, X, Y) rather than $Z = (A, X, Y, R)$, so the MAR assumption is required. See Lemma S15 for an explanation of this. Therefore, starting with the treatment confounder score $s(A, X)$, we can set the operator as (ignoring the

parameter θ for simplicity, since it does not affect the result)

$$T_{\mathcal{O},1}[s(A,X)] \coloneqq \mathbb{E}\left[s(A,X) \mid \mathcal{O}\right] = \sum_r \mathbb{I}\{R=r\}\,\mathbb{E}\left[s(A,X) \mid A, X_R, Y, R=r\right].$$

For the outcome score $s(Y \mid A,X)$, the operator is defined as

$$T_{\mathcal{O},2}[s(Y \mid A,X)] \coloneqq \mathbb{E}\left[s(Y \mid A,X) + \mathbb{I}\{R=1_p\}\Lambda_s(A,X) \mid \mathcal{O}\right],$$

which can be rewritten as

$$\begin{aligned} T_{\mathcal{O},2}[s(Y \mid A,X)] = \mathbb{I}\{R=1_p\}&\Big[s(Y \mid A,X) + \Lambda_s(A,X)\Big] \\ &+ \sum_r \mathbb{I}\{R\neq 1_p\}\mathbb{E}[s(Y \mid A,X) \mid A, X_R, Y, R=r] \end{aligned}$$

where $\mathbb{E}[s(Y \mid A,X) \mid A,X,Y,R=1_p] = s(Y \mid A,X)$ and $\mathbb{E}\left[\Lambda_s(A,X) \mid A,X,Y,R=1_p\right] = \Lambda_s(A,X)$. And finally,

$$T_{\mathcal{O},3}[s(R \mid A,X)] = \sum_r \mathbb{I}\{R\neq 1_p\}\mathbb{E}[s(R \mid A,X) \mid A, X_R, Y, R=r]$$

The observed data tangent space, obtained as the closure of the scores of regular submodels under the imposed constraints, is given by

$$\begin{aligned} \mathscr{T}_{\mathcal{O}} = \Big\{ & T_{\mathcal{O},1}[g(A,X)] + T_{\mathcal{O},2}[g(A,X,Y)] + T_{\mathcal{O},3}[g(A,X,R)] : \\ & \mathbb{E}[g(A,X)] = 0, \mathbb{E}[g(A,X,Y) \mid A,X] = 0, \mathbb{E}[g(A,X,R) \mid A,X] = 0, g(A,X,Y) \in \mathcal{H} \Big\}^{\text{cl}} \\ = \Big\{ & T_{\mathcal{O},1}[g(A,X)] + T_{\mathcal{O},2}[g(A,X,Y)] + T_{\mathcal{O},3}[g(A,X,R)] : \\ & \mathbb{E}[g(A,X)] = 0, \mathbb{E}[g(A,X,Y) \mid A,X] = 0, \mathbb{E}[g(A,X,R) \mid A,X] = 0 \Big\}^{\text{cl}} \end{aligned}$$

where the operator cl denotes the closure of a set in the Hilbert space $L^2(P)$, and $\mathcal{H} = \Big\{ a\, s_\theta(Y \mid A,X) : f_\theta(Y \mid A,X)$ satisfies Assumptions 5 and 7, $a \in \mathbb{R} \Big\}$ is not restricted on directions as shown by Canay et al. (2013). We can set

$$\check{g}(A,X) = m_1(X) - m_0(X) - \tau,$$

which can be mapped onto the R indicator

$$\begin{aligned} T_{\mathcal{O},1}[\check{g}(A,X)] &= \mathbb{E}[m_1(X) - m_0(X) \mid \mathcal{O}] - \tau \\ &= \mu_1(A, X_R, R) - \mu_0(A, X_R, R) - \tau + \Delta(A, X_R, Y, R) \end{aligned}$$

where

$$\Delta(A, X_R, Y, R) \coloneqq \mathbb{E}[m_1(X) - m_0(X) \mid \mathcal{O}] - \mathbb{E}[m_1(X) - m_0(X) \mid A, X_R, R].$$

Then, we set

$$\check{g}(A,X,Y) = \underbrace{-\mathbb{E}[\tau(X) \mid A,X,Y] + \mathbb{E}[\tau(X) \mid A,X]}_{g_0(A,X,Y)} + u(A,X,Y)$$

where $u(A,X,Y)$ satisfies

$$\begin{aligned} &u(A,X,Y) + \Lambda_u(A,X) = \rho(A,X,Y) - g_0(A,X,Y) - \Lambda_{g_0}(A,X) \\ &\mathbb{E}[u(A,X,Y) \mid A, X_r, Y, R = r] = 0 \quad \text{for every } r \neq 1_p, \\ &\mathbb{E}\left[u(A,X,Y) \mid A, X\right] = 0. \end{aligned}$$

Then $\check{g}(A,X,Y)$ has the property that

$$\begin{aligned} \mathbb{E}[\check{g}(A,X,Y) \mid A,X,Y,R=1] &= \check{g}(A,X,Y), \\ \mathbb{E}[\check{g}(A,X,Y) \mid A,X_r,Y,R=r] &= -\Delta(A,X_r,Y,r). \end{aligned}$$

We thus have $\varphi(\mathcal{O}) = T_{\mathcal{O},1}[\check{g}(A,X)] + T_{\mathcal{O},2}[\check{g}(A,X,Y)]$. Therefore, $\varphi(\mathcal{O}) \in \mathscr{T}_{\mathcal{O}}$.

S.12 Auxiliary lemmas and extra results

This subsection consolidates several key results briefly discussed in the main paper, along with lemmas that are utilized across various proofs.

Lemma S9. *Let $\widehat{U}_n$ be an estimator of the target function U, and define*

$$\Delta_n \coloneqq \|\hat{U}_n(X) - U(X)\|_2.$$

Let r_n be a positive sequence. Then:

(a) If $\mathbb{E}\left(\Delta_n^2\right) = O\left(r_n^2\right)$, then $\Delta_n = O_p\left(r_n\right)$.

(b) If $\mathbb{E}\left(\Delta_n^2\right) = o\left(r_n^2\right)$, then $\Delta_n = o_p\left(r_n\right)$.

Proof. By assumption $\mathbb{E}(\Delta_n^2) \preceq r_n^2$, where $a_n \preceq b_n$ if $a_n = O(b_n)$. Then by Markov's inequality

$$P(\Delta_n > M r_n) = P(\Delta_n^2 > M^2 r_n^2) \leq \frac{E(\Delta_n^2)}{M^2 r_n^2} \preceq M^{-2}.$$

Choosing M large enough such that $M^{-2} < \varepsilon$ yields

$$P(\Delta_n > M\, r_n) \preceq \varepsilon,$$

which implies $\Delta_n = O_p(r_n)$. The little-o_p case is analogous. □

Lemma S10. *Under Assumptions 1-4,*

$$\mu_a(A, X, R = 1_p) = \mathbb{E}[Y(a) \mid X]$$

for any $a = 0, 1$.

Proof. Assumptions 1-4 ensure the conditional expectation is well defined. Recall $\mu_a(A, X_R, R) \coloneqq \int \mathbb{E}(Y \mid A = a, X) f(X_{\bar{R}} \mid A, X_R, R)\, d\nu(X_{\bar{R}})$ then we have

$$\mu_a(A, X, R = 1_p) = \mathbb{E}(Y \mid A = a, X) = \mathbb{E}[Y(a) \mid X]$$

where the first equality is by definition of $\mu_a(A, X_R, R)$ and the second equality is from Assumption 3. □

Lemma S11. *Under Assumptions 1-4,*

$$\mathbb{E}\{f(X) A\, \mathbb{I}\{R = 1_p\} [Y - \mu_1(A, X, R)]\} = 0$$
$$\mathbb{E}\{f(X)(1 - A)\, \mathbb{I}\{R = 1_p\} [Y - \mu_0(A, X, R)]\} = 0$$

for any Borel measurable function f.

Proof. $\mathbb{E}\{f(X) A\, \mathbb{I}\{R = 1_p\} [Y - \mu_1(A, X, R)]\} = \mathbb{E}\{f(X) A\, \mathbb{I}\{R = 1_p\} [Y - \mu_1(A, X, R = 1_p)]\}$ because of $\mathbb{I}\{R = 1_p\}$. By Lemma S10 and the law of iterated expectation, it is equal to,

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{E}\left(f(X) A\, \mathbb{I}\{R = 1_p\} \{Y - \mathbb{E}[Y(1) \mid X]\} \mid A, X\right)\right] \\
&= \mathbb{E}\left\{ f(X) A\, \mathbb{E}\left(\mathbb{I}\{R = 1_p\} \{Y - \mathbb{E}[Y(1) \mid X]\} \mid A, X\right)\right\} \\
&= \mathbb{E}\left\{f(X) A\, \mathbb{E}\left(\mathbb{I}\{R = 1_p\} \{Y - \mathbb{E}[Y(1) \mid X]\} \mid A = 1, X\right) \mathbf{P}(A = 1)\right\}.
\end{aligned}$$

Next, we observe

$$
\begin{aligned}
&\mathbb{E}\left(\mathbb{I}\{R=1_p\}\left\{Y-\mathbb{E}[Y(1)\mid X]\right\}\mid A=1,X\right)\\
&=\mathbb{E}\left\{\mathbb{I}\{R=1_p\}Y-\mathbb{I}\{R=1_p\}\mathbb{E}[Y(1)\mid X]\mid A=1,X\right\}\\
&=\mathbb{E}\left(\mathbb{I}\{R=1_p\}Y\mid A=1,X\right)-\mathbb{E}\left\{\mathbb{I}\{R=1_p\}\mathbb{E}[Y(1)\mid X]\mid A=1,X\right\}\\
&=\mathbb{E}\left(\mathbb{I}\{R=1_p\}\mid A=1,X\right)\mathbb{E}\left(Y\mid A=1,X\right)-\mathbb{E}\left\{\mathbb{I}\{R=1_p\}\mid A=1,X\right\}\mathbb{E}[Y(1)\mid X]\\
&=0,
\end{aligned}
$$

where the third equality is by Assumption 3 and $\mathbb{E}[Y(1)\mid X]$ is a function of X. Hence,

$$
\mathbb{E}\{f(X)A\,\mathbb{I}\{R=1_p\}\left[Y-\mu_1(A,X,R)\right]\}=0.
$$

□

Lemma S12. *Under Assumptions 1-4, for $\tau(\mathbf{Z})$ and $\psi(\mathbf{Z})$ in (S2) and (S3), we have $\mathbb{E}[\tau(\mathbf{Z})\psi(\mathbf{Z})]=0$ and $\mathbb{E}[\psi(\mathbf{Z})]=0$.*

Proof. $\mathbb{E}[\psi(\mathbf{Z})]=0$ holds by Lemma S11 by setting $f(X)=\frac{1}{e_a(X)}$ for

$$
\begin{aligned}
&\mathbb{E}\{f(X)A\,\mathbb{I}\{R=1_p\}\left[Y-\mu_1(A,X,R)\right]\}=0,\\
&\mathbb{E}\{f(X)(1-A)\,\mathbb{I}\{R=1_p\}\left[Y-\mu_0(A,X,R)\right]\}=0,
\end{aligned}
$$

respectively. We also have $\mathbb{E}[\tau(\mathbf{Z})\psi(\mathbf{Z})]=0$ by noticing

$$
\mu_a(A,X,R=1_p)=\mathbb{E}[Y(a)\mid X]
$$

from Lemma S10 and set $f(X)=\frac{\mathbb{E}[Y(a)|X]}{e_a(X)}$. □

Lemma S13. *Let $(\Omega,\mathcal{F},P)$ be a probability space, $Z\in L^2(\Omega,\mathcal{F},P)$, and $\mathcal{G}\subseteq\mathcal{H}\subseteq\mathcal{F}$ be σ-fields. Then $Var\Big(\mathbb{E}[Z\mid\mathcal{H}]\Big)\geq Var\Big(\mathbb{E}[Z\mid\mathcal{G}]\Big)$.*

Proof. This is an immediate consequence of the law of total variance. See Theorem 5.1.1 in Durrett (2019). □

Lemma S14. *Let $\{X_\lambda:\lambda\in\Lambda\}$ be a family of integrable random variables on a probability space $(\Omega,\mathcal{F},P)$ which is uniformly integrable, i.e., $\lim_{B\to\infty}\sup_{\lambda\in\Lambda}E\Big[|X_\lambda|\,\mathbb{I}\{|X_\lambda|>B\}\Big]=0$. Let $\mathcal{G}_\lambda\subseteq\mathcal{F}$ be any sub-σ-algebra (which may depend on λ). Then the family of conditional expectations $\{E[X_\lambda\mid\mathcal{G}]:\lambda\in\Lambda\}$ is also uniformly integrable.*

Proof. This is immediate from the discussion of Chapter 14 in Williams (1991). □

Lemma S15. *The parametric submodel observed-data score with respect to θ is given by*

$$s_\theta(A, X_r, Y, R = r) = \mathbb{E}[s_\theta(A, X, Y, R = r) \mid A, X_r, Y, R = r]$$

Proof. The density for the observed data is

$$f_\theta(A, X_r, Y, R = r) = \int f_\theta(A, X, Y, R = r) d\nu(X_{\bar{r}})$$

and the log-likelihood for the observed data is

$$\log f_\theta(A, X_r, Y, R = r) = \log \int f_\theta(A, X, Y, R = r) d\nu(X_{\bar{r}}).$$

Therefore, by definition, the score with respect to θ is

$$\begin{aligned} s_\theta(A, X_r, Y, R = r) &= \frac{\partial}{\partial \theta} \log f_\theta(A, X_r, Y, R = r) \mid_{\theta=\theta_0} \\ &= \frac{\partial}{\partial \theta} \log \int f_\theta(A, X, Y, R = r) d\nu(X_{\bar{r}}) \mid_{\theta=\theta_0} . \end{aligned}$$

By the chain rule and the fundamental theorem, we find that the score is

$$\begin{aligned} \frac{\int \frac{\partial}{\partial \theta} f_\theta(A, X, Y, R = r) d\nu(X_{\bar{r}}) \mid_{\theta=\theta_0}}{\int f(A, X, Y, R = r) d\nu(X_{\bar{r}})} &= \frac{\int \frac{\partial}{\partial \theta} f_\theta(A, X, Y, R = r) d\nu(X_{\bar{r}}) \mid_{\theta=\theta_0}}{f(A, X_r, Y, R = r)} \\ &= \frac{\int s_\theta(A, X, Y, R = r) f(A, X, Y, R = r) d\nu(X_{\bar{r}})}{f(A, X_r, Y, R = r)} \end{aligned}$$

where the second equality is from dividing and multiplying by $f(A, X, Y, R = r)$. Finally, it is equivalent to

$$\int s_\theta(A, X, Y, R = r) f(X_{\bar{r}} \mid A, X_r, Y, R = r) d\nu(X_{\bar{r}}) = \mathbb{E}[s_\theta(A, X, Y, R = r) \mid A, X_r, Y, R = r].$$

□