PUATE: Efficient ATE Estimation from Treated (Positive) and Unlabeled Units

Masahiro Kato **Fumiaki Kozai** **Ryo Inokuchi**
Mizuho-DL Financial Technology Co., Ltd.
Chiyoda–ku, Tokyo 102–0083
masahiro-kato@fintec.co.jp

Abstract

The estimation of *average treatment effects* (ATEs), defined as the difference in expected outcomes between treatment and control groups, is a central topic in causal inference. This study develops semiparametric efficient estimators for ATE in a setting where only a treatment group and an unlabeled group—consisting of units whose treatment status is unknown—are observed. This scenario constitutes a variant of learning from positive and unlabeled data (*PU learning*) and can be viewed as a special case of ATE estimation with missing data. For this setting, we derive the *semiparametric efficiency bounds*, which characterize the lowest achievable asymptotic variance for regular estimators. We then construct semiparametric *efficient ATE estimators* that attain these bounds. Our results contribute to the literature on causal inference with missing data and weakly supervised learning.

1 Introduction

The estimation of *average treatment effects* (ATEs), defined as the difference in expected outcomes between treatment and control groups, is a fundamental problem in causal inference (Imbens & Rubin, 2015). Estimating ATEs enables researchers to quantify the causal impact of a treatment, intervention, or policy on an outcome of interest. This problem is of paramount importance across various fields, including economics, epidemiology, and machine learning.

Standard ATE estimation typically assumes access to both treatment and control groups, along with complete information on treatment assignment. However, in many practical situations, this assumption does not hold. In some cases, only a *treatment group* and an *unknown group*—comprising units for which treatment assignment is unobserved—are available. Such scenarios arise in various applications, including recommendation systems with implicit feedback, electronic health records, and marketing campaigns, where the absence of explicit treatment information poses significant challenges for causal inference.

We present the following examples for applications of our proposed method:

- We are interested in how building an online store affects product sales. In a recommendation system, a customer who purchases a product through the website is known to have visited the site, whereas a customer who buys the product in a physical store may or may not have visited the website. When the construction of the online store is regarded as the treatment, customers who purchase in-store belong to the unknown group.
- In a similar vein, consider the effect of distributing coupons on product sales. Because the logging system is imperfect, we observe two datasets: purchase records for customers who used coupons, and purchase records that mix customers who did and did not use coupons. Here, customers who appear only in the mixed dataset constitute the unknown group.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

- A company unintentionally sold defective products, for example, vehicles with faults or food containing impurities. Consumers who notice the defect report both the defect and any resulting damage, such as health issues or accidents, to the company. Consumers who do not notice the defect file no report. In this setting, consumers who report the defect form the treatment group, while those who do not notice the defect belong to the unknown group.

Our method is also useful when the treatment can be defined only by contrasting it with a non-treatment condition. For example, consider examining how excessive work hours affect workers' mental health. Courts and labor inspectors can identify illegal overtime, but it is difficult to decide whether other working hours are excessive. Firms found to have engaged in illegal overtime can therefore serve as the treatment group, whereas all other firms constitute the unknown group. Similarly, by taking units with easily observable outcomes as the treatment group and all others as the unknown group, one can define the comparison group simply as not treated. This idea is widely adopted in anomaly-detection approaches that rely on density-ratio or PU learning.

Content of this study. We address the problem of ATE estimation using only a treatment group and an unknown group. This setting is closely related to learning from positive and unlabeled data (*PU learning*, Sugiyama et al., 2022), where the goal is to train a classifier using only positive and unlabeled instances. In our context, the challenge lies in efficiently estimating ATEs using the treatment (positive) and unknown groups. We refer to our setting and methodology as *PUATE*.

For this problem, we first derive *semiparametric efficiency bounds*, which are theoretical lower bounds on the asymptotic variance of regular estimators under the given data-generating processes (DGPs).[1] These bounds serve as benchmarks for evaluating estimator performance. As part of this derivation, we compute the *efficient influence function*, which provides insight into the construction of *efficient estimators*.

Using the efficient influence function, we develop semiparametric efficient ATE estimators that are $\sqrt{n}$-consistent and whose asymptotic variance achieves the efficiency bounds. These estimators are thus optimal under the semiparametric framework.

In this study, we consider two DGPs relevant to the PU setup: the censoring setting and the case-control setting (Elkan & Noto, 2008; du Plessis et al., 2015). In the censoring setting, we are given a single dataset in which some units have missing treatment information while others are confirmed to have received treatment. In the case-control setting, we are provided with two datasets: one containing treated units and another comprising units with unknown treatment status.

Specifically, our contributions are as follows:

- We formulate the ATE estimation problem with missing data using the PU learning framework.
- We derive efficiency bounds under both the censoring and case-control settings.
- We propose novel efficient estimators.
- We establish connections between ATE estimation with missing data and PU learning.
- We also propose alternative candidate estimators.

This study is organized as follows. Section 2 formulates our problem. Section 3 introduces simple ATE estimators, which are later shown to be inefficient. In Section 4, we derive efficiency bounds, propose an efficient estimator, and establish its asymptotic properties under the censoring setting. Due to space limitations, we show the ATE estimation in the case-control setting briefly in Section 5 and mainly in Appendix D. Section 6 presents simulation studies. We introduce related work in Appendix A. The details of PU learning methods are explained in Appendix E

2 Problem setting

2.1 Potential outcomes and parameter of interest

We consider binary treatments, 1 and 0. For each treatment $d \in \{1, 0\}$, there exists a potential outcome $Y(d) \in \mathbb{R}$. The outcome is observed only when the corresponding treatment is assigned

[1] For regular estimators, see p.366 in van der Vaart (1998).

to a unit. Each unit has p-dimensional covariates $X \in \mathcal{X} \subset \mathbb{R}^p$. This setting is called the Neyman-Rubin causal model (Neyman, 1923; Rubin, 1974).

We denote by $P_0 \in \mathcal{P}$ a distribution of $Y(d)$, X, and other random variables introduced below, which we call the true distribution, where $\mathcal{P}$ is the set of distributions. For simplicity, we assume that P_0 has a density. We denote the conditional density of $Y(d)$ given X by $p_{Y(d),0}(y(d) \mid X)$, and the marginal density of X by $\zeta_0(x)$.

We consider n units, each of which receives either treatment or control. Let $Y_i(d)$ and X_i be i.i.d. copies of $Y(d)$ and X. Throughout this study, for a random variable R, let R_i be its i.i.d. copy under P_0. If unit i receives treatment d, we observe $Y_i(d)$ but not the counterfactual outcome.

We refer to the group of units who receive treatment 1 as the *treatment group* and those who receive treatment 0 as the *control group*. As formulated in the next subsection, we consider a scenario where part of the treatment group and a *mixture* of the treatment and control groups are observable, where the treatment indicator is unobservable. We refer to this mixed group as the *unknown group*.

Our objective is to estimate the ATE under P_0 using observed data, defined as $\tau_0 := \mathbb{E}[Y(1) - Y(0)]$, where $\mathbb{E}$ denotes the expectation under P_0.

2.2 Observations with two DGPs

In our setting, the observations are non-standard. We can only observe part of the treatment group and the unknown group, a mixture of the treatment and control groups. This setting is a variant of PU learning. PU learning encompasses two settings: the *censoring setting* and the *case-control setting*. In the censoring setting, we consider a single dataset with i.i.d. observations, where treatment labels contain missing values. Specifically, while part of the treatment group is observed, the mixture of the treated and control groups is also present. In the case-control setting, we observe two independent datasets: one consisting solely of the treatment group and the other comprising the unknown group. The censoring and case-control settings are also referred to as *one-sample* and *two-sample* settings, respectively (Niu et al., 2016). The case-control setting can also be regarded as a form of stratified sampling, studied in Imbens & Lancaster (1996) and Wooldridge (2001).

2.3 Censoring setting

In the censoring setting, we observe a single dataset $\mathcal{D}$, defined as follows:

$$\mathcal{D} := \left\{ \left(X_i, O_i, Y_i\right) \right\}_{i=1}^{n} \text{ with } \left(X_i, O_i, Y_i\right) \sim p_0(x, o, y),$$

where $O_i \in \{1, 0\}$ is an observation indicator with the *observation probability* $\pi_0(O \mid X)$, Y_i is defined as

$$Y_i := \mathbb{1}[O_i = 1] Y_i(1) + \mathbb{1}[O_i = 0] \widetilde{Y},$$

$\widetilde{Y}_i$ is the observation of the unknown groups and defined as

$$\widetilde{Y}_i := \mathbb{1}[D_i = 1] Y_i(1) + \mathbb{1}[D_i = 0] Y_i(0),$$

$D_i \in \{1, 0\}$ is a (unobserved) treatment indicator, and we denote the conditional probability of D_i given X and $O_i = 0$ as $g_0(D \mid X) = \mathbb{P}(D \mid X, O = 0)$. We refer to $g_0(d \mid X)$ as the *censoring propensity score*. Here, the density $p_0(x, o, y)$ is given as $p_0(x, o, y) = \zeta_0(x)\pi_0(o \mid X = x)p_{Y,0}(y \mid O = o, X = x)$, where $p_{Y,0}$ is the density of Y.

2.4 Case-control setting

In the case-control setting, we observe two stratified datasets, $\mathcal{D}_{\mathrm{T}}$ and $\mathcal{D}_{\mathrm{U}}$:

$$\mathcal{D}_{\mathrm{T}} := \left\{ \left(X_{\mathrm{T},j}, Y_j(1)\right) \right\}_{j=1}^{m} \text{ with } \left(X_{\mathrm{T},j}, Y_j(1)\right) \sim p_{\mathrm{T},0}(x, y(1)) \text{ and}$$

$$\mathcal{D}_{\mathrm{U}} := \left\{ \left(X_k, Y_{\mathrm{U},k}\right) \right\}_{k=1}^{l} \text{ with } \left(X_k, Y_{\mathrm{U},k}\right) \sim p_{\mathrm{U},0}(x, y_{\mathrm{U}}),$$

where m and l are fixed sample sizes of each dataset such that $m + l = n$, $X_{\mathrm{T},j}$ represents the covariates of the treatment group, $Y_{\mathrm{U},k}$ is the observed outcome defined as

$$Y_{\mathrm{U},k} = \mathbb{1}[D_k = 1] Y_k(1) + \mathbb{1}[D_k = 0] Y_k(0),$$

and $D_k \in \{1, 0\}$ is a treatment indicator with probability $e_0(D \mid X)$. We refer to $e_0(d \mid X = x)$ as the *case-control propensity score*. The densities $p_{\mathrm{T},0}(x, y(1))$ and $p_{\mathrm{U},0}(x, y_{\mathrm{U}})$ satisfy $p_{\mathrm{T},0}(x, y(1)) = \zeta_{\mathrm{T},0}(x) p_{Y(1),0}(y(1) \mid X = x)$ and $p_{\mathrm{U},0}(x, y_{\mathrm{U}}) = \zeta_0(x) p_{Y_{\mathrm{U}},0}(y_{\mathrm{U}} \mid X = x)$ respectively, where $p_{Y_{\mathrm{U}},0}(y_{\mathrm{U}} \mid X = x)$ denotes the density of $Y_{\mathrm{U},k}$ given $X = x$, and $\zeta_{\mathrm{T},0}(x)$ represents the density of the covariates X in the treatment group.

Although the ATE estimation in the case-control setting is also crucially important, due to the space limitation, we show the main results almost in Appendix D.

2.5 Difference between the two settings

We illustrate the concept of the censoring and case-control settings in Figure 1. A summary of the differences is provided below:

Censoring setting: A single dataset is observed, containing partial treatment information and a mixture of treated and control groups.

Case-control setting: Two stratified datasets are observed—one consisting of the treatment group and the other comprising the unknown group.

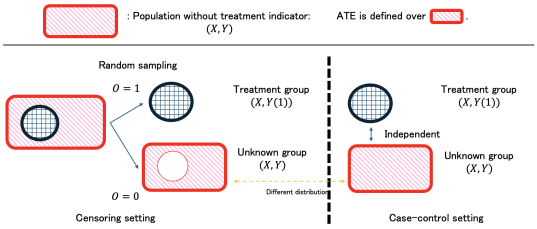

Figure 1: Illustration of the censoring and case-control settings

The key distinction lies in the randomness of treatment group observations. In the censoring setting, the observation of the treatment group is a random event, where the observation indicator O_i follows a probability $\pi_0(O \mid X)$. In contrast, in the case-control setting, the label observation is deterministic, and the treatment and unknown groups are drawn independently. This difference impacts the estimator design and efficiency bounds.

Note that the definitions and meanings of the propensity scores in the censoring and case-control settings are also different, and its difference stems from the definition of the observations indicators.

Notation. We summarize the notations above and introduce new notations. Let $\mathbb{E}$, $\mathbb{P}$, and Var be an expectation operator, a probability law, and a variance operator. For both settings, let us define $\mu_{\mathrm{T},0}(X) \coloneqq \mathbb{E}[Y(1) \mid X]$, $\mu_{\mathrm{C},0}(X) \coloneqq \mathbb{E}[Y(0) \mid X]$, and let $\tau_0(X) \coloneqq \mu_{\mathrm{T},0}(X) - \mu_{\mathrm{C},0}(X)$ be the conditional ATE. If a function f depends on the true distribution P_0, we denote it by f_0.

In the censoring setting, we use $\pi_0(o \mid X) \coloneqq \mathbb{P}(O = o \mid X)$, $g_0(d \mid X) \coloneqq \mathbb{P}(D = d \mid X, O = 0)$, and $\nu_0(X) \coloneqq \mathbb{E}[\widetilde{Y} \mid X, O = 0]$. Here, $\mathbb{E}[\mathbb{1}[O = 1]Y \mid X] = \pi_0(1 \mid X)\mu_{\mathrm{T},0}(X)$ and $\mathbb{E}[\mathbb{1}[O = 0]Y \mid X] = \pi_0(0 \mid X)\nu_0(X) = \pi_0(0 \mid X)g_0(1 \mid X)\mu_{\mathrm{T},0}(X) + \pi_0(0 \mid X)g_0(0 \mid X)\mu_{\mathrm{C},0}(X)$ hold under Assumption 3.2, defined later.

In the case-control setting, we use $e_0(d \mid X) \coloneqq \mathbb{P}(D = d \mid X)$ and $\mu_{\mathrm{U},0}(X) \coloneqq \mathbb{E}[Y_{\mathrm{U}} \mid X]$. Here, $\mu_{\mathrm{U},0}(X) = e_0(1 \mid X)\mu_{\mathrm{T},0}(X) + e_0(0 \mid X)\mu_{\mathrm{C},0}(X)$ holds under Assumption C.3. Let $r_0(X) \coloneqq \frac{\zeta_0(X)}{\zeta_{\mathrm{T},0}(X)}$ be the density ratio between the covariate densities.

3 Example of ATE estimators in the censoring setting

In this section, as a preliminary, we suggest a simple ATE estimator in the censoring setting. In Appendix D, we show the detailed result of the estimator and also propose a simple ATE estimator in the case-control setting. Note that as discussed in the subsequent subsections, the estimators are not efficient, i.e., there exist ATE estimators whose asymptotic variance is smaller.

3.1 The Inverse probability weighting estimator

We first consider the censoring setting. We begin by the arguments from the estimation of the propensity score. To estimate the propensity score, we employ the result in PU learning. Elkan & Noto (2008) addresses PU learning in the censoring setting. In that work, they show that under the Selected Completely At Random (SCAR) assumption defined below (Elkan & Noto, 2008). This assumption is analogy of the Missing Completely At Random assumption (MCAR), which is common

with the missing data literature (Little & Rubin, 2002; Rubin, 1974; Bekker & Davis, 2020). Several works such as Bekker & Davis (2018) attempt to relax the SCAR assumption, but to relax the assumption, we usually require additional assumptions.

Assumption 3.1 (SCAR)**.** *It holds that* $\mathbb{P}(D = 1, O = o \mid X) = \mathbb{P}(D = 1 \mid X)\mathbb{P}(O = o \mid D = 1)$.

From the assumption, we have $\pi_0(1 \mid x) = \mathbb{P}(D = 1 \mid X)\mathbb{P}(O = 1 \mid D = 1)$ since $\mathbb{P}(D = 0, O = 1 \mid X) = 0$ holds by definition of the DGP. Thus, under this assumption, the propensity scores $g_0(d \mid x)$ and $e_0(d \mid x)$ can be estimated using PU learning methods under some conditions (du Plessis et al., 2015; Elkan & Noto, 2008)[2]. If we know their true values, such assumptions are unnecessary. Note that the censoring PU learning studies aim to estimate $\mathbb{P}(D = d \mid X)$ not $g_0(d \mid X) = \mathbb{P}(D = d \mid X, O = 0)$. However, once we obtain an estimate of $\mathbb{P}(D = 1 \mid X)$ and $\pi_0(0 \mid X)$, we can obtain $g_0(1 \mid X)$ from $g_0(1 \mid X) = \mathbb{P}(O = 0 \mid D = 1)\mathbb{P}(D = 1 \mid X)/\pi_0(0 \mid X)$.

We further make the following unconfoundedness and common support assumptions, which are common in ATE estimation.

Assumption 3.2 (Unconfoundedness in the censoring setting)**.** *The potential outcomes* $(Y(1), Y(0))$ *are independent of treatment assignment given covariates:* $(Y(1), Y(0)) \perp\!\!\!\perp (O, D) \mid X$.

Assumption 3.3 (Common support in the censoring setting)**.** *There exists a constant* c *independent of* n *such that for all* $x \in \mathcal{X}$, $\pi_0(o \mid x), g_0(d \mid x), \zeta_0(x) > c$ *hold.*

Under these assumptions, the ATE τ_0 is estimable by replacing the following two quantities with sample approximation: $\mathbb{E}[Y(1)] = \mathbb{E}\left[\frac{\mathbb{1}[O=1]Y}{\pi_0(1|X)}\right] = \mathbb{E}\left[\mu_{\mathrm{T},0}(X)\right]$ and $\mathbb{E}[Y(0)] = \mathbb{E}\left[\frac{\mathbb{1}[O=0]Y}{g_0(0|X)\pi_0(0|X)} - \frac{g_0(1|X)\mathbb{1}[O=1]Y}{g_0(0|X)\pi_0(1|X)}\right] = \mathbb{E}\left[\frac{1}{g_0(0|X)}\nu_0(X) - \frac{g_0(1|X)}{g_0(0|X)}\mu_{\mathrm{T},0}(X)\right]$, where g_0 and π_0 can be estimated, and expectations can be approximated by sample averages. Such an estimator is a variant of the inverse probability weighting (IPW) estimator and shown in Remark 4.4.

3.2 Toward efficient estimators

Thus, we can estimate the ATE in the censoring setting (and in the case-control setting, as shown in Section 5 and Appendix C.2). However, it is unclear whether it is efficient; that is, the (asymptotic) variance is sufficiently small. In the subsequent subsections, we investigate efficient estimators and develop efficiency bounds, which work as a lower bound for the regular estimators. The efficiency bound also suggest the construction of efficient estimators, and we actually propose an estimator whose asymptotic variance aligns with the efficiency bound.

4 Semiparametric efficient ATE estimation under the censoring setting

This section presents a method for ATE estimation under the censoring setting. First, we derive the efficiency bound in Section 4.1. Then, we propose our estimator in Section 4.2 and show the consistency in Section 4.3 and the asymptotic normality in Section 4.4. Finally, in Section 4.5, we discuss issues related to the estimation of the propensity score.

4.1 Efficient influence function and efficiency

First, we derive the efficiency bound for regular estimators, which provides a lower bound on asymptotic variances. The efficiency bound is characterized via the efficient influence function (van der Vaart, 1998), derived as follows (Proof is provided in Appendix H):

Assumption 4.1 (Regularity conditions)**.** *The outcome* Y *and covariates* X *have finite variances. There exists a constant* $C > 0$ *independent of* n *such that* $\nu_0(X), \mu_{\mathrm{T},0}(X) \in [-C, C]$.

[2]Identifiability of the propensity scores differs between the censoring and case-control settings, and in both cases, it further depends on specific assumptions. Elkan & Noto (2008) shows that, under the censoring setting, the conditional class probability can be learned without the class prior, provided that the SCAR assumption holds. In the case-control setting, existing studies often require the class prior in advance of the conditional probability estimation. Some studies demonstrate that both the propensity score and the class prior can be identified simultaneously if parametric models are used for the propensity score (Lancaster & Imbens, 1996; Kato et al., 2018). There are also studies that investigate class prior estimation as an independent problem (Ramaswamy et al., 2016; du Plessis & Sugiyama, 2014).

Lemma 4.2. *If Assumptions 3.2–3.3, 4.1 hold, then the efficient influence function is given as* $\Psi^{\mathrm{cens}}(X, O, Y; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0, \tau_0)$*, where*

$$\Psi^{\mathrm{cens}}(X, O, Y; \mu_{\mathrm{T}}, \nu_0, \pi_0, g_0, \tau_0) \coloneqq S^{\mathrm{cens}}(X, O, Y; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) - \tau_0,$$

$$S^{\mathrm{cens}}(X, O, Y; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) \coloneqq \frac{\mathbb{1}[O=1]\big(Y - \mu_{\mathrm{T},0}(X)\big)}{\pi_0(1 \mid X)} - \frac{\mathbb{1}[O=0]\big(Y - \nu_0(X)\big)}{g_0(0 \mid X)\pi_0(0 \mid X)}$$
$$+ \frac{g_0(1 \mid X)\mathbb{1}[O=1]\big(Y - \mu_{\mathrm{T},0}(X)\big)}{g_0(0 \mid X)\pi_0(1 \mid X)} + \mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu_0(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X).$$

Here, note that the efficient influence function depends on unknown $\mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0$, which are referred to as *nuisance parameters*. Since the efficient influence function satisfies the equation $\mathbb{E}\left[\Psi^{\mathrm{cens}}(X, O, Y; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0, \tau_0)\right] = 0$, if the nuisance parameters are known and the exact expectation is computed, we can obtain τ_0 by solving for τ_0 that satisfies $\mathbb{E}\left[\Psi^{\mathrm{cens}}(X, O, Y; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0, \tau_0)\right] = 0$. Thus, the efficient influence function provides significant insights for constructing an efficient estimator. Furthermore, the accuracy of the estimation of the nuisance parameters affects the estimation of τ_0, the parameter of interest.

From Theorem 25.20 in van der Vaart (1998), Lemma 4.2 yields the following result about the efficiency bound.

Theorem 4.3 (Efficiency bound in the censoring setting)**.** *If Assumptions 3.2–3.3, 4.1 hold, then the asymptotic variance of any regular estimator is lower bounded by*

$$V^{\mathrm{cens}} \coloneqq \mathbb{E}\left[\Psi^{\mathrm{cens}}(X, O, Y; \mu_0, \nu_0, \pi_0, g_0, \tau_0)^2\right]$$
$$= \mathbb{E}\left[\left(1 - \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\right)^2 \frac{\mathrm{Var}(Y(1) \mid X)}{\pi_0(1 \mid X)} + \frac{\mathrm{Var}(\widetilde{Y} \mid X)}{g_0(0 \mid X)^2 \pi_0(0 \mid X)} + \Big(\tau_0(X) - \tau_0\Big)^2\right].$$

We say that an estimator is efficient if its asymptotic variance aligns with V^{cens}.

4.2 Semiparametric efficient estimator

Based on the efficient influence function, we propose an ATE estimator defined as $\widehat{\tau}_n^{\text{cens-eff}} \coloneqq \frac{1}{n}\sum_{i=1}^n S^{\mathrm{cens}}(X_i, O_i, Y_i; \widehat{\mu}_{\mathrm{T},n,i}, \widehat{\nu}_{n,i}, \widehat{\pi}_{n,i}, \widehat{g}_{n,i})$, where $\widehat{\mu}_{\mathrm{T},n,i}$, $\widehat{\nu}_{n,i}$, $\widehat{\pi}_{n,i}$ and $\widehat{g}_{n,i}$ are estimators of $\mu_{\mathrm{T},0}$, ν_0, π_0, and g_0. Note that the estimators can depend on i. This estimator is an extension of the augmented inverse probability weighting estimator, also called a doubly robust estimator (Bang & Robins, 2005).

Remark (Estimation equation)**.** *There exist several intuitive explanations for* $\widehat{\tau}_n^{\text{cens-eff}}$*. One of the typical explanations is the one from the viewpoint of the estimation equation. Given* $\widehat{\mu}_{\mathrm{T},n,i}$*,* $\widehat{\nu}_{n,i}$*,* $\widehat{\pi}_{n,i}$*, and* $\widehat{g}_{n,i}$*, the estimator* $\widehat{\tau}_n^{\text{cens-eff}}$ *is obtained by solving the following equation:* $\frac{1}{n}\sum_{i=1}^n \Psi^{\mathrm{cens}}(X_i, O_i, Y_i; \widehat{\mu}_{\mathrm{T},n,i}, \widehat{\nu}_{n,i}, \widehat{\pi}_{n,i}, \widehat{g}_{n,i}, \widehat{\tau}_n^{\text{cens-eff}}) = 0$*. Such a derivation of the efficient estimator as the estimation equation approach as explained in (Schuler & van der Laan, 2024).*

4.3 Consistency and double robustness

First, we prove the consistency result; that is, $\widehat{\tau}_n^{\text{cens-eff}} \xrightarrow{\mathrm{p}} \tau_0$ holds as $n \to \infty$. We can obtain this result relatively easily compared to the asymptotic normality. We make the following assumption that holds for most estimators of the nuisance parameters.

Assumption 4.4. *There exist constants* $C_1, C_2 > 0$ *independent of* n *such that* $\widehat{g}_{n,i}(d \mid X), \widehat{\pi}_{n,i}(d \mid X) \in (C_1, 1 - C_1)$ *and* $\widehat{\nu}_{n,i}(X), \widehat{\mu}_{\mathrm{T},n,i}(X) \in [-C_2, C_2]$ *holds almost surely. As* $n \to \infty$*,* $\left\|\widehat{g}_{n,i} - g_0\right\|_2 = o_p(1)$ *holds. Additionally, either of the followings holds for all* $i \in \{1, 2, \ldots, n\}$*:*

- $\left\|\widehat{\pi}_{n,i} - \pi_0\right\|_2 = o_p(1)$.
- $\left\|\widehat{\nu}_{n,i} - \nu_0\right\|_2 = o_p(1)$ *and* $\left\|\widehat{\mu}_{\mathrm{T},n,i} - \mu_{\mathrm{T},0}\right\|_2 = o_p(1)$.

For the estimation of the censoring propensity score, we can employ the existing PU learning methods in the censoring setting, such as Elkan & Noto (2008). Note that we can also apply methods for the case-control PU learning, such as du Plessis et al. (2015), since, as the classification problem, the case-control setting is more general than the censoring setting (Niu et al., 2016). Note that the case-control PU learning methods typically require the class prior $\mathbb{P}(D = 1)$, which can be

Algorithm 1 Cross-fitting in the censoring setting

Input: Observations $\mathcal{D} \coloneqq \left\{\left(X_i, O_i, Y_i\right)\right\}_{i=1}^{n}$, number of folds L, and estimation methods for $\mu_{\mathrm{T},0}, \nu_0, \pi_0$. Let $\mathcal{I} = \{1, 2, \ldots, n\}$ be the index set.
Randomly split $\mathcal{I}$ into L roughly equal-sized folds, $(\mathcal{I}^{(\ell)})_{\ell \in \mathcal{L}}$. Note that $\bigcup_{\ell \in \mathcal{L}} \mathcal{I}^{(\ell)} = \mathcal{I}$.
for $\ell \in \mathcal{L}$ **do**
  Set the training data as $\mathcal{I}^{(-\ell)} = \{1, 2, \ldots, n\} \setminus \mathcal{I}^{(\ell)}$.
  Construct estimators of nuisance parameters on $\mathcal{I}^{(-\ell)}$, denoted by $\widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}$.
end for
Output: Obtain an ATE estimate $\widehat{\tau}_n^{\text{cens-eff}}$ using $\widehat{\mu}_{\mathrm{T},n}^{(\ell)}$, $\widehat{\nu}_n^{(\ell)}$, and $\widehat{\pi}_n^{(\ell)}$.

estimated under several additional assumptions, even if we do not know it (du Plessis & Sugiyama, 2014; Ramaswamy et al., 2016; Kato et al., 2018).

Then, the following consistency result holds. The proof directly follows from the one for Theorem 4.8.

Theorem 4.5 (Consistency in the censoring setting)**.** *If Assumptions 3.2–3.3, and 4.1–4.4 hold, then* $\widehat{\tau}_n^{\text{cens-eff}} \xrightarrow{\mathrm{p}} \tau_0$ *holds as* $n \to \infty$.

Double robustness. There exists double-robustness structure such that given $\left\|\widehat{g}_{n,i} - g_0\right\|_2 = o_p(1)$, if either $\left\|\widehat{\pi}_{n,i} - \pi_0\right\|_2 = o_p(1)$ or $\left\|\widehat{\nu}_{n,i} - \nu_0\right\|_2 = o_p(1)$ and $\left\|\widehat{\mu}_{\mathrm{T},n,i} - \mu_{\mathrm{T},0}\right\|_2 = o_p(1)$ holds, then $\widehat{\tau}_n^{\text{cens-eff}} \xrightarrow{\mathrm{p}} \tau_0$ holds. Here, note that we need to estimate the propensity score consistently to estimate the ATE and the double robustness holds between the estimators of the observation probability π_0 and the expected outcomes $\mu_{\mathrm{T},0}$ and ν_0.[3] This is because, in our setting, the treatment indicator is unobservable. Under this setting, to identify the ATE, we need to use the propensity score and cannot avoid its estimation. Also see Appendix I.3.

4.4 Asymptotic normality

Next, we prove the asymptotic normality. Unlike consistency, we need to make a stronger assumption on the nuisance estimators, especially for the propensity score.

To establish the asymptotic normality or $\sqrt{n}$-consistency of the estimator, it is necessary to control the complexity of the estimators of the nuisance parameter. One of the simplest approaches is to assume the Donsker condition, but it is well known that the Donsker condition does not hold in several cases, such as high-dimensional regression settings. In such cases, asymptotic normality can still be established through sample splitting, a technique in this field (Klaassen, 1987), which has been recently refined by Chernozhukov et al. (2018) as cross-fitting.

Cross-fitting. Cross-fitting is a variant of sample splitting (Chernozhukov et al., 2018). We randomly partition $\mathcal{D}$ into $L > 0$ folds (subsamples), and for each fold $\ell \in \mathcal{L} \coloneqq \{1, 2, \ldots, L\}$, the nuisance parameters are estimated using all other folds. We estimate $\mu_{\mathrm{T},0}, \nu_0, \pi_0$, assuming that the propensity score g_0 is known. Let us denote the estimators in fold $\ell \in \mathcal{L}$ as $\widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}$. Let $\mathcal{I}^{(\ell)}$ be the set of the sample index belonging to fold ℓ.

Various estimation methods can be employed, including neural networks and Lasso, provided they satisfy the convergence rate conditions specified in Assumption 4.7. We later relax the assumption of a known propensity score. It is important to note that issues related to the propensity score estimation cannot be fully addressed even with cross-fitting. The pseudocode is in Algorithm 4.4.

Asymptotic normality. We describe the results only for the case with cross-fitting, but similar results hold for the case when we assume the Donsker condition.

We make the following assumptions.

Assumption 4.6. *The propensity score* g_0 *is known (*$\widehat{g}_{n,i} = g_0$*).*

Assumption 4.7. *For each* $\ell \in \mathcal{L}$*, as* $n \to \infty$*, the followings hold:*

[3] In the standard setting, the double robustness holds between the estimators of the propensity score and the expected outcome.

- $\left\|\pi_0(d \mid X) - \widehat{\pi}_n^{(\ell)}(d \mid X)\right\|_2 = o_p(1)$ *for* $d \in \{1,0\}$, $\left\|\mu_{\mathrm{T},0}(X) - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\right\|_2 = o_p(1)$, *and* $\left\|\nu_0(X) - \widehat{\nu}_n^{(\ell)}(X)\right\|_2 = o_p(1)$.
- $\left\|\pi_0(d \mid X) - \widehat{\pi}_n^{(\ell)}(d \mid X)\right\|_2 \left\|\mu_{\mathrm{T},0}(X) - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\right\|_2 = o_p(n^{-1/2})$ *for* $d \in \{1,0\}$.
- $\left\|\pi_0(0 \mid X) - \widehat{\pi}_n^{(\ell)}(0 \mid X)\right\|_2 \left\|\nu_0(X) - \widehat{\nu}_n^{(\ell)}(X)\right\|_2 = o_p(n^{-1/2})$.

Then, we construct the estimator as $\widehat{\tau}_n^{\text{cens-eff}} := \frac{1}{n}\sum_{\ell\in\mathcal{L}}\sum_{i\in\mathcal{I}^{(\ell)}} S^{\text{cens}}(X_i, O_i, Y_i; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0)$ and show the asymptotic normality holds as follows:

Theorem 4.8 (Asymptotic normality in the censoring setting)**.** *Consider the censoring setting. Suppose that Assumptions 3.2–3.3, 4.1, 4.6–4.7 hold; that is,* $\widehat{g}_{n,i} = g_0$, *and* $\widehat{\mu}_{\mathrm{T},n,i} = \widehat{\mu}_{\mathrm{T},n}^{(\ell)}$, $\widehat{\nu}_{n,i} = \widehat{\nu}_n^{(\ell)}$, *and* $\widehat{\pi}_{n,i} = \widehat{\pi}_n^{(\ell)}$ *are constructed via cross-fitting with certain convergence rates. Then, we have*

$$\sqrt{n}\left(\widehat{\tau}_n^{\text{cens-eff}} - \tau_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V^{\text{cens}}) \ \text{ as } \ n \to \infty.$$

The proof is provided in Appendix I. The asymptotic variance of $\widehat{\tau}_n^{\text{cens-eff}}$ matches the efficiency bound. Therefore, Theorem 4.8 also implies that the estimator $\widehat{\tau}_n^{\text{cens-eff}}$ is asymptotically efficient.

We discuss the other candidates of ATE estimators below.

Remark (Inefficiency of the Inverse Probability Weighting (IPW) estimator)**.** *We can define the IPW estimator as* $\widehat{\tau}_n^{\text{cens-IPW}} := \frac{1}{n}\sum_{i=1}^n \left(\frac{\mathbb{1}[O_i=1]Y_i}{\widehat{\pi}_{n,i}(1|X_i)} - \frac{\mathbb{1}[O_i=0]Y_i}{\widehat{g}_{n,i}(0|X_i)\widehat{\pi}_{n,i}(0|X_i)} + \frac{\widehat{g}_{n,i}(1|X_i)\mathbb{1}[O_i=1]Y_i}{\widehat{g}_{n,i}(0|X_i)\widehat{\pi}_{n,i}(1|X_i)}\right)$. *Compared to our proposed efficient estimator, this estimator does not use the conditional outcome estimators (Horvitz & Thompson, 1952). If* g_0 *and* π_0 *are known, this estimator is unbiased. However, it incurs a large asymptotic variance, given as* $V^{\text{IPW}} := \mathbb{E}\Big[\left(1 - \frac{g_0(1|X)}{g_0(0|X)}\right)^2 \frac{\mathbb{E}[Y(1)^2|X]}{\pi_0(1|X)} + \frac{\mathbb{E}[\widetilde{Y}^2|X]}{g_0(0|X)^2\pi_0(0|X)}\Big]$. *Here, it holds that* $V^{\text{IPW}} \geq V^{\text{cens}}$, *where the equality holds when* $\mu_{\mathrm{T},0}(x) = 0$ *and* $\nu_0 = 0$ *hold for all* x. *Thus, the IPW estimator is inefficient compared to* $\widehat{\tau}_n^{\text{cens-eff}}$. *Additionally, if* π_0 *is unknown, the IPW estimator requires more restrictive conditions for the asymptotic normality than* $\widehat{\tau}_n^{\text{cens-eff}}$.

Remark (Direct Method (DM) estimator)**.** *Another candidate is a DM estimator, defined as* $\widehat{\tau}_n^{\text{cens-DM}} := \widehat{\mu}_{\mathrm{T},n,i}(X) - \frac{1}{\widehat{g}_{n,i}(0|X)}\widehat{\nu}_{n,i}(X) + \frac{\widehat{g}_{n,i}(1|X)}{\widehat{g}_{n,i}(0|X)}\widehat{\mu}_{\mathrm{T},n,i}(X)$, *which is also referred to as a naive plug-in estimator. The asymptotic normality significantly depends on the estimators* $\widehat{\mu}_{\mathrm{T},n,i}$ *and* $\widehat{g}_{n,i}$. *Additionally, the DM estimator is known to be sensitive to model misspecification.*

4.5 Unknown propensity score

We have assumed that the propensity score g_0 is known. This is because we cannot establish $\sqrt{n}$-consistency even if we assume the Donsker condition or employ cross-fitting if g_0 is estimated. However, this assumption can be relaxed by utilizing an additional dataset to estimate g_0.

Several practical scenarios exist. For instance, consider that the following additional dataset is available: $\mathcal{D}^{\text{aux}} := \left\{\left(X_{i'}, O_{i'}\right)\right\}_{i'=1}^{n^{\text{aux}}}$, $\left(X_{i'}, O_{i'}\right) \sim \zeta_0(x)\pi_0(o \mid x)$.

Such a dataset can be less costly since it does not have the outcome data. Let $\widehat{g}_{n^{\text{aux}}}$ be an estimator obtained from $\mathcal{D}^{\text{aux}}$ and consider the following assumption:

Assumption 4.9. *It holds that* $\|\widehat{g}_{n^{\text{aux}}} - g_0\|_2 = o_p(1)$ *as* $n^{\text{aux}} \to \infty$.

If n^{aux} approaches infinity independently of n, under Assumption 3.1, we can establish the asymptotic normality without assuming the propensity score is known.

Corollary 4.10 (Asymptotic normality in the censoring setting)**.** *Consider the censoring setting. Suppose that Assumptions 3.2–3.3, 4.1, 4.7, and 4.9 hold. Then, it holds that* $\sqrt{n}\left(\widehat{\tau}_n^{\text{cens-eff}} - \tau_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V^{\text{cens}})$ *as* $n \to \infty$.

We can also use $\{(X_i, O_i)\}_{i=1}^n$ from $\mathcal{D}$ to estimate g_0 with $\mathcal{D}^{\text{aux}}$. The inclusion of $\{(X_i, O_i)\}_{i=1}^n$ can improve empirical performance.

Another practical scenario involves an auxiliary dataset with treatment indicators and missing outcomes, given as $\mathcal{D}^{\text{aux}'} := \left\{\left(X_{i'}, D_{i'}\right)\right\}_{i'=1}^{n^{\text{aux}}}$, $\left(X_{i'}, D_{i'}\right) \sim \zeta_0(x)g_0(o \mid x)$.

Table 1: Experimental results. Left: censoring setting; Right: case-control setting.

Censoring	IPW	DM	Efficient	IPW	DM	Efficient
	(estimated g_0)			(true g_0)		
MSE	0.31	0.08	0.06	0.06	0.01	0.01
Bias	-0.26	0.16	0.12	-0.06	0.03	0.00
Cov. ratio	0.95	0.07	0.78	1.00	0.09	0.93

Case-control	IPW	DM	Efficient	IPW	DM	Efficient
	(estimated e_0)			(true e_0)		
MSE	10.85	0.07	0.06	0.03	0.01	0.00
Bias	1.44	0.11	0.07	0.00	0.03	0.00
Cov. ratio	0.57	0.61	0.73	0.98	0.95	0.95

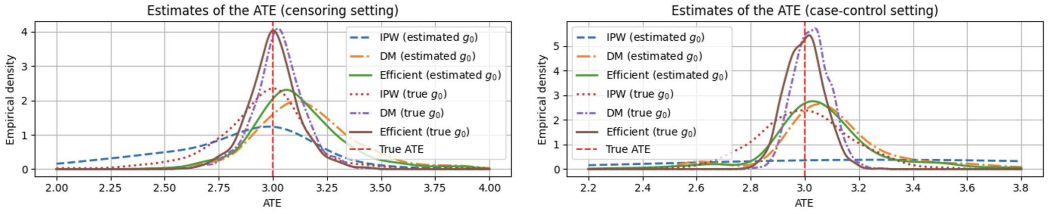

Figure 2: Empirical distributions of ATE estimates.

Remark (Double machine learning)**.** *Semiparametric efficient estimators typically rely on two ingredients: a doubly robust structure and control of function complexity. In our setting, the doubly robust structure links the observation probability and the conditional expected outcome, but it does not involve the propensity score. Specifically, conditional on a consistent estimator of the propensity score, if either the observation probability or the conditional expected outcome is consistently estimated, the ATE can be estimated consistently. The doubly robust structure contributes not only to consistency but also to rapid bias reduction through Neyman orthogonality. In the standard ATE setup, one can remove bias by appropriately estimating the conditional expected outcome and the propensity score and plugging them into a Neyman-orthogonal estimating equation. In PUATE, however, although a doubly robust relation holds between the observation probability and the conditional expected outcome, it does not extend to the propensity score and the remaining nuisance parameters. Consequently, the estimation error of the propensity score may persist. Specifically,* $\sqrt{n}(\widehat{\tau} - \tau) = G + Bias1 + Bias2$*, where* G *is asymptotically normal,* $Bias1$ *is the product of the estimation errors for the observation probability and the conditional expected outcome, and* $Bias2$ *is the estimation error of the propensity score.* $Bias1$ *can be* $o_p(n^{-1/2})$*, but* $Bias2$ *is not* $o_p(n^{-1/2})$ *unless the propensity score is estimated at a rate faster than the standard parametric rate* $O_p(n^{-1/2})$*. This is why our analysis assumes that the propensity score is known or is estimated more rapidly using an independent dataset. When using such an independent dataset, in our proof we take the limit with* $m \to \infty$ *first, followed by* $n \to \infty$*. Alternatively, one can consider an asymptotic regime in which both* n^{aux} *and* n *diverge (i.e.,* $n^{aux}, n \to \infty$*); in that case,* m *must diverge faster than* n*.*

5 Semiparametric efficient ATE estimation under the case-control setting

Here, we briefly introduce the ATE estimator in the case-control setting. More detailed results are shown in Appendix D.

We define $\widehat{\tau}_n^{\text{cc-eff}} := \frac{1}{m}\sum_{j=1}^{m}\left(1 - \frac{\widehat{e}_{n,j}(1|X_j)}{\widehat{e}_{n,j}(0|X_j)}\right)\left(Y_j(1) - \widehat{\mu}_{\mathrm{T},n,j}(X)\right)\widehat{r}_{n,j}(X_j) + \frac{1}{l}\sum_{k=1}^{l}\left(\frac{Y_{\mathrm{U},k} - \widehat{\mu}_{\mathrm{U},n,k}(X_k)}{\widehat{e}_{n,k}(0|X_k)} + \widehat{\mu}_{\mathrm{T},n,k}(X_k) - \frac{\widehat{\mu}_{\mathrm{U},n,k}(X_k)}{\widehat{e}_{n,k}(0|X_k)} + \frac{\widehat{e}_{n,k}(1|X_k)\widehat{\mu}_{\mathrm{T},n,k}(X_k)}{\widehat{e}_{n,k}(0|X_k)}\right)$ as an ATE estimator in the case-control setting. Here, $\widehat{\mu}_{\mathrm{T},n,j}$, $\widehat{\mu}_{\mathrm{U},n,k}$, $\widehat{e}_{n,j}$, and $\widehat{r}_{n,j}$ are estimators of $\mu_{\mathrm{T},0}$, $\mu_{\mathrm{U},0}$, e_0, and r_0, where m and l denote the dependence on each dataset.

For the estimator, we show the following theorem, which is an informal version of Theorem D.8

Theorem 5.1 (Asymptotic normality in the case-control setting (Informal))**.** *Fix* $\alpha \in (0,1)$*. For* $n > 0$*, consider the case-control setting with sample sizes* m, l *such that* $m = \alpha n$ *and* $l = (1-\alpha)n$*. If the case-control propensity score* e_0 *and the density ratio are known (*$\widehat{e}_{n,i} = e_0$ *and* $\widehat{r}_{n,i} = r_0$*), and* $\widehat{\mu}_{\mathrm{T},n,i} = \widehat{\mu}^{(\ell)}_{\mathrm{T},m}$*,* $\widehat{\mu}_{\mathrm{U},n,i} = \widehat{\mu}^{(\ell)}_{\mathrm{U},l}$ *are consistent estimators constructed via cross-fitting. Then, under regularity conditions (see Theorem D.8), we have* $\sqrt{n}\left(\widehat{\tau}_n^{\text{cc-eff}} - \tau_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V^{\mathrm{cc}})$ *as* $n \to \infty$*, where* $V^{\mathrm{cc}} > 0$ *is the efficiency bound defined in Theorem D.3.*

6 Simulation studies

This section investigates the empirical performance of the proposed estimators. We also show the experimental results using semi-synthetic data in Appendix M.

6.1 Censoring setting

We generate synthetic data under the censoring setting, where the covariates X are drawn from a multivariate normal distribution as $X \sim \zeta_0(x)$, where $\zeta_0(x)$ is the density of $\mathcal{N}(0, I_p)$, and I_p denotes the $(p \times p)$ identity matrix. We set $p = 3$. Set $\mathbb{P}(D \mid X) = \mathrm{trunc}(\mathrm{sigmoid}(X^\top \beta), 0.1, 0.9)$, where β is a coefficient sampled from $\mathcal{N}(0, 0.5 I_p)$, and $\mathrm{trunc}(t, a, b)$ truncates t by a and b ($a < b$). Treatment D is sampled from the probability. The observation indicator O is generated from a Bernoulli distribution with probability c if $D_i = 1$ and $O_i = 0$ if $D_i = 0$. Here, c is generated from a uniform distribution with support $[0, 1]$. The outcome is generated as $Y = X^\top \beta + 1.1 + \tau_0 \cdot D + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$, where we set $\tau_0 = 3$.

The nuisance parameters are estimated using linear regression and (linear) logistic regression. We compared our proposed estimator, $\widehat{\tau}_n^{\text{cens-eff}}$, with the other candidates, the IPW estimator $\widehat{\tau}_n^{\text{cens-IPW}}$ and the DM estimator $\widehat{\tau}_n^{\text{cens-DM}}$, defined in Remarks 4.4 and 4.4, respectively. Note that all of these estimators are proposed by us, and our goal is not to confirm $\widehat{\tau}_n^{\text{cens-eff}}$ outperforms the others, while our recommendation is $\widehat{\tau}_n^{\text{cens-eff}}$. We consider both cases where the propensity score is either estimated using the method proposed by Elkan & Noto (2008) or assumed to be known.

We set $n = 3000$. We conduct 5000 trials and report the empirical mean squared errors (MSEs) and biases for the true ATE and the coverage ratio (Cov. ratio) computed from the confidence intervals in Table 1. We also present the empirical distributions of the ATE estimates in Figure 2.

As the theory suggests, $\widehat{\tau}_n^{\text{cens-eff}}$ exhibits smaller MSEs compared to other methods. Interestingly, when the propensity score is estimated, the MSEs decrease, a phenomenon reported in existing studies. The coverage ratio is also accurate. The empirical distribution of the ATE estimates demonstrates the asymptotic normality.

6.2 Case-control setting

In the case-control setting, covariates for the treatment and unknown groups are generated from different p-dimensional normal distributions: $X_{\mathrm{T}} \sim \zeta_{\mathrm{T},0}(x)$ and $X \sim \zeta_0(x) = e_0(1)\zeta_{\mathrm{T},0}(x) + e_0(0)\zeta_{\mathrm{C}}(x)$, where we set $p = 3$, $\zeta_{\mathrm{T},0}(x)$ and $\zeta_{\mathrm{C}}(x)$ are the densities of normal distributions $\mathcal{N}(\mu_p \mathbf{1}_p, I_p)$ and $\mathcal{N}(\mu_n \mathbf{1}_p, I_p)$, $\mu_p = 0.5$ and $\mu_n = 0$, $\mathbf{1}_p = (1\ 1\ \cdots\ 1)^\top$, and $e_0(1)$ is the class prior set as $e_0(1) = 0.3$. By definition, the propensity score $e_0(d \mid x)$ is given as $e_0(1 \mid x) = e_0(1)\zeta_{\mathrm{T},0}(x)/\zeta_0(x)$. The outcome is generated similarly to the censoring setting $Y = X^\top \beta + 1.1 + \tau_0 D + \varepsilon$, where $\tau_0 = 3$.

We set $m = 1000$ and $l = 2000$ and compute the same evaluation metrics as in the censoring setting. Although logistic regression is used, the propensity score model is misspecified, while the expected conditional outcome follows a linear model.

Overall, $\widehat{\tau}_n^{\text{cc-eff}}$ demonstrates robust performance in terms of MSE, bias, and coverage ratio. The poor performance of the IPW estimator is attributed to model misspecification.

We investigate non-linear settings in Appendix L.

7 Conclusion

In this study, we investigated PUATE, the problem of ATE estimation in the presence of missing treatment indicators. We formulated the problem using the censoring and case-control settings, inspired by PU learning. For each setting, we derived the efficiency bound and developed an efficient estimator. Our analysis revealed that achieving asymptotic normality and efficiency. Future research directions include extending our approach to the semi-supervised setting, handling additional missing values, and the relaxation of assumptions regarding the missingness mechanism.

References

Jaeil Ahn, Bhramar Mukherjee, Stephen B. Gruber, and Samiran Sinha. Missing exposure data in stereotype regression model: Application to matched case–control study with disease subclassification. *Biometrics*, 67(2):546–558, 2011.

Ahmed M. Alaa and Mihaela van der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. In *Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3427–3435. Curran Associates Inc., 2017.

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

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data under the selected at random assumption. In *Proceedings of the Second International Workshop on Learning with Imbalanced Domains: Theory and Applications*, 2018.

Jessa Bekker and Jesse Davis. Learning from positive and unlabeled data: a survey. *Machine Learning*, 109(4):719–760, 2020.

Sandra E. Black, Paul J. Devereux, and Kjell G. Salvanes. Staying in the classroom and out of the maternity ward? the effect of compulsory schooling laws on teenage births. *The Economic Journal*, 118(530):1025–1054, 2008.

David Bruns-Smith, Oliver Dukes, Avi Feller, and Elizabeth L Ogburn. Augmented balancing weights as linear regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 04 2025.

Abhishek Chakrabortty and Guorong Dai. A general framework for treatment effect estimation in semi-supervised and high dimensional settings, 2024. arXiv:2201.00468.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 2018.

Victor Chernozhukov, Whitney K. Newey, and Rahul Singh. Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027, 2022.

Victor Chernozhukov, Whitney K. Newey, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via riesz regression, 2024. arXiv:2104.14737.

Alicia Curth and Mihaela van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

Issa J Dahabreh, Sarah E Robertson, Eric J Tchetgen, Elizabeth A Stuart, and Miguel A Hernán. Generalizing causal inferences from individuals in randomized trials to all trial-eligible individuals. *Biometrics*, 75(2):685–694, 2019.

Marthinus Christoffel du Plessis and M. Sugiyama. Class prior estimation from positive and unlabeled data. *IEICE Transactions on Information and Systems*, E97-D(5):1358–1362, 2014.

Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Analysis of learning from positive and unlabeled data. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 703–711, 2014.

Marthinus Christoffel du Plessis, Gang. Niu, and Masashi Sugiyama. Convex formulation for learning from positive and unlabeled data. In *International Conference on Machine Learning (ICML)*, 2015.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, 2008.

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

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

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2020.

Jerry Hausman. Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives*, 15(4), 2001.

James Heckman. Shadow prices, market wages, and labor supply. *Econometrica*, 42(4):679–694, 1974.

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

Masayuki Henmi and Shinto Eguchi. A paradox concerning nuisance parameters and projected estimating functions. *Biometrika*, 2004.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Keisuke Hirano, Guido Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 2003.

Daniel G. Horvitz and Donovan J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Yu-Guan Hsieh, Gang Niu, and Masashi Sugiyama. Classification from positive, unlabeled and biased negative data. In *International Conference on Machine Learning (ICML)*, 2019.

Kosuke Imai and Marc Ratkovic. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443 – 470, 2013.

Guido W. Imbens and Tony Lancaster. Efficient estimation and stratified sampling. *Journal of Econometrics*, 74(2):289–318, 1996.

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

Guido W. Imbens and Jeffrey M. Wooldridge. Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, 47(1):5–86, 2009.

Fredrik D. Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029, 2016.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009.

Masahiro Kato. Direct bias-correction term estimation for propensity scores and average treatment effect estimation, 2025. arXiv: 2509.22122.

Masahiro Kato and Takeshi Teshima. Non-negative bregman divergence minimization for deep direct density ratio estimation. In *International Conference on Machine Learning (ICML)*, 2021.

Masahiro Kato, Liyuan Xu, Gang Niu, and Masashi Sugiyama. Alternate estimation of a classifier and the class-prior from positive and unlabeled data, 2018. arXiv:1809.05710.

Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *International Conference on Learning Representations (ICLR)*, 2019.

Masahiro Kato, Takuya Ishihara, Junya Honda, and Yusuke Narita. Efficient adaptive experimental design for average treatment effect estimation, 2020. arXiv:2002.05308.

Masahiro Kato, Kenichiro McAlinn, and Shota Yasui. The adaptive doubly robust estimator and a paradox concerning logging policy. In *International Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

Masahiro Kato, Akihiro Oga, Wataru Komatsubara, and Ryo Inokuchi. Active adaptive experimental design for treatment effect estimation with covariate choice. In *International Conference on Machine Learning (ICML)*, 2024.

Edward H. Kennedy. Semiparametric theory and empirical processes in causal inference, 2016. arXiv: 1510.04740.

Edward H. Kennedy. Efficient nonparametric causal inference with missing exposure information. *The International Journal of Biostatistics*, 16(1), 2020.

Edward H. Kennedy. Semiparametric doubly robust targeted double machine learning: a review, 2023. arXiv: 2203.06469.

Edward H. Kennedy, Sivaraman Balakrishnan, James M. Robins, and Larry Wasserman. Minimax rates for heterogeneous causal effect estimation. *The Annals of Statistics*, 52(2):793 – 816, 2024.

Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. Positive-unlabeled learning with non-negative risk estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Chris A. J. Klaassen. Consistent estimation of the influence function of locally asymptotically linear estimators. *Annals of Statistics*, 15, 1987.

Milan Kuzmanovic, Tobias Hatt, and Stefan Feuerriegel. Estimating conditional average treatment effects with missing treatment information. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Tony Lancaster and Guido Imbens. Case-control studies with contaminated controls. *Journal of Econometrics*, 71(1):145–160, 1996.

Arthur Lewbel. Estimation of average treatment effects with misclassification. *Econometrica*, 75 (2):537–551, 2007.

Roderick Little and Donald B. Rubin. *Statistical analysis with missing data*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. Wiley, 2002.

Aprajit Mahajan. Identification and estimation of regression models with misclassification. *Econometrica*, 74(3):631–665, 2006.

Charles F. Manski. Identification problems in the social sciences. *Sociological Methodology*, 23: 1–56, 1993.

Charles F. Manski. *Partial Identification in Econometrics*, pp. 178–188. Palgrave Macmillan UK, 2010.

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

Jerzy Neyman. Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes. *Statistical Science*, 5:463–472, 1923.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108, 2020.

Gang Niu, Marthinus Christoffel du Plessis, Tomoya Sakai, Yao Ma, and Masashi Sugiyama. Theoretical comparisons of positive-unlabeled learning against positive-negative learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embeddings of distributions. In *International Conference on Machine Learning (ICML)*, 2016.

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*, 89:846–866, 1994.

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

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

Tomoya Sakai, Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. Semi-supervised classification based on classification from positive and unlabeled data. In *International Conference on Machine Learning (ICML)*, 2017.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4), 2020.

Alejandro Schuler and Mark van der Laan. Introduction to modern causal inference, 2024. URL `https://alejandroschuler.github.io/mci/introduction-to-modern-causal-inference.html`.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning (ICML)*, pp. 3076–3085, 2017.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2019.

Dan Steinberg and N. Scott Cardell. Estimating logistic regression models when the dependent variable has no variance. *Communications in Statistics - Theory and Methods*, 21(2):423–450, 1992.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Masashi Sugiyama, Han Bao, Takashi Ishida, Nan Lu, and Tomoya Sakai. *Machine Learning from Weak Supervision: An Empirical Risk Minimization Approach (Adaptive Computation and Machine Learning series)*. The MIT Press, 2022.

Paweł Teisseyre, Timo Martens, Jessa Bekker, and Jesse Davis. Learning from biased positive-unlabeled data via threshold calibration. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2025.

A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer Series in Statistics. Springer New York, 2007.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.

Masatoshi Uehara, Masahiro Kato, and Shota Yasui. Off-policy evaluation and learning for external validity under a covariate shift. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

van der Laan. Targeted maximum likelihood learning, 2006. U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 213. `https://biostats.bepress.com/ucbbiostat/paper213/`.

Aad W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

Aad W. van der Vaart. Semiparametric statistics, 2002. URL `https://sites.stat.washington.edu/jaw/COURSES/EPWG/stflour.pdf`.

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

Jeffrey M. Wooldridge. Asymptotic properties of weighted m-estimation for standard stratified samples. *Econometric Theory*, 2001.

Ikko Yamane, Florian Yger, Jamal Atif, and Masashi Sugiyama. Uplift modeling from separate labels. In *International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 9949–9959. Curran Associates Inc., 2018.

Yachong Yang, Arun Kumar Kuchibhotla, and Eric Tchetgen Tchetgen. Doubly robust calibration of prediction sets under covariate shift. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):943–965, 03 2024.

Zhiwei Zhang, Wei Liu, Bo Zhang, Li Tang, and Jun Zhang. Causal inference with missing exposure information: Methods and applications to an obstetric study. *Statistical methods in medical research*, 25:1003–1014, 12 2013.

Anqi Zhao and Peng Ding. To adjust or not to adjust? estimating the average treatment effect in randomized experiments with missing covariates. *Journal of the American Statistical Association*, 119(545):450–460, 2024.

Qingyuan Zhao. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47(2):965 – 993, 2019.

Wenjing Zheng and Mark J van der Laan. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer New York, NY, 2011.

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

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Both the abstract and introduction clearly state our contributions and the important assumption that we introduce. For the summary of our contributions, see Section 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: For example, our proposed method requires the convergence rate for the estimators of the nuisance parameters. We explain this limitation in Sections 4 and 5, and other related Sections.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We describe the details of the assumptions. For example, for the censoring setting, see Sections 3–4.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We explain the details of the experimental settings. For example, see Section 6 and Appendix L.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will organize and provide the experimental code until the camera-ready.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: Our experiment does not use test data since our interest is to estimate the ATE (point).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes] Replace by [Yes] , [No] , or [NA] .

   Justification: We provide the MSE, bias, coverage ratios, and the empirical distributions of the ATE estimates.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: Our experiment is not computationally hard, and we only use a personal MacBook Pro PC

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper conforms, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our paper may have a societal impact since we consider applications to marketing and medicine. However, the impact is within the standard data science and not so serious.

    Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our study poses no such risks

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use the IHDP dataset with citing Hill (2011).

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our study does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our study does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We only use LLM for the grammatical check.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Related work

The ATE estimation problem has long been studied in statistics, epidemiology, economics, and machine learning (Imbens & Rubin, 2015). While randomized controlled trials are considered the gold standard, it is extremely important to estimate the ATE in observational studies. In ATE estimation with observational data, one of the basic approaches is to employ the IPW estimator, which allows us to correct for selection bias (Horvitz & Thompson, 1952).

Although the IPW estimator is a powerful tool, it is known that its asymptotic variance exceeds the efficiency bound even when the true propensity score is used (Hahn, 1998).[4]

Another powerful estimator in this context is the doubly robust estimator (Bang & Robins, 2005), which also plays an important role in the literature on missing data (Yang et al., 2024). The doubly robust estimator not only satisfies the double robustness property but also achieves asymptotic efficiency; that is, its asymptotic variance reaches the efficiency bound (Hahn, 1998). This property is closely related to the efficient influence function in the derivation of the efficiency bound (van der Vaart, 1998; Tsiatis, 2007). The doubly robust estimator is defined as a sample average of the efficient influence function with estimated nuisance parameters.

The doubly robust estimator has been refined in several directions. For example, van der Laan (2006) propose the targeted maximum likelihood framework, which can improve the finite-sample performance of the efficient ATE estimator. Chernozhukov et al. (2018) propose a debiased machine learning framework, which is further generalized as automatic debiased machine learning (Chernozhukov et al., 2022) and Riesz regression (Chernozhukov et al., 2024). Covariate balancing scores are also important attempts to improve performance (Imai & Ratkovic, 2013; Hainmueller, 2012; Zubizarreta, 2015; Zhao, 2019), and Bruns-Smith et al. (2025) and Kato (2025) show the equivalence between covariate balancing methods and Riesz regression.

To construct efficient estimators, convergence rate conditions and complexity restrictions for the nuisance estimators are usually required (Schuler & van der Laan, 2024). In particular, to satisfy the complexity restrictions, researchers often assume the Donsker condition or apply sample splitting (van der Vaart, 2002; Klaassen, 1987; Zheng & van der Laan, 2011). These approaches are further developed in the double machine learning framework by Chernozhukov et al. (2018), where convergence rate conditions are relaxed through the use of Neyman orthogonality, and complexity restrictions are addressed via sample splitting, known as cross-fitting. For discussions of the relationship between double machine learning and other frameworks, such as targeted maximum likelihood estimation, see Kennedy (2016, 2023).

It is important to note that Neyman orthogonality and cross-fitting play different roles. Cross-fitting is used to ensure that the nuisance estimators (e.g., propensity score, outcome models) are independent of the observations to which they are applied. However, cross-fitting alone is not sufficient to guarantee asymptotic normality. The issue is that nuisance estimators typically converge at rates slower than $\sqrt{n}$. In doubly robust estimation, the convergence rate conditions for the nuisance estimators are relaxed due to the doubly robust structure, which is also referred to as Neyman orthogonality.

Our work builds upon these arguments. However, in our setting, we cannot apply the techniques from Chernozhukov et al. (2018) to mitigate the convergence rate condition for the propensity score. In other words, our estimator is sensitive to the accuracy of the propensity score estimation. Therefore, we assume that the propensity score is known in order to derive asymptotic normality, although consistency can still be achieved when the propensity score is estimated.

CATE estimation is also an important topic related to this study (Heckman et al., 1997). Various methods have been proposed for estimating CATE including methods using neural networks (Johansson et al., 2016; Shalit et al., 2017; Shi et al., 2019; Hassanpour & Greiner, 2020; Curth & van der Schaar, 2021), gaussian process (Alaa & van der Schaar, 2017), and tree-based approaches (Wager & Athey, 2018). A critical perspective in recent literature is minimax optimality. Kennedy et al. (2024) proposes a minimax optimal CATE estimator based on the R-learner (Nie & Wager, 2020), by deriving a minimax lower bound (Tsybakov, 2008). While several di-

[4]Under certain conditions, using an estimated propensity score can reduce the asymptotic variance, as shown by Hirano et al. (2003). This phenomenon is known as the paradox of the propensity score (Henmi & Eguchi, 2004; Kato et al., 2021).

rections exist for extending our results to CATE estimation, deriving a minimax optimal CATE estimator would require further theoretical analysis, which is beyond the scope of this study.

This study employs arguments for statistical analysis under stratified sampling (Wooldridge, 2001), which includes PU learning as a special case (Imbens & Lancaster, 1996). Another example of stratified sampling is covariate shift; Uehara et al. (2020) study causal inference when the training data contain (covariates, treatment, outcomes) but the evaluation data contain only covariates.

The Heckman model (Heckman, 1974) and the framework of Robins et al. (1994) are important contributions to causal inference with missing data. The Heckman model addresses endogeneity between the outcome and the observation indicator, whereas our analysis does not consider such endogeneity. Our setting is therefore simpler in that respect, yet more challenging because the relationship between treatment assignment and observation follows the PU learning mechanism. Similarly, Kennedy (2020) allow endogeneity between treatment observation and the outcome, but the observation mechanism itself is simpler than in our work. These approaches are not nested within one another; rather, they are complementary, and combining them may yield more flexible statistical modeling.

A.1 ATE estimation with missing data

ATE estimation under missing values has been extensively studied, as the standard ATE estimation setting is itself closely related to the literature on missing data (Rubin, 1976; Bang & Robins, 2005).

In this context, various assumptions can be made about how data is missing. For example, some studies consider settings with missing covariates (Zhao & Ding, 2024). This study, however, focuses on the case in which covariates are fully observed and treatment assignment is missing. In the problem of missing treatments, Molinari (2010) presents several examples from survey analysis. Ahn et al. (2011) investigates the effect of physical activity on colorectal cancer using data in which treatment is missing for about 20% of the units. Zhang et al. (2013) estimates infant weight outcomes where the treatment—mother's body mass index (BMI)—is missing for about half of the sample. Kennedy (2020) develops a general framework for handling settings where both the observation indicator and the treatment indicator are separately observable. In contrast, in our case, we can observe only the product of the observation and treatment indicators, implying less available information than in Kennedy (2020). Kuzmanovic et al. (2023) proposes a method for conditional ATE estimation under this weaker setting.

Our problem is also related to ATE estimation from misclassified data (Lewbel, 2007). Early econometric studies focused on continuous regressors (Hausman, 2001). With regard to binary variables, Mahajan (2006) analyzes misclassification in regression models, while Lewbel (2007) develops methods for identifying and estimating ATEs under potentially misclassified treatment indicators. Researchers have also explored partial identification approaches when the exact misclassification process is unknown, providing bounds on parameters rather than point estimates (Manski, 1993, 2010). In applied settings, validation data have been used to refine causal effect estimates under potential misclassification (Black et al., 2008), demonstrating that even modest errors in treatment indicators can significantly impact policy conclusions. Yamane et al. (2018) also addresses a related problem.

Finally, we refer to semi-supervised treatment effect estimation (Chakrabortty & Dai, 2024), which primarily considers a scenario where two datasets are available: one with complete data and the other with only treatment indicators D and covariates X but no outcome data. Although the setting is not directly related, integrating insights from both areas could enhance the applicability.

A.2 Introduction of PU learning algorithms

Another related body of work comes from the literature on PU learning. PU learning is a classification method primarily designed for binary classifiers (though it can be extended to multi-class settings) in the presence of missing data. Its origins trace back to case-control studies with contaminated controls (Steinberg & Cardell, 1992; Lancaster & Imbens, 1996), which are refined in du Plessis et al. (2015) under the term case-control PU learning. In parallel, Elkan & Noto (2008) investigates PU learning in the context of the censoring setting. One of the main applications of PU learning is learning from implicit feedback, which commonly arises in marketing and recommender

systems. In such settings, user actions—such as product purchases—are observable, but non-actions do not necessarily imply disinterest in the products; therefore, we might suffer bias in a classifier for predicting the users interests if we train it using such data with regarding the action and non-action as positive and negative data. As discussed in the introduction, we consider a similar application. However, our goal is to estimate treatment effects, rather than to train a classifier.

B Reformulation of the censoring and the case-control settings

This section provides a reformulation of the case-control and censoring setting to deepen our understanding. Note that the formulation described in this section is mathematically equivalent to the ones in Sections 4 and 2.4.

B.1 Reformulation of the censoring setting

We can introduce the censoring setting with the following story:

- For each i, a sample (X_i, D_i, Y_i) is generated.
- A coin is tossed, and $\widetilde{O}_i = 1$ if it lands heads, or $\widetilde{O}_i = 0$ if it lands tails.
  - If $\widetilde{O}_i = 1$ and $D_i = 1$ (that is, $\widetilde{O}_i D_i = 1$), then we observe the treatment indicator $D_i = \widetilde{O}_i = 1$.
  - Otherwise, the treatment indicator is not observed, and we only observe (X_i, Y_i).
- Finally, we observe $(X_i, \widetilde{O}_i D_i, Y_i)$.

By denoting $\widetilde{O}_i D_i$ by O_i, we can obtain the same formulation in Section 2.3.

Remark (Kennedy (2020))**.** *For example, Kennedy (2020) considers the missing treatment information, which is essentially different from ours. Kennedy (2020) considers the following setup:*

- *For each i, a sample (X_i, D_i, Y_i) is generated, where X_i denotes covariates, D_i the treatment indicator, and Y_i the outcome.*
- *A coin is tossed, and $\widetilde{O}_i = 1$ if it lands heads, or $\widetilde{O}_i = 0$ if it lands tails.*
  - *If $\widetilde{O}_i = 1$, we observe the treatment indicator $D_i \in 1, 0$ along with (X_i, Y_i).*
  - *If $\widetilde{O}_i = 0$, the treatment indicator is unobserved, and we only observe (X_i, Y_i).*
- *Finally, we observe $(X_i, \widetilde{O}_i, \widetilde{O}_i D_i, Y_i)$.*

For each i, a sample (X_i, D_i, Y_i) is generated, where X_i denotes covariates, D_i the treatment indicator, and Y_i the outcome.

Thus, while Kennedy (2020) observes $(X_i, \widetilde{O}_i, \widetilde{O}_i D_i, Y_i)$, we observe only $(X_i, \widetilde{O}_i D_i, Y_i)$. In our case, $\widetilde{O}_i$ itself is also missing (note again that $(X_i, \widetilde{O}_i D_i, Y_i)$ is equivalent to (X_i, O_i, Y_i), where $O_i = \widetilde{O}_i D_i$). That is, in our case, we can observe only a subset of treatment labels with $D_i = 1$, while the remaining labels are missing and consist of a mixture of $D_i = 1$ and $D_i = 0$. In contrast, Kennedy (2020) can observe both $D_i = 1$ and $D_i = 0$ when the label is observed. Our setting is designed to be more suitable for applications in marketing and recommendation systems, where implicit feedback is common[5].

Note that in our study, we do not explicitly use $\widetilde{O}_i$ but instead denote $\widetilde{O}_i D_i$ by another random variable O_i; that is, $O_i = \widetilde{O}_i D_i$.

In machine learning terminology, we believe that the setup in Kennedy (2020) is close to semi-supervised learning, where both a fully labeled dataset (X_i, L_i) $(i = 1, 2, .., n)$ and an unlabeled

[5]Regarding the missingness mechanism of $\widetilde{O}_i$, Kennedy (2020) considers a more general setting than ours by allowing $\widetilde{O}_i$ to depend on the outcome Y_i. In contrast, while we do not allow such dependence, we use less information about the treatment indicator than Kennedy (2020), as explained above. Thus, we cannot say which setting is more general.

dataset X_j $(j = 1, 2, .., m)$ are available (Note, however, that in Kennedy (2020), whether a data point is labeled is itself a random event, whereas typical semi-supervised learning assumes a deterministic labeling process). PU learning is generally considered a distinct setting from semi-supervised learning though they are related. For a discussion of the relationship between these settings, see Sakai et al. (2017).

B.2 Reformulation of the case-control setting

We can introduce the case-control setting with the following story:

- There are two groups: the treatment group and the unknown group:
  - Treatment group:
    - For each j, a sample $(X_{\mathrm{T},j}, Y_j(1))$ is generated and observed by us.
  - Unknown group:
    - For each k, a sample (X_i, D_i, Y_i) is generated.
    - We observe (X_i, Y_i).

C Details of the examples of ATE estimators in censoring setting

This section provides the details of the examples shown in Section 3.

C.1 Example ATE estimator in the censoring setting

In the censoring setting, as we explained in Section 3, we can obtain an ATE estimator by replacing the following two quantities with sample approximation and nuisance estimators:

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[\frac{\mathbb{1}[O=1]Y}{\pi_0(1 \mid X)}\right],$$
$$\mathbb{E}[Y(0)] = \mathbb{E}\left[\frac{\mathbb{1}[O=0]Y}{g_0(0 \mid X)\pi_0(0 \mid X)}\right] - \mathbb{E}\left[\frac{g_0(1 \mid X)\mathbb{1}[O=1]Y}{g_0(0 \mid X)\pi_0(1 \mid X)}\right].$$

Such an estimator can be defined as follows:

$$\widehat{\tau}_n^{\text{cens-IPW}} := \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}[O_i=1]Y_i}{\widehat{\pi}_n(1 \mid X_i)} - \frac{\mathbb{1}[O_i=0]Y_i}{\widehat{g}_n(0 \mid X_i)\widehat{\pi}_{n,i}(0 \mid X_i)} + \frac{\widehat{g}_{n,i}(1 \mid X_i)\mathbb{1}[O_i=1]Y_i}{\widehat{g}_{n,i}(0 \mid X_i)\widehat{\pi}_{n,i}(1 \mid X_i)}\right),$$

where $\widehat{\pi}_{n,i}$ and $\widehat{g}_n$ are estimators of π_0 and g_0. We refer to this estimator as the inverse probability weighting (IPW) estimator, which is also shown in Remark 4.4

For this estimator, we can show the following theorem.

Theorem C.1. *Suppose that Assumptions 3.2–3.3, and 4.4. If $\|\widehat{\pi}_{n,i}(d \mid X) - \pi_0(d \mid X)\|_2 = o_p(1)$ and $\|\widehat{g}_n(d \mid X) - g_0(d \mid X)\|_2 = o_p(1)$ hold as $n \to \infty$ for $d \in \{1, 0\}$ then $\widehat{\tau}_n^{\text{cens-IPW}} \xrightarrow{\text{p}} \tau_0$ holds as $n \to \infty$.*

Proof. We have

$$\begin{aligned}
\widehat{\tau}_n^{\text{cens-IPW}} &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}[O_i=1]Y_i}{\widehat{\pi}_n(1 \mid X_i)} - \frac{\mathbb{1}[O_i=0]Y_i}{\widehat{g}_n(0 \mid X_i)\widehat{\pi}_{n,i}(0 \mid X_i)} + \frac{\widehat{g}_{n,i}(1 \mid X_i)\mathbb{1}[O_i=1]Y_i}{\widehat{g}_{n,i}(0 \mid X_i)\widehat{\pi}_{n,i}(1 \mid X_i)}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}[O_i=1]Y_i}{\pi_0(1 \mid X_i)} - \frac{\mathbb{1}[O_i=0]Y_i}{g_0(0 \mid X_i)\pi_0(0 \mid X_i)} + \frac{g_0(1 \mid X_i)\mathbb{1}[O_i=1]Y_i}{g_0(0 \mid X_i)\pi_0(1 \mid X_i)}\right) \\
&\quad - \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}[O_i=1]Y_i}{\pi_0(1 \mid X_i)} - \frac{\mathbb{1}[O_i=0]Y_i}{g_0(0 \mid X_i)\pi_0(0 \mid X_i)} + \frac{g_0(1 \mid X_i)\mathbb{1}[O_i=1]Y_i}{g_0(0 \mid X_i)\pi_0(1 \mid X_i)}\right)\right) \\
&\quad + \left(\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}[O_i=1]Y_i}{\widehat{\pi}_n(1 \mid X_i)} - \frac{\mathbb{1}[O_i=0]Y_i}{\widehat{g}_n(0 \mid X_i)\widehat{\pi}_{n,i}(0 \mid X_i)} + \frac{\widehat{g}_{n,i}(1 \mid X_i)\mathbb{1}[O_i=1]Y_i}{\widehat{g}_{n,i}(0 \mid X_i)\widehat{\pi}_{n,i}(1 \mid X_i)}\right)\right)
\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}[O_i=1]Y_i}{\pi_0(1 \mid X_i)} - \frac{\mathbb{1}[O_i=0]Y_i}{g_0(0 \mid X_i)\pi_0(0 \mid X_i)} + \frac{g_0(1 \mid X_i)\mathbb{1}[O_i=1]Y_i}{g_0(0 \mid X_i)\pi_0(1 \mid X_i)}\right) + o_p(1).$$

Here, from the law of large numbers, we have

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{1}[O_i=1]Y_i}{\pi_0(1 \mid X_i)} \xrightarrow{\mathrm{p}} \mathbb{E}\left[\frac{\mathbb{1}[O=1]Y}{\pi_0(1 \mid X)}\right] = \mathbb{E}[Y(1)]$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}[O_i=0]Y_i}{g_0(0 \mid X_i)\pi_0(0 \mid X_i)} - \frac{g_0(1 \mid X_i)\mathbb{1}[O_i=1]Y_i}{g_0(0 \mid X_i)\pi_0(1 \mid X_i)}\right)$$

$$\xrightarrow{\mathrm{p}} \mathbb{E}\left[\frac{\mathbb{1}[O=0]Y}{g_0(0 \mid X)\pi_0(0 \mid X)}\right] - \mathbb{E}\left[\frac{g_0(1 \mid X)\mathbb{1}[O=1]Y}{g_0(0 \mid X)\pi_0(1 \mid X)}\right] = \mathbb{E}[Y(0)].$$

Thus, the proof is complete. □

C.2 Example ATE estimator in the case-control setting

In the case-control setting, PU learning methods have been investigated by Imbens & Lancaster (1996) and du Plessis et al. (2015). In that works, we typically make the following assumption, which corresponds to the SCAR assumption in the censoring setting.

Assumption C.2. *It holds that* $\zeta_{\mathrm{T},0}(x) = \zeta_0(x \mid D=1)$*, where* $\zeta_0(x \mid D=d) = \frac{e_0(d|x)\zeta_0(x)}{e_0(d)}$.

This assumption is also attempted to be relaxed by existing work, such as Kato et al. (2019) and Hsieh et al. (2019) introduce their approaches. As well as the censoring setting, various relaxations exist depending on the application, and there are trade-offs between the strengths of assumptions and identification (Manski, 1993).

We further make the following assumptions.

Assumption C.3 (Unconfoundedness in the case-control setting)**.** *The potential outcomes* $(Y(1), Y(0))$ *are independent of treatment assignment given covariates:*

$$(Y(1), Y(0)) \perp\!\!\!\perp D \mid X.$$

Assumption C.4 (Common support in the case-control setting)**.** *There exists a constant* c *independent of* n *such that for all* $x \in \mathcal{X}$*,* $e_0(d \mid x), \zeta_{\mathrm{T},0}(x), \zeta_0(x) > c$ *hold.*

Under these assumptions, the ATE τ_0 is estimable by replacing the following two quantities with sample approximation:

$$\mathbb{E}[Y(1)] = \mathbb{E}\left[Y r_0(X)\right]$$

and

$$\mathbb{E}[Y(0)] = \mathbb{E}\left[\frac{Y}{e_0(0 \mid X)}\right] - \mathbb{E}\left[\frac{e_0(1 \mid X)Y}{e_0(0 \mid X)}\right],$$

where recall that $r_0(X) = \frac{\zeta_0(X)}{\zeta_{\mathrm{T},0}(X)}$, e_0 can be estimated using PU learning methods, and expectations can be approximated by sample averages.

D Semiparametric efficient ATE estimation under the case-control setting

In this section, we consider efficient ATE estimation under the case-control setting. Similar to the censoring setting, we first derive the efficiency bound and then propose an efficient estimator, providing theoretical guarantees for its consistency and asymptotic normality. Throughout the arguments, we assume $m = \alpha n$ and $l = (1-\alpha)n$, where $\alpha \in (0, 1)$.

D.1 Efficient influence function and efficiency bound

Using efficiency arguments under the stratified sampling scheme (Uehara et al., 2020), we derive the following efficient influence function (see Appendix J for the proof).

Assumption D.1 (Regularity conditions)**.** *The outcome* Y *and covariates* X *have finite variances. There exists a constant* $C > 0$ *independent of* n *such that* $\mu_{\mathrm{T},0}(X), \mu_{\mathrm{U},0}(X) \in [-C, C]$.

Lemma D.2. *If Assumptions C.2–C.4, D.1 hold, then the efficient influence functions are given as* $\Psi^{\mathrm{cc\ (T)}}(X, Y(1); \mu_{\mathrm{T},0}, e_0, r_0)$ *and* $\Psi^{\mathrm{cc\ (U)}}(X, Y_{\mathrm{U}}; \mu_{\mathrm{T},0}, \mu_{\mathrm{U},0}, e_0, \tau_0)$*, where*

$$\Psi^{\mathrm{cc\ (T)}}(X, Y(1); \mu_{\mathrm{T},0}, e_0, r_0) \coloneqq S^{\mathrm{cc\ (T)}}(X, Y(1); \mu_{\mathrm{T},0}, e_0, r_0),$$
$$\Psi^{\mathrm{cc\ (U)}}(X, Y_{\mathrm{U}}; \mu_{\mathrm{T},0}, \mu_{\mathrm{U},0}, e_0, \tau_0) \coloneqq S^{\mathrm{cc\ (U)}}(X, Y_{\mathrm{U}}; \mu_{\mathrm{T},0}, \mu_{\mathrm{U},0}, e) - \tau_0,$$
$$S^{\mathrm{cc\ (T)}}(X, Y(1); \mu_{\mathrm{T},0}, e_0, r_0) \coloneqq \left(1 - \frac{e_0(1 \mid X)}{e_0(0 \mid X)}\right)\Big(Y(1) - \mu_{\mathrm{T},0}(X)\Big)r_0(X),$$
$$S^{\mathrm{cc\ (U)}}(X, Y_{\mathrm{U}}; \mu_{\mathrm{T},0}, \mu_{\mathrm{U},0}, e_0) \coloneqq -\frac{Y_{\mathrm{U}} - \mu_{\mathrm{U},0}(X)}{e_0(0 \mid X)} + \mu_{\mathrm{T},0}(X) - \frac{\mu_{\mathrm{U},0}(X)}{e_0(0 \mid X)} + \frac{e_0(1 \mid X)\mu_{\mathrm{T},0}(X)}{e_0(0 \mid X)},$$

and recall that $r_0(X) \coloneqq \frac{\zeta_0(X)}{\zeta_{\mathrm{T},0}(X)}$.

Then, we obtain the result on the efficiency bound.

Theorem D.3 (Efficiency bound in the case-control setting)**.** *If Assumptions C.2–C.4, D.1 hold, then the asymptotic variance of any regular estimator is lower bounded by*

$$V^{\mathrm{cc}} \coloneqq \frac{1}{\alpha}\mathbb{E}\left[\Psi^{\mathrm{cc\ (T)}}(X, O, Y; \mu_{\mathrm{T},0}, e_0, r_0)^2\right] + \frac{1}{1-\alpha}\mathbb{E}\left[\Psi^{\mathrm{cc\ (U)}}(X, O, Y; \mu_{\mathrm{U},0}, e_0)^2\right]$$
$$= \frac{1}{\alpha}\mathbb{E}\left[\left(1 - \frac{e_0(1 \mid X)}{e_0(0 \mid X)}\right)^2 \mathrm{Var}(Y(1) \mid X)r_0(X)^2\right] + \frac{1}{1-\alpha}\mathbb{E}\left[\frac{\mathrm{Var}(Y_{\mathrm{U}} \mid X)}{e_0(0 \mid X)^2} + \Big(\tau_0(X) - \tau_0\Big)^2\right]$$

where $\alpha = m/n$.

D.2 Semiparametric efficient estimator

Based on the efficient influence function, we define

$$\widehat{\tau}_n^{\text{cc-eff}} \coloneqq \frac{1}{m}\sum_{j=1}^{m} S^{\mathrm{cc\ (T)}}(X_j, Y_j; \widehat{\mu}_{\mathrm{T},n,i}, \widehat{e}_{n,i}, \widehat{r}_{n,i}) + \frac{1}{l}\sum_{k=1}^{l} S^{\mathrm{cc\ (U)}}(X_k, Y_k; \widehat{\mu}_{\mathrm{U},n,i}, \widehat{e}_{n,i}.$$

Here, $\widehat{\mu}_{\mathrm{T},n,i}$, $\widehat{\mu}_{\mathrm{U},n,i}$, $\widehat{e}_{n,i}$, and $\widehat{r}_{n,i}$ are estimators of $\mu_{\mathrm{T},0}$, $\mu_{\mathrm{U},0}$, e_0, and r_0, where m and l denote the dependence on each dataset. Unlike the censoring setting, we do not use the observation indicator O, as it is deterministic whether a unit belongs to the treatment group or the control group. This distinction leads to differences in the theoretical analysis compared to the censoring setting.

D.3 Consistency

We make the following assumption.

Assumption D.4. *As* $n \to \infty$*, it holds that* $\left\|\widehat{e}_{n,i} - e_0\right\|_2 = o_p(1)$ *and* $\left\|\widehat{r}_{n,i} - r_0\right\|_2 = o_p(1)$.

Then, the following consistency result holds.

Theorem D.5 (Consistency in the case-control setting)**.** *If Assumptions C.2–C.4, D.1, and D.4 holds, then* $\widehat{\tau}_n^{\text{cc-eff}} \xrightarrow{\mathrm{p}} \tau_0$ *as* $n \to \infty$.

Interestingly, to achieve consistency, it is sufficient to obtain consistent $\widehat{e}_{n,i}$. Compared to Assumption D.4 in the censoring setting, consistency of the expected outcome estimators $\widehat{\mu}_{\mathrm{T},m}$ and $\widehat{\nu}_\ell$ is not required. This is because, in this setting, the observation probability can be treated as known (1 and 0 for each dataset).

D.4 Asymptotic normality

Next, we establish the asymptotic normality of the estimator. Similar to the censoring setting, we assume that the propensity score e_0 is known and obtain estimators of $\mu_{\mathrm{T},0}$ and $\mu_{\mathrm{U},0}$ via cross-fitting.

Assumption D.6. *The propensity score* e_0 *and the density ratio* r_0 *are known and used in constructing* $\widehat{\tau}_n^{\text{cc-eff}}$*, i.e.,* $\widehat{e}_{n,i} = e_0$ *and* $\widehat{r}_{n,i} = r_0$.

Assumption D.7. *For each $\ell \in \mathcal{L}$, the following conditions hold as $n \to \infty$:* $\left\|\mu_{\mathrm{T},0}(X) - \widehat{\mu}^{(\ell)}_{\mathrm{T},m}(X)\right\|_2 = o_p(1)$ *and* $\left\|\mu_{\mathrm{U},0}(X) - \widehat{\mu}^{(\ell)}_{\mathrm{U},l}(X)\right\|_2 = o_p(1)$.

We establish the asymptotic normality in the following theorem with the proof in Appendix K. In this result, we consider the scenario where the sample sizes m and l approach infinity while maintaining a fixed ratio $m : l = \alpha : (1 - \alpha)$.

Theorem D.8 (Asymptotic normality in the case-control setting)**.** *Fix $\alpha \in (0, 1)$. For $n > 0$, consider the case-control setting with sample sizes m, l such that $m = \alpha n$ and $l = (1 - \alpha)n$. Suppose that Assumptions C.3–C.4, D.1, D.6, and D.7 hold; that is, $\widehat{e}_{n,i} = e_0$ and $\widehat{r}_{n,i} = r_0$, and $\widehat{\mu}_{\mathrm{T},n,i} = \widehat{\mu}^{(\ell)}_{\mathrm{T},m}$, $\widehat{\mu}_{\mathrm{U},n,i} = \widehat{\mu}^{(\ell)}_{\mathrm{U},l}$ are consistent estimators constructed via cross-fitting. Then, we have*

$$\sqrt{n}\left(\widehat{\tau}_n^{\text{cc-eff}} - \tau_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V^{\mathrm{cc}}) \ \text{ as } \ n \to \infty$$

Thus, the proposed estimator is efficient with respect to the efficiency bound derived in Theorem D.3.

D.5 Comparison with the censoring setting

Unlike the censoring setting, we do not require a specific convergence rate for the nuisance estimators if the propensity score is known. This is because, in the case-control setting, the group membership—whether a unit belongs to the treatment group or the unknown group—is deterministically known. This scenario can be interpreted as a case in which the observation probability is known, meaning that only the consistency of the expected outcome estimators is required (Kato et al., 2020, 2021).

In other words, we can intuitively consider that in the case-control case, the observation probability is given as one for the treatment group, while it is given as zero for the unknown group. Since we know the true probabilities, we can ignore the estimation error unlike the censoring setting. Note that this interpretation may be mathematically confusing since it gives us impression that the case-control setting is a special case of the censoring setting where the observation probability is given one or zero. This understanding is not correct because in the case-control setting, the treated and unknown groups are different datasets; that is, the sampling scheme is completely different. This sampling scheme has extensively studied as stratified sampling scheme from 1990s to 2000s by existing studies such as Imbens & Lancaster (1996) and Wooldridge (2001).

E PU learning algorithms

We review representative PU learning methods. For all methods, the goal is not to obtain a conditional class probability (propensity score) but rather to obtain a better classifier. However, under specific loss functions, including logistic loss, the obtained classifiers can be interpreted as estimators of the probability (Elkan & Noto, 2008; Kato et al., 2019; Kato & Teshima, 2021).

E.1 Censoring PU learning

In Elkan & Noto (2008), it is assumed that only a fraction of the truly positive instances are labeled as positive.

Let O denote the event "labeled as positive," and let $D = 1$ indicate true positivity. In our study, O is called an observation indicator, and D is called a treatment indicator.

First, we make the following assumption, which plays a central role in the method of Elkan & Noto (2008):

$$\mathbb{P}(O = 1 \mid D = 1, x) = c\Big(= \mathbb{P}(O = 1 \mid D = 1)\Big) \quad \forall x \in \mathcal{X}, \tag{1}$$

where $c \in (0, 1]$ is a constant (Assumption 3.1). Intuitively, c represents the *labeling probability* or *censoring rate*, which denotes the fraction of positive instances that are observed (uncensored) in the labeled dataset. If we relax this assumption, we may not pointy identify the ATE without different assumptions. There are various approaches proposed to address the relaxation (Bekker & Davis, 2018).

The learning procedure proposed by Elkan & Noto (2008) consists of three main steps (for details, see Elkan & Noto (2008)):

Estimation of π_0: First, the observation probability π_0 is estimated using standard regression methods, such as logistic regression.

Estimation of c: Next, c is estimated using an estimator $\widehat{\pi}_n$ of π_0. Under Assumption 3.1, c can be estimated by taking the sample average of $\widehat{\pi}_n$ over positively labeled samples.

Correction of the observation probability: From Assumption 3.1, we have $\pi_0(1 \mid X) = c\mathbb{P}(D = 1 \mid X)$. Using this relationship and the estimators of π_0 and c, $\mathbb{P}(D = 1 \mid X)$ is estimated as $\widehat{\pi}_n(1 \mid X)/\widehat{c}$.

Remark. *Violation of the assumptions The impact of violating assumptions depends on how the data deviates from the assumed conditions. For example, if treatment labels are not missing at random, the original estimator of Elkan & Noto (2008) may no longer be valid. In such cases, alternative methods—such as those proposed by Bekker & Davis (2018) and Teisseyre et al. (2025)—may be applicable, although they rely on different assumptions. In the case-control setting, du Plessis et al. (2014) provides a sensitivity analysis of the trained classifier when the class prior is misspecified.*

Remark (Time complexity)**.** *The computational cost of estimating the conditional class probability via PU learning is comparable to that of standard logistic regression. For example, in the censoring setting, Elkan & Noto (2008) proposes a method based on logistic regression. In the case-control setting, du Plessis et al. (2015) presents an unbiased PU learning approach which, under the log-loss, has the same complexity as logistic regression. Specifically, for linear-in-parameter models, logistic regression has a time complexity of order $O((dn^2 + n^3)\log(1/\epsilon))$, where d is the feature dimension and ϵ is the target optimization accuracy. If Newton's method is used with T iterations, the total time complexity becomes $O((dn^2 + n^3)T)$, which dominates the final averaging step.*

E.2 Case-control PU learning

A different perspective is provided by du Plessis et al. (2015) and subsequent studies, often referred to as case-control PU learning. In this approach, the labeled positive data follow a distribution $\zeta_{\mathrm{T},0}(x)$, whereas the unlabeled data are drawn from $\zeta_0(x)$, a mixture of positive and negative instances.

Let h be a classifier. In conventional supervised learning, the classification risk is defined as:

$$R(h) = e_0(1)R_+(h, +1) + (1 - e_0(1))R_-(h, -1),$$

where $e_0(1)$ is the prior probability of being positive, and $R_+(h, +1)$ and $R_-(h, -1)$ denote the risks over the positive and negative distributions, respectively. Specifically, $R_+(h, +1)$ represents the expected loss when predicting class -1 while the true label is $+1$ in the positive distribution, and $R_-(h, -1)$ represents the expected loss when predicting class $+1$ while the true label is -1 in the negative distribution.

Since negative examples are unavailable, du Plessis et al. (2015) re-express $R_-(h, -1)$ as:

$$(1 - e_0(1))R_-(h, -1) = R_{\mathrm{U}}(h, -1) - e_0(1)R_+(h, -1),$$

where $R_{\mathrm{U}}(h, -1)$ and $R_+(h, -1)$ denote the risks over the unlabeled and positive distributions, respectively. The term $R_{\mathrm{U}}(h, -1)$ corresponds to the expected loss when predicting class $+1$ while the true label is -1 in the unlabeled distribution, and $R_+(h, -1)$ is the expected loss under the same prediction and true label in the positive distribution. Note that $R_+(h, +1)$ and $R_+(h, -1)$ are distinct: they consider different true labels while expectations are taken over the same positive distribution.

Substituting the above expression into the original risk gives the following classification risk:

$$R(h) = e_0(1)R_+(h, +1) + R_{\mathrm{U}}(h, -1) - e_0(1)R_+(h, -1).$$

A sample-based approximation of this formulation is referred to as an *unbiased risk estimator*.

For example, using the logistic loss, the unbiased risk estimator becomes:

$$\widehat{R}(h) = e_0(1)\frac{1}{m}\sum_{j=1}^{m} \log\left(1 + \exp(-h(X_j))\right)$$

$$+\frac{1}{l}\sum_{k=1}^{l}\log\big(1+\exp(h(X_k))\big)-e_0(1)\frac{1}{m}\sum_{j=1}^{m}\log\big(1+\exp(h(X_j))\big).$$

Then, we can train a classifier as $\widehat{h}:=\arg\min_{h\in\mathcal{H}}\widehat{R}(h)$, where $\mathcal{H}$ is a given hypothesis set.

Note that we assume the class prior $e_0(1)$ is known. Several methods have been proposed to estimate it under additional assumptions (du Plessis & Sugiyama, 2014).

E.3 Density-ratio estimation

Since the density ratio can be estimated in the case-control setting, we introduce related methods. Density-ratio estimation has emerged as a powerful technique in machine learning and statistics, providing a principled approach for estimating the ratio of two probability density functions (Sugiyama et al., 2012). Let $X_i, Z_i \in \mathcal{X}$ be random variables. Specifically, if $\{X_i\}_{i=1}^n$ are drawn from $p_0(x)$ and $\{Z_j\}_{j=1}^m$ are drawn from $q_0(z)$, the goal is to estimate

$$r_0(x)=\frac{p_0(x)}{q_0(x)}$$

directly, *without* first estimating $p_0(x)$ and $q_0(z)$ separately.

Estimating $p_0(x)$ and $q_0(z)$ individually can be challenging and may introduce unnecessary modeling complexities if only the ratio $r_0(x)$ is required. By *directly* estimating the density ratio, more stable and accurate estimates can often be obtained, avoiding potential compounding errors from separately learned density models.

Various algorithms have been proposed for direct density-ratio estimation, including the Kullback–Leibler Importance Estimation Procedure (KLIEP, Sugiyama et al., 2008) and Least-Squares Importance Fitting (LSIF, Kanamori et al., 2009). These methods typically optimize a criterion that ensures the estimated ratio closely approximates the true ratio in a specific divergence sense, such as the Kullback–Leibler divergence or squared error, which can be generalized as a Bregman divergence minimization problem (Sugiyama et al., 2012).

From the Bregman divergence minimization perspective, PU learning methods can also be seen as variants of density-ratio estimation, as demonstrated in Kato et al. (2019) and Kato & Teshima (2021).

F Remark on the nuisance parameter estimation in the censoring setting

In the censoring setting, by applying the method of Elkan & Noto (2008), we can obtain an estimator of $\mathbb{P}(D=1\mid X)$ from an estimator of $\pi_0(o\mid x)=\mathbb{P}(O=o\mid X)$. However, our objective is to estimate

$$g_0(1\mid X)=\mathbb{P}(D=1\mid X,O=0)=\frac{\mathbb{P}(D=1,O=0\mid X)}{\mathbb{P}(O=0\mid X)}=\frac{\mathbb{P}(O=0\mid D=1)\mathbb{P}(D=1\mid X)}{\mathbb{P}(O=0\mid X)}$$

rather than $\mathbb{P}(D=1\mid X)$.

Let $\widehat{\kappa}_n(1\mid X)$ be an estimator of $\mathbb{P}(D=1\mid X)$. We can then obtain an estimator of $g_0(1\mid X)$ as follows:

$$\widehat{g}_n(1\mid X)=\frac{(1-\widehat{c})\widehat{\kappa}_n(1\mid X)}{\widehat{\pi}_n(0\mid X)},$$

where $\widehat{c}$ is an estimate of $c=\mathbb{P}(O=0\mid D=1)$. Notably, under Assumption 3.1, c can be estimated by taking the mean of $\widehat{\pi}_n(0\mid X)$ over the positively labeled sample ($O_i=1$).

G Pseudo-code for ATE estimation in the case-control setting

We explain how we construct the estimators of the nuisance parameters in the case-control setting.

We can estimate $\mu_{\mathrm{T},0}$ and $\mu_{\mathrm{U},0}$ using standard regression methods, including logistic regression and nonparametric regression. Specifically, for estimating $\mu_{\mathrm{T},0}$, we typically use the dataset $\big\{\big(X_{\mathrm{T},j},Y_j(1)\big)\big\}_{j=1}^m$, while for estimating $\mu_{\mathrm{U},0}$, we use $\big\{\big(X_k,Y_{\mathrm{U},k}\big)\big\}_{k=1}^l$.

Algorithm 2 Cross-fitting in the case-control setting

Input: Observations $\mathcal{D}_{\mathrm{T}} = \left\{\left(X_{\mathrm{T},j}, Y_j(1)\right)\right\}_{j=1}^{m}$ and $\mathcal{D}_{\mathrm{U}} = \left\{\left(X_k, Y_{\mathrm{U},k}\right)\right\}_{k=1}^{l}$, number of folds L, and estimation methods for $\mu_{\mathrm{T},0}, \mu_{\mathrm{U},0}, e_0, r_0$.
Randomly partition $\mathcal{D}_{\mathrm{T}}$ into L roughly equal-sized folds, $(\mathcal{J}^{(\ell)})_{\ell\in\mathcal{L}}$. Note that $\bigcup_{\ell\in\mathcal{L}} \mathcal{J}^{(\ell)} = \mathcal{D}$.
Randomly partition $\mathcal{D}_{\mathrm{U}}$ into L roughly equal-sized folds, $(\mathcal{K}^{(\ell)})_{\ell\in\mathcal{L}}$. Note that $\bigcup_{\ell\in\mathcal{L}} \mathcal{K}^{(\ell)} = \mathcal{D}$.
for $\ell \in \mathcal{L}$ **do**
  Set the training data as $\mathcal{J}^{(-\ell)} = \{1, 2, \ldots, n\} \setminus \mathcal{J}^{(\ell)}$.
  Set the training data as $\mathcal{K}^{(-\ell)} = \{1, 2, \ldots, n\} \setminus \mathcal{K}^{(\ell)}$.
  Construct estimators of nuisance parameters on $\mathcal{J}^{(-\ell)}$ and $\mathcal{K}^{(-\ell)}$.
  Construct an ATE estimate $\widehat{\tau}_n^{\text{cc-eff}(\ell)}$ using $\mathcal{J}^{(\ell)}$, $\mathcal{K}^{(\ell)}$and the nuisance estimates $\widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\mu}_{\mathrm{U},n}^{(\ell)}, \widehat{e}_n^{(\ell)}, \widehat{r}_n^{(\ell)}$.
end for
Output: Combine $(\widehat{\tau}_n^{\text{cc-eff}(\ell)})_{\ell\in\mathcal{L}}$ to form $\widehat{\tau}_n^{\text{cc-eff}}$.

To estimate e_0, we can apply case-control PU learning methods, such as convex PU learning proposed by du Plessis et al. (2015). For estimating r_0, density-ratio estimation methods can be employed (Sugiyama et al., 2012).

Notably, if $\zeta_{\mathrm{T},0}(x) = \zeta_0(x \mid d = 1)$, then r_0 can be estimated from an estimator of e_0 using the relationship $r_0 = 1/(e_0(1 \mid x)e_0(1))$.

In cross-fitting, we split $\mathcal{D}_{\mathrm{T}}$ and $\mathcal{D}_{\mathrm{U}}$, respectively, as performed in Uehara et al. (2020). The pseudo-code is shown in Algorithm G.

H Proof of Lemma 4.2

We prove Lemma 4.2. Our proof procedure is inspired by the one in Hahn (1998).

Proof. Recall that the density function for (X, O, Y) is given as

$$p_0(x, o, y) = \zeta_0(x)\Big(\pi_0(1 \mid x)p_{Y(1),0}(y \mid x)\Big)^{\mathbb{1}[o=1]}\Big(\pi_0(0 \mid x)p_{\widetilde{Y},0}(y \mid x)\Big)^{\mathbb{1}[o=0]},$$

where $p_{Y(1),0}(y \mid x)$, $p_{Y(0),0}(y \mid x)$, and $p_{\widetilde{Y},0}(y \mid x)$ are the conditional densities of $Y(1)$, $Y(0)$, and $\widetilde{Y}$ in the censoring seting.

For this density function, we consider the parametric submodels:

$$\mathcal{P}^{\text{sub}} \coloneqq \{P_\theta \in \mathcal{P} \colon \theta \in \mathbb{R}\},$$

where P_θ has the following density:

$$p_0(x, o, y; \theta) = \zeta_0(x; \theta)\Big(\pi_0(1 \mid x; \theta)p_{Y(1),0}(y \mid x; \theta)\Big)^{\mathbb{1}[o=1]}\Big(\pi_0(0 \mid x; \theta)p_{\widetilde{Y}}(y \mid x; \theta)\Big)^{\mathbb{1}[o=0]},$$

while there exists $\theta_0 \in \mathbb{R}$ such that

$$p(x, o, y; \theta_0) = p_0(x, o, y).$$

Then, we define scores as follows:

$$\begin{aligned} S(x, o, y; \theta) &\coloneqq \frac{\partial}{\partial\theta} \log p(x, o, y; \theta) \\ &= S_X(x; \theta) + \mathbb{1}[o = 1]\left(S_{Y(1)}(y \mid x; \theta) + \frac{\dot{\pi}(1 \mid x; \theta)}{\pi(1 \mid x; \theta)}\right) + \mathbb{1}[o = 0]\left(S_{\widetilde{Y}}(y \mid x; \theta) + \frac{\dot{\pi}(0 \mid x; \theta)}{\pi(0 \mid x; \theta)}\right), \end{aligned}$$

where

$$S_X(x; \theta) \coloneqq \frac{d}{d\theta} \log \zeta(x; \theta),$$

$$S_{Y(1)}(y \mid x;\theta) \coloneqq \frac{d}{d\theta}\log p_{Y(1)}(y \mid x;\theta),$$
$$S_{\widetilde{Y}}(y \mid x;\theta) \coloneqq \frac{d}{d\theta}\log p_{\widetilde{Y}}(y \mid x;\theta),$$
$$\dot{\pi}(0 \mid x;\theta) \coloneqq \frac{d}{d\theta}\pi(o \mid x;\theta).$$

Let $\mathcal{T} \coloneqq \{S(x,o,y;\theta)\}$ be the tangent space.

Here, note that

$$p_{\widetilde{Y}}(y \mid x;\theta) = g_0(1 \mid x)p_{Y(1)}(y \mid x;\theta) + g_0(0 \mid x)p_{Y(0)}(y \mid x;\theta).$$

We have

$$p_{\mathrm{C}}(y \mid x;\theta) = \frac{1}{g_0(0 \mid x)}\Big(p_{\widetilde{Y}}(y \mid x;\theta) - g_0(1 \mid x)p_{Y(1)}(y \mid x;\theta)\Big)$$

Using this relationship, we write the ATE under the parametric submodels as

$$\begin{aligned}\tau(\theta) &\coloneqq \iint y(1)p_{Y(1)}(y(1) \mid x;\theta)\zeta(x;\theta)\mathrm{d}y(1)\mathrm{d}x - \iint y(0)p_{Y(0)}(y(0) \mid x;\theta)\zeta(x;\theta)\mathrm{d}y(0)\mathrm{d}x \\ &= \iint y(1)p_{Y(1)}(y(1) \mid x;\theta)\zeta(x;\theta)\mathrm{d}y(1)\mathrm{d}x - \iint y(0)\frac{1}{g_0(0 \mid x)}p_{\widetilde{Y}}(y(0) \mid x;\theta)\zeta(x;\theta)\mathrm{d}y(0)\mathrm{d}x \\ &\qquad + \iint y(0)\frac{g_0(1 \mid x)}{g_0(0 \mid x)}p_{Y(1)}(y(0) \mid x;\theta)\zeta(x;\theta)\mathrm{d}y(0)\mathrm{d}x.\end{aligned}$$

Them, the derivative is given as

$$\begin{aligned}\frac{\partial\tau(\theta)}{\partial\theta} &= \mathbb{E}_\theta\Big[Y(1)S_{Y(1)}(Y(1) \mid X;\theta)\Big] - \mathbb{E}_\theta\left[\frac{1}{g_0(0 \mid X)}\widetilde{Y}S_{\widetilde{Y}}(\widetilde{Y} \mid X;\theta)\right] \\ &\quad + \mathbb{E}_\theta\left[\frac{g_0(1 \mid X)}{g_0(0 \mid X)}Y(1)S_{Y(1)}(Y(1) \mid X;\theta)\right] \\ &\quad + \mathbb{E}_\theta\Big[\tau(X;\theta)S_X(X;\theta)\Big],\end{aligned}$$

where

$$\tau(X;\theta) \coloneqq \mu(1 \mid X) - \frac{1}{g_0(0 \mid X)}\mu(U \mid X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T}}(X) = \mu(1 \mid X) - \mu_{\mathrm{C}}(X)$$

From the Riesz representation theorem, there exists a function Ψ such that

$$\frac{\partial\tau(\theta)}{\partial\theta}\Big|_{\theta=\theta_0} = \mathbb{E}\big[\Psi(X,O,Y)S(X,O,Y;\theta_0)\big]. \tag{2}$$

There exists a unique function Ψ^{cens} such that $\Psi^{\mathrm{cens}} \in \mathcal{T}$, called the efficient influence function. We specify the efficient influence function as

$$\begin{aligned}&\Psi^{\mathrm{cens}}(X,O,Y;\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0) \\ &= S^{\mathrm{cens}}(X,O,Y;\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0) - \tau_0, \\ &= \frac{\mathbb{1}[O=1]\Big(Y - \mu_{\mathrm{T},0}(X)\Big)}{\pi_0(1 \mid X)} - \frac{\mathbb{1}[O=0]\Big(Y - \nu_0(X)\Big)}{g_0(0 \mid X)\pi_0(0 \mid X)} \\ &\quad + \frac{g_0(1 \mid X)\mathbb{1}[O=1]\Big(Y - \mu_{\mathrm{T},0}(X)\Big)}{g_0(0 \mid X)\pi_0(1 \mid X)} \\ &\quad + \mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X) - \tau_0.\end{aligned}$$

We prove that $\Psi^{\mathrm{cens}}(X,O,Y;\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0)$ is actually the unique efficient influence function by verifying that Ψ^{cens} satisfies (2) and $\Psi^{\mathrm{cens}} \in \mathcal{T}$.

Proof of (2): First, we confirm that Ψ^{cens} satisfies (2). We have

$$
\begin{aligned}
&\mathbb{E}\left[\Psi^{\mathrm{cens}}(X,O,Y;\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0)S(X,O,Y;\theta_0)\right] \\
&= \mathbb{E}\Bigg[\Psi^{\mathrm{cens}}(X,O,Y;\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0) \\
&\quad\cdot\Big(S_X(X;\theta)+\mathbb{1}[O=1]\left(S_{Y(1)}(Y\mid X;\theta_0)+\frac{\dot\pi(1\mid X;\theta)}{\pi(1\mid X;\theta_0)}\right) \\
&\qquad\qquad +\mathbb{1}[O=0]\left(S_{\widetilde{Y}}(Y\mid X;\theta)+\frac{\dot\pi(0\mid X;\theta_0)}{\pi(0\mid X;\theta_0)}\right)\Big)\Bigg] \\
&= \mathbb{E}\Bigg[\Bigg(\frac{\mathbb{1}[O=1]\Big(Y-\mu_{\mathrm{T},0}(X)\Big)}{\pi_0(1\mid X)}-\frac{\mathbb{1}[O=0]\Big(Y-\nu_0(X)\Big)}{g_0(0\mid X)\pi_0(0\mid X)} \\
&\quad+\frac{g_0(1\mid X)\mathbb{1}[O=1]\Big(Y-\mu_{\mathrm{T},0}(X)\Big)}{g_0(0\mid X)\pi_0(1\mid X)} \\
&\quad+\mu_{\mathrm{T},0}(X)-\frac{1}{g_0(0\mid X)}\nu(X)+\frac{g_0(1\mid X)}{g_0(0\mid X)}\mu_{\mathrm{T},0}(X)-\tau_0\Bigg) \\
&\quad\cdot\Bigg(S_X(X;\theta)+\mathbb{1}[O=1]\left(S_{Y(1)}(Y\mid X;\theta_0)+\frac{\dot\pi(1\mid X;\theta)}{\pi(1\mid X;\theta_0)}\right) \\
&\qquad\qquad +\mathbb{1}[O=0]\left(S_{\widetilde{Y}}(Y\mid X;\theta)+\frac{\dot\pi(0\mid X;\theta_0)}{\pi(0\mid X;\theta_0)}\right)\Bigg)\Bigg] \\
&= \mathbb{E}\Bigg[\left(\mu_{\mathrm{T},0}(X)-\frac{1}{g_0(0\mid X)}\nu(X)+\frac{g_0(1\mid X)}{g_0(0\mid X)}\mu_{\mathrm{T},0}(X)-\tau_0\right)S_X(X;\theta_0) \\
&\quad+\Bigg(\frac{\mathbb{1}[O=1]\Big(Y-\mu_{\mathrm{T},0}(X)\Big)}{\pi_0(1\mid X)}+\frac{g_0(1\mid X)\mathbb{1}[O=1]\Big(Y-\mu_{\mathrm{T},0}(X)\Big)}{g_0(0\mid X)\pi_0(1\mid X)} \\
&\quad+\mu_{\mathrm{T},0}(X)-\frac{1}{g_0(0\mid X)}\nu(X)+\frac{g_0(1\mid X)}{g_0(0\mid X)}\mu_{\mathrm{T},0}(X)-\tau_0\Bigg)\mathbb{1}[O=1]\left(S_{Y(1)}(Y\mid X;\theta_0)+\frac{\dot\pi(1\mid X;\theta_0)}{\pi(1\mid X;\theta_0)}\right) \\
&\quad+\Bigg(-\frac{\mathbb{1}[O=0]\Big(Y-\nu_0(X)\Big)}{g_0(0\mid X)\pi_0(0\mid X)}+\mu_{\mathrm{T},0}(X)-\frac{1}{g_0(0\mid X)}\nu(X)+\frac{g_0(1\mid X)}{g_0(0\mid X)}\mu_{\mathrm{T},0}(X)-\tau_0\Bigg) \\
&\qquad\cdot\mathbb{1}[O=0]\left(S_{\widetilde{Y}}(Y\mid X;\theta)+\frac{\dot\pi(0\mid X;\theta_0)}{\pi(0\mid X;\theta_0)}\right)\Bigg],
\end{aligned}
$$

where we used $\mathbb{1}[O=1]\mathbb{1}[O=0]=0$, and

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\mathbb{1}[O=1]\Big(Y-\mu_{\mathrm{T},0}(X)\Big)}{\pi_0(1\mid X)}\right]=\mathbb{E}\left[\frac{\mathbb{1}[O=1]\Big(Y(1)-\mu_{\mathrm{T},0}(X)\Big)}{\pi_0(1\mid X)}\right] \\
&\quad=\mathbb{E}\left[\frac{\pi_0(1\mid X)\Big(\mu_{\mathrm{T},0}(X)-\mu_{\mathrm{T},0}(X)\Big)}{\pi_0(1\mid X)}\right]=0, \\
&\mathbb{E}\left[\frac{\mathbb{1}[O=0]\Big(Y-\nu_0(X)\Big)}{g_0(0\mid X)\pi_0(0\mid X)}\right]=\mathbb{E}\left[\frac{\mathbb{1}[O=0]\Big(\widetilde{Y}-\nu_0(X)\Big)}{g_0(0\mid X)\pi_0(0\mid X)}\right]=\mathbb{E}\left[\frac{\pi_0(0\mid X)\Big(\nu_0(X)-\nu_0(X)\Big)}{g_0(0\mid X)\pi_0(0\mid X)}\right]=0, \\
&\mathbb{E}\left[\frac{g_0(1\mid X)\mathbb{1}[O=1]\Big(Y-\mu_{\mathrm{T},0}(X)\Big)}{g_0(0\mid X)\pi_0(1\mid X)}\right]=\mathbb{E}\left[\frac{g_0(1\mid X)\pi_0(1\mid X)\Big(\mu_{\mathrm{T},0}(X)-\mu_{\mathrm{T},0}(X)\Big)}{g_0(0\mid X)\pi_0(1\mid X)}\right]=0.
\end{aligned}
$$

We have

$$
\begin{aligned}
&\mathbb{E}\Bigg[\Bigg(\mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X) - \tau_0\Bigg) S_X(X;\theta) \\
&\quad + \Bigg(\frac{\mathbb{1}[O=1]\Big(Y - \mu_{\mathrm{T},0}(X)\Big)}{\pi_0(1 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1]\Big(Y - \mu_{\mathrm{T},0}(X)\Big)}{g_0(0 \mid X)\pi_0(1 \mid X)} \\
&\qquad + \mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X) - \tau_0\Bigg) \mathbb{1}[O=1]\left(S_{Y(1)}(Y \mid X;\theta_0) + \frac{\dot{\pi}(1 \mid X;\theta)}{\pi(1 \mid X;\theta_0)}\right) \\
&\qquad + \Bigg(-\frac{\mathbb{1}[O=0]\Big(Y - \nu_0(X)\Big)}{g_0(0 \mid X)\pi_0(0 \mid X)} + \mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X) - \tau_0\Bigg) \\
&\qquad\quad \cdot \mathbb{1}[O=0]\left(S_{\widetilde{Y}}(Y \mid X;\theta) + \frac{\dot{\pi}(0 \mid X;\theta_0)}{\pi(0 \mid X;\theta_0)}\right)\Bigg] \\
&= \mathbb{E}\Bigg[\Bigg(\mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X)\Bigg) S_X(X;\theta_0) \\
&\quad + \Bigg(\frac{\mathbb{1}[O=1]\Big(Y - \mu_{\mathrm{T},0}(X)\Big)}{\pi_0(1 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1]\Big(Y - \mu_{\mathrm{T},0}(X)\Big)}{g_0(0 \mid X)\pi_0(1 \mid X)}\Bigg) S_{Y(1)}(Y \mid X;\theta_0) \\
&\qquad - \frac{\mathbb{1}[O=0]\Big(Y - \nu_0(X)\Big)}{g_0(0 \mid X)\pi_0(0 \mid X)} S_{\widetilde{Y}}(Y \mid X;\theta)\Bigg] \\
&= \mathbb{E}\Bigg[\Bigg(\mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X)\Bigg) S_X(X;\theta_0) \\
&\quad + \Bigg(\frac{\mathbb{1}[O=1]Y(1)}{\pi_0(1 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1]Y(1)}{g_0(0 \mid X)\pi_0(1 \mid X)}\Bigg) S_{Y(1)}(Y(1) \mid X;\theta) - \frac{\mathbb{1}[O=0]\widetilde{Y}}{g_0(0 \mid X)\pi_0(0 \mid X)} S_{\widetilde{Y}}(\widetilde{Y} \mid X;\theta_0)\Bigg],
\end{aligned}
$$

where we used

$$
\begin{aligned}
&\mathbb{E}\Big[\tau_0 S_X(X;\theta)\Big] = 0 \\
&\mathbb{E}\Bigg[\Big(\mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X) - \tau_0\Big)\mathbb{1}[O=1]\left(S_{Y(1)}(Y \mid X;\theta_0) + \frac{\dot{\pi}(1 \mid X;\theta_0)}{\pi(1 \mid X;\theta_0)}\right)\Bigg] \\
&\quad = 0.
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
&\mathbb{E}\Bigg[\Bigg(\mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X)\Bigg) S_X(X;\theta_0) \\
&\quad + \Bigg(\frac{\mathbb{1}[O=1]Y(1)}{\pi_0(1 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1]Y(1)}{g_0(0 \mid X)\pi_0(1 \mid X)}\Bigg) S_{Y(1)}(Y(1) \mid X;\theta_0) - \frac{\mathbb{1}[O=0]\widetilde{Y}}{g_0(0 \mid X)\pi_0(0 \mid X)} S_{\widetilde{Y}}(\widetilde{Y} \mid X;\theta_0)\Bigg] \\
&= \mathbb{E}\Big[Y(1) S_{Y(1)}(Y(1) \mid X;\theta_0)\Big] - \mathbb{E}\left[\frac{1}{g_0(0 \mid X)}\widetilde{Y} S_{\widetilde{Y}}(\widetilde{Y} \mid X;\theta_0)\right] \\
&\quad + \mathbb{E}\left[\frac{g_0(1 \mid X)}{g_0(0 \mid X)} Y(1) S_{Y(1)}(Y(1) \mid X;\theta_0)\right] \\
&\quad + \mathbb{E}_{\theta_0}\Big[\tau(X;\theta) S_X(X;\theta_0)\Big]
\end{aligned}
$$

$$= \frac{\partial \tau(\theta)}{\partial \theta}\Big|_{\theta=\theta_0}$$

Proof of $\Psi^{\mathrm{cens}} \in \mathcal{T}$: Set

$$S_{Y(1)}(y \mid x) = \frac{y - \mathbb{E}\big[Y(1) \mid X = x\big]}{\pi_0(1 \mid x)},$$
$$S_{\widetilde{Y}}(y \mid x) = \frac{y - \mathbb{E}\big[\widetilde{Y} \mid X = x\big]}{\pi_0(0 \mid x)},$$
$$S_X(X;\theta) = \mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X) - \tau_0.$$

Then, $\Psi^{\mathrm{cens}} \in \mathcal{T}$ holds. □

Remark. *Some regularity conditions are needed to guarantee that the expectations of the score functions are zero. In particular, one must justify interchanging integration and differentiation for the relevant densities. This holds, for example, when the density is integrable and has a bounded first derivative.*

More precisely, it is standard to construct a regular parametric submodel of the true distribution. Regular parametric submodels are technical tools used to derive the efficiency bound and can be constructed to satisfy the necessary regularity conditions. For a true density $p_0(x)$, a typical construction is $p_t(x) = (1 + tg(x))p_0(x)$, where $g(x)$ is a bounded score function and $t \in [0, \infty)$. This satisfies $g(x) = \frac{\partial}{\partial t}\big|_{t=0} \log p_t(x)$. If $g(x)$ is unbounded, we can still define a parametric submodel as $p_t(x) = c(t)k(tg(x))p_0(x)$, where k is a nonnegative function satisfying $k(0) = k'(0) = 1$, and $c(t)$ is a normalizing constant ensuring that $\int p_t(x)dx = 1$. For more details, see Section 25.16 of van der Vaart (1998).

In our study, we verify the regularity requirement by checking that the parametric submodel induced by the derived score satisfies the necessary properties. In our case, the required regularity holds under finite first and second moments of the outcomes (Assumptions 4.1), together with the common support condition (Assumption 3.3).

I Proof of Theorem 4.8: Semiparametric efficient ATE estimator under the censoring setting

For simplicity, we consider two-fold cross-fitting; that is, $L = 2$. Without loss of generality, we assume that the sample size n is even, and let $\overline{n} = n/2$. For each $\ell \in \{1, 2\}$, we denote the subset of the dataset in cross-fitting as

$$\mathcal{D}^{(\ell)} \coloneqq \{(\widetilde{X}_i^{\ell}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)})\}_{i=1}^{\overline{n}}.$$

We defined the estimator as

$$\widehat{\tau}_n^{\text{cens-eff}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{cens}}(X_i, O_i, Y_i; \widehat{\mu}_{\mathrm{T},n,i}, \widehat{\nu}_{n,i}, \widehat{\pi}_{n,i}, g_0),$$

where recall that

$$S^{\mathrm{cens}}(X, O, Y; \widehat{\mu}_{\mathrm{T},n,i}, \widehat{\nu}_{n,i}, \widehat{\pi}_{n,i}, g_0)$$
$$= \frac{\mathbb{1}[O=1]\Big(Y - \widehat{\mu}_{\mathrm{T},n,i}(X)\Big)}{\widehat{\pi}_{n,i}(1 \mid X)} - \frac{\mathbb{1}[O=0]\Big(Y - \widehat{\nu}_{n,i}(X)\Big)}{g_0(0 \mid X)\widehat{\pi}_{n,i}(0 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1]\Big(Y - \widehat{\mu}_{\mathrm{T},n,i}(X)\Big)}{g_0(0 \mid X)\widehat{\pi}_{n,i}(1 \mid X)}$$
$$+ \widehat{\mu}_{\mathrm{T},n,i}(X) - \frac{1}{g_0(0 \mid X)}\nu(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\widehat{\mu}_{\mathrm{T},n,i}(X).$$

We have

$$\widehat{\tau}_n^{\text{cens-eff}} = \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{cens}}(X_i, O_i, Y_i; \widehat{\mu}_{\mathrm{T},n,i}, \widehat{\nu}_{n,i}, \widehat{\pi}_{n,i}, g_0)$$

$$= \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{cens}}(X_i, O_i, Y_i; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) - \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{cens}}(X_i, O_i, Y_i; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0)$$
$$+ \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{cens}}(X_i, O_i, Y_i; \widehat{\mu}_{\mathrm{T},n,i}, \widehat{\nu}_{n,i}, \widehat{\pi}_{n,i}, g_0).$$

Here, if it holds that

$$\frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{cens}}(X_i, O_i, Y_i; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) - \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{cens}}(X_i, O_i, Y_i; \widehat{\mu}_{\mathrm{T},n,i}, \widehat{\nu}_{n,i}, \widehat{\pi}_{n,i}, g_0) = o_p(1/\sqrt{n}) \tag{3}$$

then we have

$$\sqrt{n}\Big(\widehat{\tau}_n^{\text{cens-eff}} - \tau_0\Big) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S^{\mathrm{cens}}(X_i, O_i, Y_i; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) + o_p(1)$$
$$\xrightarrow{\mathrm{d}} \mathcal{N}(0, V^{\mathrm{cens}}),$$

from the central limit theorem for i.i.d. random variables.

Therefore, we prove Theorem 4.8 by showing (3). We decompose the LHS of (3) as

$$\frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{cens}}(X_i, O_i, Y_i; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) - \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{cens}}(X_i, O_i, Y_i; \widehat{\mu}_{\mathrm{T},n,i}, \widehat{\nu}_{n,i}, \widehat{\pi}_{n,i}, g_0)$$
$$= \frac{\overline{n}}{n}\sum_{\ell\in\{1,2\}}\left(\frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0)\right.$$
$$\left. - \frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0)\right).$$

Let $\mathcal{D}^{(\ell)}$ denote the ℓ-th fold of $\mathcal{D}$. Here, we have

$$\frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) - \frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0)$$
$$= \frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) - \frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0)$$
$$- \left(\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) \mid \mathcal{D}^{(\ell)}\Big]\right.$$
$$\left. - \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0) \mid \mathcal{D}^{(\ell)}\Big]\right)$$
$$+ \left(\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) \mid \mathcal{D}^{(\ell)}\Big]\right.$$
$$\left. - \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0) \mid \mathcal{D}^{(\ell)}\Big]\right).$$

To show (3), we show the following two inequalities separately:

$$\frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) - \frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0)$$
$$- \left(\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) \mid \mathcal{D}^{(\ell)}\Big]\right.$$

$$- \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0) \mid \mathcal{D}^{(\ell)}\Big]\Bigg)$$

$$= o_p(1/\sqrt{n}), \tag{4}$$

$$\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) \mid \mathcal{D}^{(\ell)}\Big]$$

$$- \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0) \mid \mathcal{D}^{(\ell)}\Big]$$

$$= o_p(1/\sqrt{n}). \tag{5}$$

Here, the LHS of the first inequality is referred to as the empirical process term, while the LHS of the second inequality is referred to as the second-order remainder term.

I.1 Proof of (4)

Proof. We aim to show that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Bigg(\sqrt{\overline{n}}\Bigg|\frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0)$$

$$- \frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0)$$

$$- \Bigg(\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) \mid \mathcal{D}^{(\ell)}\Big]$$

$$- \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0) \mid \mathcal{D}^{(\ell)}\Big]\Bigg)\Bigg| > \varepsilon\Bigg)$$

$$= 0. \tag{6}$$

We show (6) by showing that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Bigg(\sqrt{\overline{n}}\Bigg|\frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0)$$

$$- \frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0)$$

$$- \Bigg(\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) \mid \mathcal{D}^{(\ell)}\Big]$$

$$- \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0) \mid \mathcal{D}^{(\ell)}\Big]\Bigg)\Bigg| \geq \varepsilon \mid \mathcal{D}^{(\ell)}\Bigg)$$

$$= 0. \tag{7}$$

If (7) holds, then (6) also holds from dominated convergence theorem.

We prove (7) using Chebychev's inequality. From Chebychev's inequality we have

$$\mathbb{P}\Bigg(\sqrt{\overline{n}}\Bigg|\frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) - \frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0)$$

$$- \Bigg(\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) \mid \mathcal{D}^{(\ell)}\Big]$$

$$- \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0) \mid \mathcal{D}^{(\ell)}\Big]\Bigg)\Bigg| \geq \varepsilon \mid \mathcal{D}^{(\ell)}\Bigg)$$

$$
\begin{aligned}
&\leq \frac{\overline{n}}{\varepsilon}\mathrm{Var}\Bigg(\frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0) - \frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\widehat{\mu}_{\mathrm{T},n}^{(\ell)},\widehat{\nu}_n^{(\ell)},\widehat{\pi}_n^{(\ell)},g_0)\\
&- \Bigg(\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0)\mid \mathcal{D}^{(\ell)}\Big]\\
&\qquad - \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\widehat{\mu}_{\mathrm{T},n}^{(\ell)},\widehat{\nu}_n^{(\ell)},\widehat{\pi}_n^{(\ell)},g_0)\mid \mathcal{D}^{(\ell)}\Big]\Bigg) \mid \mathcal{D}^{(\ell)}\Bigg).
\end{aligned}
$$

Since observations are i.i.d. and the conditional mean of the target part is zero, we have

$$
\begin{aligned}
&m\mathrm{Var}\Bigg(\frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0) - \frac{1}{\overline{n}}\sum_{i=1}^{m} S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\widehat{\mu}_{\mathrm{T},n}^{(\ell)},\widehat{\nu}_n^{(\ell)},\widehat{\pi}_n^{(\ell)},g_0)\\
&- \Bigg(\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0)\mid \mathcal{D}^{(\ell)}\Big]\\
&\qquad - \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\widehat{\mu}_{\mathrm{T},n}^{(\ell)},\widehat{\nu}_n^{(\ell)},\widehat{\pi}_n^{(\ell)},g_0)\mid \mathcal{D}^{(\ell)}\Big]\Bigg) \mid \mathcal{D}^{(\ell)}\Bigg)\\
&= \mathrm{Var}\Bigg(S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0) - S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\widehat{\mu}_{\mathrm{T},n}^{(\ell)},\widehat{\nu}_n^{(\ell)},\widehat{\pi}_n^{(\ell)},g_0)\\
&- \Bigg(\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0)\mid \mathcal{D}^{(\ell)}\Big]\\
&\qquad - \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\widehat{\mu}_{\mathrm{T},n}^{(\ell)},\widehat{\nu}_n^{(\ell)},\widehat{\pi}_n^{(\ell)},g_0)\mid \mathcal{D}^{(\ell)}\Big]\Bigg) \mid \mathcal{D}^{(\ell)}\Bigg)\\
&= \mathbb{E}\Bigg[\Bigg(S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0) - S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\widehat{\mu}_{\mathrm{T},n}^{(\ell)},\widehat{\nu}_n^{(\ell)},\widehat{\pi}_n^{(\ell)},g_0)\\
&- \Bigg(\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0)\mid \mathcal{D}^{(\ell)}\Big]\\
&\qquad - \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)},\widetilde{O}_i^{(\ell)},\widetilde{Y}_i^{(\ell)};\widehat{\mu}_{\mathrm{T},n}^{(\ell)},\widehat{\nu}_n^{(\ell)},\widehat{\pi}_n^{(\ell)},g_0)\mid \mathcal{D}^{(\ell)}\Big]\Bigg)\Bigg)^2 \mid \mathcal{D}^{(\ell)}\Bigg].
\end{aligned}
\tag{8}
$$

The term (8) converges to zero in probability as $n \to \infty$ if $\left\|\mu_{\mathrm{T},0} - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}\right\|_2 = o_p(1)$, $\left\|\nu_0 - \widehat{\nu}_n^{(\ell)}\right\|_2 = o_p(1)$, and $\left\|\pi_0 - \widehat{\pi}_n^{(\ell)}\right\|_2 = o_p(1)$ as $n \to \infty$. Here, we used the boundedness conditions of each function and the following computation. Them, we complete the proof.

We explain the last step of the above proof below. Let A and B denote the first and second terms in the expectation of (8), respectively. Then, we have

$$
(8) = \mathbb{E}\left[\left(A - B - \mathbb{E}\left[A - B \mid \mathcal{D}^{(\ell)}\right]\right)^2 \mid \mathcal{D}^{(\ell)}\right].
$$

Here, we have

$$
(8) = \mathbb{E}\left[(A-B)^2 \mid \mathcal{D}^{(\ell)}\right] - \left(\mathbb{E}\left[A - B \mid \mathcal{D}^{(\ell)}\right]\right)^2 \leq \mathbb{E}\left[(A-B)^2 \mid \mathcal{D}^{(\ell)}\right].
$$

By showing that $\mathbb{E}\left[(A-B)^2 \mid \mathcal{D}^{(\ell)}\right] = o_p(1)$, we prove the statement. To show $\mathbb{E}\left[(A-B)^2 \mid \mathcal{D}^{(\ell)}\right] = o_p(1)$, we use the following concrete form of S^{cens}:

$$
S^{\mathrm{cens}}(X, O, Y; \mu_T, v, \pi, g)
$$

$$= \frac{\mathbb{1}[O=1](Y-\mu_T(X))}{\pi(1 \mid X)} - \frac{\mathbb{1}[O=0](Y-\nu(X))}{g(0 \mid X)\pi(0 \mid X)} + \frac{g(1 \mid X)\mathbb{1}[O=1](Y-\mu_T(X))}{g(0 \mid X)\pi(1 \mid X)}$$
$$+ \mu_T(X) - \frac{1}{g(0 \mid X)}\nu(X) + \frac{g(1 \mid X)}{g(0 \mid X)}\mu_T(X).$$

Then, we have

$$A - B$$
$$= \frac{\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{\pi_0(1 \mid X)} - \frac{\mathbb{1}[O=0](Y-\nu_0(X))}{g_0(0 \mid X)\pi_0(0 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{g_0(0 \mid X)\pi_0(1 \mid X)}$$
$$+ \mu_{T,0}(X) - \frac{1}{g_0(0 \mid X)}\nu_0(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{T,0}(X)$$
$$- \left(\frac{\mathbb{1}[O=1](Y-\widehat{\mu}^{(\ell)}_{\mathrm{T},n}(X))}{\widehat{\pi}^{(\ell)}_n(1 \mid X)} - \frac{\mathbb{1}[O=0](Y-\widehat{\nu}^{(\ell)}_n(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(0 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\widehat{\mu}^{(\ell)}_{\mathrm{T},n}(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(1 \mid X)} \right.$$
$$\left. + \widehat{\mu}^{(\ell)}_{\mathrm{T},n}(X) - \frac{1}{g_0(0 \mid X)}\widehat{\nu}^{(\ell)}_n(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\widehat{\mu}^{(\ell)}_{\mathrm{T},n}(X) \right)$$

Here, we can show that the following term converges to zero in probability, which follows directly from the convergence in probability of each nuisance-parameter estimator:

$$\left(\mu_{T,0}(X) - \frac{1}{g_0(0 \mid X)}\nu_0(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{T,0}(X) \right)$$
$$- \left(\widehat{\mu}_{T,0}(X) - \frac{1}{g_0(0 \mid X)}\widehat{\nu}^{(\ell)}_n(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\widehat{\mu}^{(\ell)}_{\mathrm{T},n}(X)) \right).$$

Then, we show that the remaining parts converge to zero in probability. Let us denote the parts as

$$(\star) = \frac{\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{\pi_0(1 \mid X)} - \frac{\mathbb{1}[O=0](Y-\nu_0(X))}{g_0(0 \mid X)\pi_0(0 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{g_0(0 \mid X)\pi_0(1 \mid X)}$$
$$- \left(\frac{\mathbb{1}[O=1](Y-\widehat{\mu}^{(\ell)}_{\mathrm{T},n}(X))}{\widehat{\pi}^{(\ell)}_n(1 \mid X)} - \frac{\mathbb{1}[O=0](Y-\widehat{\nu}^{(\ell)}_n(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(0 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\widehat{\mu}^{(\ell)}_{\mathrm{T},n}(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(1 \mid X)} \right).$$

Next, we have

$$(\star) = \frac{\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{\pi_0(1 \mid X)} - \frac{\mathbb{1}[O=0](Y-\nu_0(X))}{g_0(0 \mid X)\pi_0(0 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{g_0(0 \mid X)\pi_0(1 \mid X)}$$
$$- \left(\frac{\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{\widehat{\pi}^{(\ell)}_n(1 \mid X)} - \frac{\mathbb{1}[O=0](Y-\nu_0(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(0 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(1 \mid X)} \right)$$
$$+ \left(\frac{\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{\widehat{\pi}^{(\ell)}_n(1 \mid X)} - \frac{\mathbb{1}[O=0](Y-\nu_0(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(0 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(1 \mid X)} \right)$$
$$- \left(\frac{\mathbb{1}[O=1](Y-\widehat{\mu}^{(\ell)}_{\mathrm{T},n}(X))}{\widehat{\pi}^{(\ell)}_n(1 \mid X)} - \frac{\mathbb{1}[O=0](Y-\widehat{\nu}^{(\ell)}_n(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(0 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\widehat{\mu}^{(\ell)}_{\mathrm{T},n}(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(1 \mid X)} \right).$$

Then, from the parallelogram law, we have

$$(\star)^2 \leq 2 \left(\frac{\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{\pi_0(1 \mid X)} - \frac{\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{\widehat{\pi}^{(\ell)}_n(1 \mid X)} \right)^2$$
$$+ 2 \left(\frac{\mathbb{1}[O=0](Y-\nu_0(X))}{g_0(0 \mid X)\pi_0(0 \mid X)} - \frac{\mathbb{1}[O=0](Y-\nu_0(X))}{g_0(0 \mid X)\widehat{\pi}^{(\ell)}_n(0 \mid X)} \right)^2$$
$$+ \cdots$$

$$+ 2\left(\frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{g_0(0 \mid X)\widehat{\pi}_n^{(\ell)}(1 \mid X)} - \frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X))}{g_0(0 \mid X)\widehat{\pi}_n^{(\ell)}(1 \mid X)}\right)^2.$$

Here, we can bound

$$2\mathbb{E}\left[\left(\frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\mu_{T,0}(X))}{g_0(0 \mid X)\widehat{\pi}_n^{(\ell)}(1 \mid X)} - \frac{g_0(1 \mid X)\mathbb{1}[O=1](Y-\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X))}{g_0(0 \mid X)\widehat{\pi}_n^{(\ell)}(1 \mid X)}\right)^2 \mid \mathcal{D}^{(\ell)}\right]$$

by

$$C\mathbb{E}(\mu_{T,0}(X) - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X))^2],$$

where $C > 0$ is constant independent of n, and we used the boundedness of $\widehat{g}$ and $\widehat{\pi}$. Similarly, we can bound each of the remaining terms. Thus, we complete the proof. □

I.2 Proof of (5)

Proof. We have

$$\begin{aligned}
&\mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \mu_{\mathrm{T},0}, \nu_0, \pi_0, g_0) \mid \mathcal{D}^{(\ell)}\Big] - \mathbb{E}\Big[S^{\mathrm{cens}}(\widetilde{X}_i^{(\ell)}, \widetilde{O}_i^{(\ell)}, \widetilde{Y}_i^{(\ell)}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\nu}_n^{(\ell)}, \widehat{\pi}_n^{(\ell)}, g_0) \mid \mathcal{D}^{(\ell)}\Big] \\
&= \mathbb{E}\Bigg[\frac{\mathbb{1}[O=1]\Big(Y-\mu_{\mathrm{T},0}(X)\Big)}{\pi_0(1 \mid X)} - \frac{\mathbb{1}[O=0]\Big(Y-\nu_0(X)\Big)}{g_0(0 \mid X)\pi(0 \mid X)} + \frac{g_0(1 \mid X)\mathbb{1}[O=1]\Big(Y-\mu_{\mathrm{T},0}(X)\Big)}{g_0(0 \mid X)\pi_0(0 \mid X)} \\
&\qquad + \mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu_0(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X) \mid \mathcal{D}^{(\ell)}\Bigg] \\
&\quad - \mathbb{E}\Bigg[\frac{\mathbb{1}[O=1]\Big(Y-\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\Big)}{\widehat{\pi}_n^{(\ell)}(1 \mid X)} - \frac{\mathbb{1}[O=0]\Big(Y-\widehat{\nu}_n^{(\ell)}(X)\Big)}{g_0(0 \mid X)\widehat{\pi}_n^{(\ell)}(0 \mid X)} \\
&\qquad + \frac{g_0(1 \mid X)\mathbb{1}[O=1]\Big(Y-\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\Big)}{g_0(0 \mid X)\widehat{\pi}_n^{(\ell)}(0 \mid X)} \\
&\qquad + \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X) - \frac{1}{g_0(0 \mid X)}\widehat{\nu}_n^{(\ell)}(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X) \mid \mathcal{D}^{(\ell)}\Bigg] \\
&= \mathbb{E}\left[\mu_{\mathrm{T},0}(X) - \frac{1}{g_0(0 \mid X)}\nu_0(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X)\right] \\
&\quad - \mathbb{E}\Bigg[\frac{\pi_0(1 \mid X)\Big(\mu_{\mathrm{T},0}(X)-\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\Big)}{\widehat{\pi}_n^{(\ell)}(1 \mid X)} - \frac{\pi_0(0 \mid X)\Big(\nu_0(X)-\widehat{\nu}_n^{(\ell)}(X)\Big)}{g_0(0 \mid X)\widehat{\pi}_n^{(\ell)}(0 \mid X)} \\
&\qquad + \frac{g_0(1 \mid X)\pi_0(1 \mid X)\Big(\mu_{\mathrm{T},0}(X)-\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\Big)}{g_0(0 \mid X)\widehat{\pi}_n^{(\ell)}(0 \mid X)} \\
&\qquad + \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X) - \frac{1}{g_0(0 \mid X)}\widehat{\nu}_n^{(\ell)}(X) + \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\Bigg] \\
&= \mathbb{E}\left[\left(1 - \frac{\pi_0(1 \mid X)}{\widehat{\pi}_n^{(\ell)}(1 \mid X)}\right)\Big(\mu_{\mathrm{T},0}(X) - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\Big)\right] \\
&\quad + \mathbb{E}\left[\frac{1}{g_0(0 \mid X)}\widehat{\nu}_n^{(\ell)}(X) - \frac{1}{g_0(0 \mid X)}\nu_0(X) - \frac{\pi_0(0 \mid X)\Big(\widehat{\nu}_n^{(\ell)}(X)-\nu_0(X)\Big)}{g_0(0 \mid X)\widehat{\pi}_n^{(\ell)}(0 \mid X)}\right] \\
&\quad + \mathbb{E}\left[\frac{g_0(1 \mid X)}{g_0(0 \mid X)}\mu_{\mathrm{T},0}(X) - \frac{g_0(1 \mid X)}{g_0(0 \mid X)}\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X) - \frac{g_0(1 \mid X)\pi_0(1 \mid X)\Big(\mu_{\mathrm{T},0}(X)-\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\Big)}{g_0(0 \mid X)\widehat{\pi}_n^{(\ell)}(0 \mid X)}\right]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(1 - \frac{\pi_0(1 \mid X)}{\widehat{\pi}_n^{(\ell)}(1 \mid X)}\right)\left(\mu_{\mathrm{T},0}(X) - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\right)\right] \\
&\quad + \mathbb{E}\left[\frac{1}{g_0(0 \mid X)}\left(1 - \frac{\pi_0(0 \mid X)}{\widehat{\pi}_n^{(\ell)}(0 \mid X)}\right)\left(\widehat{\nu}_n^{(\ell)}(X) - \nu_0(X)\right)\right] \\
&\quad + \mathbb{E}\left[\frac{g_0(1 \mid X)}{g_0(0 \mid X)}\left(1 - \frac{\pi_0(1 \mid X)}{\widehat{\pi}_n^{(\ell)}(0 \mid X)}\right)\left(\mu_{\mathrm{T},0}(X) - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\right)\right] \\
&\leq C\sqrt{\mathbb{E}\left[\left(\widehat{\pi}_n^{(\ell)}(1 \mid X) - \pi_0(1 \mid X)\right)^2\right]\mathbb{E}\left[\left(\mu_{\mathrm{T},0}(X) - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\right)^2\right]} \\
&\quad + C\sqrt{\mathbb{E}\left[\left(\widehat{\pi}_n^{(\ell)}(0 \mid X) - \pi_0(0 \mid X)\right)^2\right]\mathbb{E}\left[\left(\widehat{\nu}_n^{(\ell)}(X) - \nu_0(X)\right)^2\right]} \\
&\quad + C\sqrt{\mathbb{E}\left[\left(\widehat{\pi}_n^{(\ell)}(0 \mid X) - \pi_0(1 \mid X)\right)^2\right]\mathbb{E}\left[\left(\mu_{\mathrm{T},0}(X) - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\right)^2\right]} \\
&= o_p(1/\sqrt{n}),
\end{aligned}
$$

where we used Hölder's inequality. □

I.3 Remark on the double robustness

Indeed, a key structural feature specific to the PUATE setting is that double robustness holds only when the propensity score is correctly specified. That is, in the PUATE framework, double robustness can be stated as follows: given a consistent estimator of the propensity score $g_0(D \mid X)$, if either the observation probability (e.g., $\pi(O \mid X)$) or the conditional expected outcome (e.g., $\mu_{\mathrm{T},0}(X) = \mathbb{E}[Y(1) \mid X]$ and $\mu_{\mathrm{C},0}(X) = \mathbb{E}[Y(0) \mid X]$) is consistently estimated, then the ATE estimator is consistent.

This arises because estimating the conditional expected outcome for the control group, $\mu_{\mathrm{C},0}(X) = \mathbb{E}[Y(0) \mid X]$, requires the propensity score $g_0(D \mid X)$. In the standard ATE setting, $\mathbb{E}[Y(0) \mid X]$ can be estimated directly via regression on outcomes with $D = 0$. However, in the PUATE setting, we must rely on the following identity to recover $\mathbb{E}[Y(0) \mid X]$.

In the standard doubly robust estimator, $E[Y(0)]$ can be estimated by taking the sample mean of

$$
\frac{\mathbb{1}[D_i = 0](Y_i - \widehat{\mu}_{\mathrm{C}}(X_i))}{\widehat{g}(0 \mid X_i)} + \widehat{\mu}_C(X_i).
$$

In contrast, for PUATE under the censoring setting, $E[Y(0)]$ can be estimated by taking the sample mean of

$$
\frac{\mathbb{1}[O_i = 0](Y_i - \widehat{\nu}(X_i))}{\widehat{g}(0 \mid X_i)\widehat{\pi}(0 \mid X_i)} - \frac{\widehat{g}(1 \mid X_i)\mathbb{1}[O_i = 1](Y_i - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X_i))}{\widehat{g}(0 \mid X_i)\widehat{\pi}(1 \mid X_i)} + \frac{1}{\widehat{g}(0 \mid X_i)}\widehat{\nu}(X_i) - \frac{\widehat{g}(1 \mid X_i)}{\widehat{g}(0 \mid X_i)}\widehat{\mu}_{\mathrm{T}}(X_i).
$$

Put differently, in the standard ATE setting, $\mathbb{E}[Y(0)]$ can be estimated either by the sample mean of

$$
\frac{\mathbb{1}[D = 0]Y}{\widehat{g}(0 \mid X)}
$$

or by

$$
\widehat{\mu}_{\mathrm{C}}(X).
$$

One relies solely on the propensity score $g(0 \mid X)$, and the other solely on the outcome model $\widehat{\mu}_C(X)$. However, in the PUATE setting, $E[Y(0)]$ can be estimated either by the sample mean of

$$
\frac{\mathbb{1}[O_i = 0]Y_i}{\widehat{g}(0 \mid X_i)\widehat{\pi}(0 \mid X_i)} - \frac{\widehat{g}(1 \mid X_i)\mathbb{1}[O_i = 1]Y_i}{\widehat{g}(0 \mid X_i)\widehat{\pi}(1 \mid X_i)}
$$

or

$$
\frac{1}{\widehat{g}(0 \mid X_i)}\widehat{\nu}(X_i) - \frac{\widehat{g}(1 \mid X_i)}{\widehat{g}(0 \mid X_i)}\widehat{\mu}_{\mathrm{T}}(X_i).
$$

Both estimators depend on the estimated propensity score $\widehat{g}$.

This highlights the intuition behind the lack of double robustness in the PUATE setting: to estimate (or identify) $\mathbb{E}[Y(0) \mid X]$, one must necessarily rely on an estimator of g_0. This structural dependency underlies why double robustness, in the usual sense, does not hold in PUATE. Even in cases where the treatment indicator contains missing values, if $\mathbb{E}[Y(0) \mid X]$ can be estimated without relying on $\widehat{g}$, it is still possible to construct a doubly robust estimator involving the product of the propensity score error and the outcome model error. However, in the PUATE setting, we observe only treated units and a mixture of treated and control units. This makes the estimation of $\mathbb{E}[Y(0) \mid X]$ particularly challenging. Also see Dahabreh et al. (2019).

J Proof of Lemma D.2

Our proof is inspired by those in Uehara et al. (2020) and Kato et al. (2024). Uehara et al. (2020) revisits the efficiency bound under the stratified sampling scheme, a generalization of the case-control setting, studied by Wooldridge (2001) and Imbens & Wooldridge (2009). In the stratified sampling, we define an efficiency bound by regarding $\left(\mathcal{D}_{\mathrm{T}}, \mathcal{D}_{\mathrm{U}}\right)$ as one sample.

Their proof considers a nonparametric model for the distribution of potential outcomes and defines regular subparametric models. Then, (i) we characterize the tangent set for all regular parametric submodels, (ii) verify that the parameter of interest is pathwise differentiable, (iii) verify that a guessed semiparametric efficient influence function lies in the tangent set, and (iv) calculate the expected square of the influence function.

In the case-control setting, the observations are generated as follows:

$$\mathcal{D}_{\mathrm{T}} \coloneqq \left\{\left(X_{\mathrm{T},j}, Y_j(1)\right)\right\}_{j=1}^{m}, \quad \left(X_{\mathrm{T},j}, Y_j(1)\right) \sim p_{\mathrm{T},0}(x, y(1)) = \zeta_{\mathrm{T},0}(x) p_{Y(1),0}(y(1) \mid x),$$
$$\mathcal{D}_{\mathrm{U}} \coloneqq \left\{\left(X_k, Y_{\mathrm{U},k}\right)\right\}_{k=1}^{l}, \quad \left(X_k, Y_{\mathrm{U},k}\right) \sim p_{\mathrm{U},0}(x, y_{\mathrm{U}}) = \zeta_0(x) p_{Y_{\mathrm{U}},0}(y_{\mathrm{U}} \mid x).$$

We derive the efficiency bound by regarding

$$\mathcal{E} = \left(\mathcal{D}_{\mathrm{T}}, \mathcal{D}_{\mathrm{U}}\right)$$

as one observation.

We define regular parametric submodels

$$\mathcal{P}^{\mathrm{sub}} \coloneqq \{P_{\mathrm{T},\theta}, P_{\mathrm{U},\theta} \colon \theta \in \mathbb{R}\},$$

where $P_{\mathrm{T},\theta}$ is a parametric submodel for the distribution of $X_{\mathrm{T},j}, Y_j(1)$ and $P_{\mathrm{U},\theta}$ is a parametric submodel for the distribution of $X_k, Y_{\mathrm{U},k}$.

We denote the probability densities under $P_{\mathrm{T},\theta}$ and $P_{\mathrm{U},\theta}$ by

$$p_{\mathrm{T}}(x, y; \theta) = \zeta_{\mathrm{T}}(x; \theta) p_{Y(1),0}(y(1) \mid x; \theta),$$
$$p_{\mathrm{U}}(x, y; \theta) = \zeta(x; \theta) p_{Y_{\mathrm{U}}}(y_{\mathrm{U}} \mid x; \theta).$$

We consider the joint log-likelihood of $\mathcal{D}_T$ and $\mathcal{D}_S$, which is defined as

$$\sum_{j=1}^{m} \log\left(p_{\mathrm{T}}(X_{\mathrm{T},j}, Y_j(1); \theta)\right) + \sum_{k=1}^{l} \log\left(p_{\mathrm{U}}(X_k, Y_{\mathrm{U},k}; \theta)\right).$$

By taking the derivative of $\sum_{j=1}^{m} \log\left(p_{\mathrm{T}}(X_{\mathrm{T},j}, Y_j(1); \theta)\right) + \sum_{k=1}^{l} \log\left(p_{\mathrm{U}}(X_k, Y_{\mathrm{U},k}; \theta)\right)$ with respect to β, we can obtain the corresponding score as

$$S(\mathcal{E}; \theta) \coloneqq \frac{\mathrm{d}}{\mathrm{d}\theta}\left(\sum_{j=1}^{m} \log\left(p_{\mathrm{T}}(X_{\mathrm{T},j}, Y_j(1); \theta)\right) + \sum_{k=1}^{l} \log\left(p_{\mathrm{U}}(X_k, Y_{\mathrm{U},k}; \theta)\right)\right)$$
$$= \sum_{j=1}^{m} S_{X_{\mathrm{T}}}(X_{\mathrm{T},j}; \theta) + \sum_{j=1}^{m} S_{Y(1)}(Y_j(1) \mid X_{\mathrm{T},j}; \theta) + \sum_{k=1}^{l} S_X(X_k; \theta) + \sum_{k=1}^{l} S_{Y_{\mathrm{U}}}(Y_{\mathrm{U},k} \mid X_k; \theta).$$

where

$$S_{X_{\mathrm{T}}}(x;\theta) \coloneqq \frac{d}{d\theta}\log \zeta_{\mathrm{T}}(x;\theta),$$
$$S_{Y(1)}(y \mid x;\theta) \coloneqq \frac{d}{d\theta}\log p_{Y(1)}(y \mid x;\theta),$$
$$S_{X}(x;\theta) \coloneqq \frac{d}{d\theta}\log \zeta(x;\theta),$$
$$S_{Y_{\mathrm{U}}}(y \mid x;\theta) \coloneqq \frac{d}{d\theta}\log p_{Y_{\mathrm{U}}}(y \mid x;\theta),$$

Let us also define

$$S^{\mathrm{cc\ (T)}}(x,y;\mu_{\mathrm{T}},e,r) = S_{X_{\mathrm{T}}}(x;\theta) + S_{Y(1)}(y \mid x;\theta),$$
$$S^{\mathrm{cc\ (U)}}(X,Y_{\mathrm{U}};\mu_{\mathrm{T}},\mu_{\mathrm{U}},e) = S_{X}(x;\theta) + S_{Y_{\mathrm{U}}}(y \mid x;\theta).$$

Here, note that

$$p_{Y_{\mathrm{U}}}(y \mid x;\theta) = e_0(1 \mid x)p_{Y(1)}(y \mid x;\theta) + e_0(0 \mid x)p_{Y(0)}(y \mid x;\theta).$$

We have

$$p_{Y(0)}(y \mid x;\theta) = \frac{1}{e_0(0 \mid x)}\Big(p_{Y_{\mathrm{U}}}(y \mid x;\theta) - e_0(1 \mid x)p_{Y(1)}(y \mid x;\theta)\Big)$$

Using this relationship, we write the ATE under the parametric submodels as

$$\begin{aligned}\tau(\theta) &\coloneqq \iint y(1)p_{Y(1)}(y(1) \mid x;\theta)\zeta(x;\theta)\mathrm{d}y(1)\mathrm{d}x - \iint y(0)p_{Y(0)}(y(0) \mid x;\theta)\zeta(x;\theta)\mathrm{d}y(0)\mathrm{d}x \\ &= \iint y(1)p_{Y(1)}(y(1) \mid x;\theta)\zeta(x;\theta)\mathrm{d}y(1)\mathrm{d}x - \iint y(0)\frac{1}{e_0(0 \mid x)}p_{Y_{\mathrm{U}}}(y(0) \mid x;\theta)\zeta(x;\theta)\mathrm{d}y(0)\mathrm{d}x \\ &\qquad + \iint y(0)\frac{e_0(1 \mid x)}{e_0(0 \mid x)}p_{Y(1)}(y(0) \mid x;\theta)\zeta(x;\theta)\mathrm{d}y(0)\mathrm{d}x.\end{aligned}$$

The tangent space for this parametric submodel at $\theta = \theta_0$ is given as

$$\mathcal{T} \coloneqq$$
$$\left\{\sum_{j=1}^{m} S_{X_{\mathrm{T}}}(X_{\mathrm{T},j};\theta_0) + \sum_{j=1}^{m} S_{Y(1)}(Y_j(1) \mid X_{\mathrm{T},j};\theta_0) + \sum_{k=1}^{l} S_X(X_k;\theta_0) + \sum_{k=1}^{l} S_{Y_{\mathrm{U}}}(Y_{\mathrm{U},k} \mid X_k;\theta_0) \in L_2(\mathcal{E})\right\}.$$

From the Riesz representation theorem, there exists a function $\widetilde{\Psi}$ such that

$$\frac{\partial \tau(\theta)}{\partial \theta}\Big|_{\theta=\theta_0} = \mathbb{E}\big[\widetilde{\Psi}(\mathcal{E})S(\mathcal{E};\theta_0)\big]. \tag{9}$$

Here, we use the assumption that each random variable has finite variance, which guarantees the existence of the function.

There exists a unique function Ψ^{cc} such that $\Psi^{\mathrm{cc}} \in \mathcal{T}$, called the efficient influence function. We specify the efficient influence function as

$$\begin{aligned}&\widetilde{\Psi}^{\mathrm{cc}}(\mathcal{E};\mu_{\mathrm{T},0},\mu_{\mathrm{U},0},e_0,r_0), \\ &= \frac{1}{m}\sum_{j=1}^{m}\left(\left(1 - \frac{e_0(1 \mid X_{\mathrm{T},j})}{e_0(0 \mid X_{\mathrm{T},j})}\right)\Big(Y_j(1) - \mu_{\mathrm{T},0}(X)\Big)\right)r_0(X_{\mathrm{T},j}), \\ &\quad + \frac{1}{l}\sum_{k=1}^{l}\left(-\frac{\Big(Y_{\mathrm{U},k} - \mu_{\mathrm{U},0}(X_k)\Big)}{e_0(0 \mid X_k)} + \mu_{\mathrm{T},0}(X_k) - \frac{1}{e_0(0 \mid X_k)}\mu_{\mathrm{U},0}(X) + \frac{e_0(1 \mid X_k)}{e_0(0 \mid X_k)}\mu_{\mathrm{T},0}(X_k)\right) - \tau_0.\end{aligned}$$

We prove that $\widetilde{\Psi}^{\mathrm{cc}}(X,O,Y;\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0)$ is actually the unique efficient influence function by verifying that $\widetilde{\Psi}^{\mathrm{cc}}$ satisfies (9) and $\widetilde{\Psi}^{\mathrm{cc}} \in \mathcal{T}$.

Proof of (9): First, we confirm that $\widetilde{\Psi}^{\mathrm{cc}}$ satisfies (9). We have

$$
\begin{aligned}
&\mathbb{E}\left[\widetilde{\Psi}^{\mathrm{cc}}(\mathcal{E};\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0)S(\mathcal{E};\theta_0)\right] \\
&= \mathbb{E}\Bigg[\Psi^{\mathrm{cc}}(\mathcal{E};\mu_{\mathrm{T},0},\nu_0,\pi_0,g_0) \\
&\quad\cdot\left(\sum_{j=1}^{m} S_{X_{\mathrm{T}}}(X_{\mathrm{T},j};\theta_0)+\sum_{j=1}^{m} S_{Y(1)}(Y_j(1)\mid X_{\mathrm{T},j};\theta_0)+\sum_{k=1}^{l} S_X(X_k;\theta_0)+\sum_{k=1}^{l} S_{Y_{\mathrm{U}}}(Y_{\mathrm{U},k}\mid X_k;\theta_0)\right)\Bigg] \\
&= \mathbb{E}\Bigg[\left(\frac{1}{m}\sum_{j=1}^{m}\left(\left(1-\frac{e_0(1\mid X_{\mathrm{T},j})}{e_0(0\mid X_{\mathrm{T},j})}\right)\left(Y_j(1)-\mu_{\mathrm{T}}(X_{\mathrm{T},j})\right)\right) r_0(X_{\mathrm{T},j})\right. \\
&\quad\left.+\frac{1}{l}\sum_{k=1}^{l}\left(-\frac{\left(Y_{\mathrm{U},k}-\mu_{\mathrm{U},0}(X_k)\right)}{e_0(0\mid X_k)}+\mu_{\mathrm{T},0}(X_k)-\frac{1}{e_0(0\mid X_k)}\mu_{\mathrm{U}}(X_k)+\frac{e_0(1\mid X_k)}{e_0(0\mid X_k)}\mu_{\mathrm{T},0}(X_k)\right)-\tau_0\right) \\
&\quad\cdot\left(\sum_{j=1}^{m} S_{X_{\mathrm{T}}}(X_{\mathrm{T},j};\theta_0)+\sum_{j=1}^{m} S_{Y(1)}(Y_j(1)\mid X_{\mathrm{T},j};\theta_0)+\sum_{k=1}^{l} S_X(X_k;\theta_0)+\sum_{k=1}^{l} S_{Y_{\mathrm{U}}}(Y_{\mathrm{U},k}\mid X_k;\theta_0)\right)\Bigg].
\end{aligned}
$$

Since $\mathcal{D}_{\mathrm{T}}$ and $\mathcal{D}_{\mathrm{U}}$ are independent and observations are i.i.d., we have

$$
\begin{aligned}
&\mathbb{E}\Bigg[\left(\frac{1}{m}\sum_{j=1}^{m}\left(\left(1-\frac{e_0(1\mid X_{\mathrm{T},j})}{e_0(0\mid X_{\mathrm{T},j})}\right)\left(Y_j(1)-\mu_{\mathrm{T}}(X_{\mathrm{T},j})\right)\right) r_0(X_{\mathrm{T},j})\right. \\
&\quad\left.+\frac{1}{l}\sum_{k=1}^{l}\left(-\frac{\left(Y_{\mathrm{U},k}-\mu_{\mathrm{U},0}(X_k)\right)}{e_0(0\mid X_k)}+\mu_{\mathrm{T},0}(X_k)-\frac{1}{e_0(0\mid X_k)}\mu_{\mathrm{U}}(X_k)+\frac{e_0(1\mid X_k)}{e_0(0\mid X_k)}\mu_{\mathrm{T},0}(X_k)-\tau_0\right)\right) \\
&\quad\cdot\left(\sum_{j=1}^{m} S_{X_{\mathrm{T}}}(X_{\mathrm{T},j};\theta_0)+\sum_{j=1}^{m} S_{Y(1)}(Y_j(1)\mid X_{\mathrm{T},j};\theta_0)+\sum_{k=1}^{l} S_X(X_k;\theta_0)+\sum_{k=1}^{l} S_{Y_{\mathrm{U}}}(Y_{\mathrm{U},k}\mid X_k;\theta_0)\right)\Bigg] \\
&= \mathbb{E}\left[\left(1-\frac{e_0(1\mid X_{\mathrm{T},j})}{e_0(0\mid X_{\mathrm{T},j})}\right)\left(Y_j(1)-\mu_{\mathrm{T},0}(X_{\mathrm{T},j})\right) r_0(X_{\mathrm{T},j})\left(S_{X_{\mathrm{T}}}(X_{\mathrm{T},j};\theta_0)+S_{Y(1)}(Y_j(1)\mid X_{\mathrm{T},j};\theta_0)\right)\right] \\
&\quad+\mathbb{E}\Bigg[\left(-\frac{\left(Y_{\mathrm{U},k}-\mu_{\mathrm{U},0}(X_k)\right)}{e_0(0\mid X_k)}+\mu_{\mathrm{T},0}(X_k)-\frac{1}{e_0(0\mid X_k)}\mu_{\mathrm{U},0}(X_k)+\frac{e_0(1\mid X_k)}{e_0(0\mid X_k)}\mu_{\mathrm{T},0}(X_k)-\tau_0\right) \\
&\qquad\cdot\left(S_X(X_k;\theta_0)+S_{Y_{\mathrm{U}}}(Y_{\mathrm{U},k}\mid X_k;\theta_0)\right)\Bigg].
\end{aligned}
$$

Because the density ratio allows us to change the measure, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(1-\frac{e_0(1\mid X_{\mathrm{T},j})}{e_0(0\mid X_{\mathrm{T},j})}\right)\left(Y_j(1)-\mu_{\mathrm{T},0}(X_{\mathrm{T},j})\right) r_0(X_{\mathrm{T},j})\left(S_{X_{\mathrm{T}}}(X_{\mathrm{T},j};\theta_0)+S_{Y(1)}(Y_j(1)\mid X;\theta_0)\right)\right] \\
&= \mathbb{E}\left[\left(1-\frac{e_0(1\mid X)}{e_0(0\mid X)}\right)\left(Y(1)-\mu_{\mathrm{T},0}(X)\right)\left(S_X(X_{\mathrm{T},j};\theta_0)+S_{Y(1)}(Y(1)\mid X;\theta_0)\right)\right]
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(1-\frac{e_0(1\mid X)}{e_0(0\mid X)}\right)\left(Y(1)-\mu_{\mathrm{T}}(X)\right)\left(S_X(X;\theta_0)+S_{Y(1)}(Y(1)\mid X;\theta_0)\right)\right] \\
&\quad+\mathbb{E}\left[\left(-\frac{\left(Y_{\mathrm{U}}-\mu_{\mathrm{U},0}(X)\right)}{e_0(0\mid X)}+\mu_{\mathrm{T},0}(X)-\frac{1}{e_0(0\mid X)}\mu_{\mathrm{U}}(X)+\frac{e_0(1\mid X)}{e_0(0\mid X)}\mu_{\mathrm{T},0}(X)-\tau_0\right)\right.
\end{aligned}
$$

$$
\begin{aligned}
&\quad \cdot \left(S_X(X;\theta_0) + S_{Y_{\mathrm{U}}}(Y_{\mathrm{U}} \mid X;\theta_0)\right)\Big] \\
&= \mathbb{E}_\theta\Big[Y(1)S_{Y(1)}(Y(1) \mid X;\theta_0)\Big] - \mathbb{E}_\theta\left[\frac{1}{e_0(0 \mid X)}Y_{\mathrm{U}}S_{Y_{\mathrm{U}}}(Y_{\mathrm{U}} \mid X;\theta_0)\right] \\
&\quad + \mathbb{E}_\theta\left[\frac{e_0(1 \mid X)}{e_0(0 \mid X)}Y(1)S_{Y(1)}(Y(1) \mid X;\theta_0)\right] \\
&\quad + \mathbb{E}_\theta\Big[\tau(X;\theta_0)S_X(X;\theta_0)\Big] \\
&= \frac{\partial \tau(\theta)}{\partial \theta}\Big|_{\theta=\theta_0},
\end{aligned}
$$

where we defined $\mu_{\mathrm{T},0}(X) - \frac{1}{e_0(0|X)}\mu_{\mathrm{U}}(X) + \frac{e_0(1|X)}{e_0(0|X)}\mu_{\mathrm{T},0}(X) = \tau(X;\theta_0)$, and from the first to the second equality, we used the followings:

$$
\begin{aligned}
&\mathbb{E}\left[\left(1 - \frac{e_0(1 \mid X)}{e_0(0 \mid X)}\right)\Big(Y(1) - \mu_{\mathrm{T}}(X)\Big)S_X(X;\theta_0) \mid X\right] \\
&\qquad = \left(1 - \frac{e_0(1 \mid X)}{e_0(0 \mid X)}\right) S_X(X;\theta_0)\mathbb{E}\left[Y(1) - \mu_{\mathrm{T}}(X) \mid X\right] = 0, \\
&\mathbb{E}\left[\frac{\Big(Y_{\mathrm{U}} - \mu_{\mathrm{U},0}(X)\Big)}{e_0(0 \mid X)}S_X(X;\theta_0)\right] = 0, \\
&\mathbb{E}\left[\tau_0\left(S_X(X;\theta_0) + S_{Y_{\mathrm{U}}}(Y_{\mathrm{U}} \mid X;\theta_0)\right)\right] = 0.
\end{aligned}
$$

Proof of $\widetilde{\Psi}^{\mathrm{cc}} \in \mathcal{T}$: Set

$$
\begin{aligned}
&S^{\mathrm{cc}\ (\mathrm{T})}(X, Y(1); \mu_{\mathrm{T},0}, e_0, r_0) = \left(\left(1 - \frac{e_0(1 \mid X)}{e_0(0 \mid X)}\right)\Big(Y(1) - \mu_{\mathrm{T},0}(X)\Big)\right) r(X), \\
&S^{\mathrm{cc}\ (\mathrm{U})}(X, Y_{\mathrm{U}}; \mu_{\mathrm{T},0}, \mu_{\mathrm{U},0}, e_0) = \frac{\Big(Y_{\mathrm{U}} - \mu_{\mathrm{U},0}(X)\Big)}{e_0(0 \mid X)} + \mu_{\mathrm{T},0}(X) - \frac{1}{e_0(0 \mid X)}\mu_{\mathrm{U},0}(X) + \frac{e_0(1 \mid X)}{e_0(0 \mid X)}\mu_{\mathrm{T},0}(X),
\end{aligned}
$$

Then, $\widetilde{\Psi}^{\mathrm{cc}} \in \mathcal{T}$ holds.

K Proof of Theorem D.8: : Semiparametric efficient ATE estimator under the case-control setting

Recall that we have defined the ATE estimators as

$$
\widehat{\tau}_n^{\text{cc-eff}} = \frac{1}{m}\sum_{j=1}^{m} S^{\mathrm{cc}\ (\mathrm{T})}(X_j, Y_j; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{e}_n^{(\ell)}, \widehat{r}_n^{(\ell)}) + \frac{1}{l}\sum_{k=1}^{l} S^{\mathrm{cc}\ (\mathrm{U})}(X_k, Y_k; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\mu}_{\mathrm{U},n}^{(\ell)}, \widehat{e}_n^{(\ell)}).
$$

We aim to show

$$
\sqrt{n}\left(\widehat{\tau}_n^{\text{cc-eff}} - \tau_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}(0, V^{\mathrm{cc}}) \quad \text{as } n \to \infty.
$$

Recall that

$$
\begin{aligned}
&S^{\mathrm{cc}\ (\mathrm{T})}(X, Y(1); \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{e}_n^{(\ell)}, \widehat{r}_n^{(\ell)}) \\
&\quad = \left(1 - \frac{\widehat{e}_n^{(\ell)}(1 \mid X)}{\widehat{e}_n^{(\ell)}(0 \mid X)}\right)\Big(Y(1) - \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X)\Big)\widehat{r}_n^{(\ell)}(X), \\
&S^{\mathrm{cc}\ (\mathrm{U})}(X, Y_{\mathrm{U}}; \widehat{\mu}_{\mathrm{T},n}^{(\ell)}, \widehat{\mu}_{\mathrm{U},n}^{(\ell)}, \widehat{e}_n^{(\ell)}) \\
&\quad = \frac{\Big(Y_{\mathrm{U}} - \widehat{\mu}_{\mathrm{U},n}^{(\ell)}(X)\Big)}{\widehat{e}_n^{(\ell)}(0 \mid X)} + \widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X) - \frac{1}{\widehat{e}_n^{(\ell)}(0 \mid X)}\widehat{\mu}_{\mathrm{U},n}^{(\ell)}(X) + \frac{\widehat{e}_n^{(\ell)}(1 \mid X)}{\widehat{e}_n^{(\ell)}(0 \mid X)}\widehat{\mu}_{\mathrm{T},n}^{(\ell)}(X).
\end{aligned}
$$

We have

$$\begin{aligned}
\widehat{\tau}_n^{\text{cc-eff}} &= \frac{1}{m}\sum_{j=1}^{m} S^{\text{cc (T)}}(X_j, Y_j; \widehat{\mu}_{\text{T},n}^{(\ell)}, \widehat{e}_n^{(\ell)}, \widehat{r}_n^{(\ell)}) + \frac{1}{l}\sum_{k=1}^{l} S^{\text{cc (U)}}(X_k, Y_k; \widehat{\mu}_{\text{U},n}^{(\ell)}, \widehat{e}_n^{(\ell)}) \\
&= \frac{1}{m}\sum_{j=1}^{m} S^{\text{cc (T)}}(X_j, Y_j; \mu_{\text{T},0}, e_0, r_0) + \frac{1}{l}\sum_{k=1}^{l} S^{\text{cc (U)}}(X_k, Y_k; \mu_{\text{U},0}, e_0) \\
&\quad - \frac{1}{m}\sum_{j=1}^{m} S^{\text{cc (T)}}(X_j, Y_j; \mu_{\text{T},0}, e_0, r_0) - \frac{1}{l}\sum_{k=1}^{l} S^{\text{cc (U)}}(X_k, Y_k; \mu_{\text{U},0}, e_0) \\
&\quad + \frac{1}{m}\sum_{j=1}^{m} S^{\text{cc (T)}}(X_j, Y_j; \widehat{\mu}_{\text{T},n}^{(\ell)}, \widehat{e}_n^{(\ell)}, \widehat{r}_n^{(\ell)}) + \frac{1}{l}\sum_{k=1}^{l} S^{\text{cc (U)}}(X_k, Y_k; \widehat{\mu}_{\text{U},n}^{(\ell)}, \widehat{e}_n^{(\ell)}).
\end{aligned}$$

Here, if it holds that

$$\frac{1}{m}\sum_{j=1}^{m} S^{\text{cc (T)}}(X_j, Y_j; \widehat{\mu}_{\text{T},n}^{(\ell)}, \widehat{e}_n^{(\ell)}, \widehat{r}_n^{(\ell)}) - \frac{1}{m}\sum_{j=1}^{m} S^{\text{cc (T)}}(X_j, Y_j; \mu_{\text{T},0}, e_0, r_0) = o_p(1/\sqrt{m}), \quad (10)$$

$$\frac{1}{l}\sum_{k=1}^{l} S^{\text{cc (U)}}(X_k, Y_k; \widehat{\mu}_{\text{U},n}^{(\ell)}, \widehat{e}_n^{(\ell)}) - \frac{1}{l}\sum_{k=1}^{l} S^{\text{cc (U)}}(X_k, Y_k; \mu_{\text{U},0}, e_0) = o_p(1/\sqrt{l}). \quad (11)$$

then we have

$$\begin{aligned}
&\sqrt{n}\Big(\widehat{\tau}_n^{\text{cc-eff}} - \tau_0\Big) \\
&= \sqrt{n}\frac{1}{m}\sum_{j=1}^{m} S^{\text{cc (T)}}(X_j, Y_j; \mu_{\text{T},0}, e_0, r_0) + \sqrt{n}\frac{1}{l}\sum_{k=1}^{l} S^{\text{cc (U)}}(X_k, Y_k; \mu_{\text{U},0}, e_0) + o_p(1) \\
&= \frac{1}{\sqrt{\alpha m}}\sum_{j=1}^{m} S^{\text{cc (T)}}(X_j, Y_j; \mu_{\text{T},0}, e_0, r_0) + \frac{1}{\sqrt{(1-\alpha)l}}\sum_{k=1}^{l} S^{\text{cc (U)}}(X_k, Y_k; \mu_{\text{U},0}, e_0) + o_p(1) \\
&\xrightarrow{\text{d}} \mathcal{N}(0, V^{\text{cc}}),
\end{aligned}$$

from the central limit theorem for i.i.d. random variables.

Therefore, we prove Theorem D.8 by establishing (10) and (11). These inequalities can be proved in the same manner as the proof of Theorem 4.8 and the analysis of double machine learning under the stratified scheme presented in Uehara et al. (2020). Since the procedure is nearly identical, we omit further details.

L Additional results of the simulation studies

This section investigates the case in which the expected outcomes and propensity scores follow non-linear models.

All experiments were conducted on a Mac computer equipped with an Apple M2 processor and 24 GB of RAM.

L.1 Censoring setting

We generate synthetic data under the censoring setting, where the covariates X are drawn from a multivariate normal distribution as $X \sim \zeta_0(x)$, where $\zeta_0(x)$ is the density of $\mathcal{N}(0, I_p)$, p denotes the number of covariates, and I_p is the $(p \times p)$ identity matrix. We set $p = 10$. The propensity score is given by $g_0(1 \mid X) = \text{sigmoid}(X^\top \beta_1 + X^{2\top}\beta_2)$, where X^2 is the element-wise square of X, and β_1 and β_2 are coefficient vectors sampled from $\mathcal{N}(0, 0.5 I_{2p})$. The treatment indicator D is sampled according to the propensity score.

Table 2: Experimental results. The upper and lower tables are results in the censoring and case-control settings, respectively.

Censoring	IPW	DM	Efficient	IPW	DM	Efficient
	(estimated g_0)			(true g_0)		
MSE	6.86	0.51	0.28	2.30	0.17	0.21
Bias	-1.60	0.40	0.22	0.33	0.10	0.04
Cov. ratio	0.81	0.18	0.76	0.96	0.29	0.94

Case-control	IPW	DM	Efficient	IPW	DM	Efficient
	(estimated e_0)			(true e_0)		
MSE	1.06	0.09	0.10	0.35	0.03	0.03
Bias	-0.03	0.19	0.18	-0.00	-0.01	-0.01
Cov. ratio	0.93	0.40	0.61	0.97	0.77	0.91

Table 3: Experimental results. The upper and lower tables are results in the censoring and case-control settings, respectively.

Censoring	IPW	DM	Efficient	IPW	DM	Efficient
	(estimated g_0)			(true g_0)		
MSE	5.03	0.23	0.13	1.25	0.07	0.09
Bias	-1.32	0.24	0.18	0.17	0.07	0.04
Cov. ratio	0.91	0.22	0.82	0.99	0.34	0.98

Case-control	IPW	DM	Efficient	IPW	DM	Efficient
	(estimated e_0)			(true e_0)		
MSE	0.40	0.03	0.03	0.23	0.01	0.01
Bias	-0.09	0.10	0.11	0.00	-0.02	-0.00
Cov. ratio	0.99	0.69	0.82	0.99	0.92	0.98

Censoring	IPW	DM	Efficient
MSE	297.34	6.38	5.19
Bias	-16.36	-0.58	0.56
Cov. ratio	0.00	0.10	0.22

Case-control	IPW	DM	Efficient
MSE	26.58	1.18	1.49
Bias	2.36	0.52	0.77
Cov. ratio	0.42	0.29	0.40

Table 4: Response surface A. Left: censoring setting; Right: case-control setting.

The observation indicator O is generated such that $O_i = 1$ with probability c if $D_i = 1$, and $O_i = 0$ if $D_i = 0$, where c is drawn from a uniform distribution over $[0, 1]$ before the experiment begins. The outcome is generated as $Y = (X^\top \beta)^2 + 1.1 + \tau_0 \cdot D + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, 1)$ and we set $\tau_0 = 3$.

The nuisance parameters are estimated using three-layer perceptrons with hidden layers of 100 nodes. The convergence rates satisfy Assumption 4.7 under standard conditions (Schmidt-Hieber, 2020). We compare our proposed estimator, $\widehat{\tau}_n^{\text{cens-eff}}$, with two alternative estimators: the IPW estimator $\widehat{\tau}_n^{\text{cens-IPW}}$ and the DM estimator $\widehat{\tau}_n^{\text{cens-DM}}$, as defined in Remarks 4.4 and 4.4, respectively. Note that all of these estimators are proposed by us, and our objective is not to demonstrate that $\widehat{\tau}_n^{\text{cens-eff}}$ outperforms the others, although we recommend it in practice. We consider both cases in which the propensity score is either estimated using the method proposed by Elkan & Noto (2008) or assumed to be known.

We set $n = 3000$. We conduct 5000 trials and report the empirical mean squared errors (MSEs) and biases for the true ATE, as well as the coverage ratio computed from the confidence intervals in Table 2 for $n = 3000$ and Table 3 for $n = 500$. We also present the empirical distributions of the ATE estimates in Figure 3 for $n = 3000$ and Figure 4 for $n = 5000$.

L.2 Case-control setting

In the case-control setting, covariates for the treatment and unknown groups are generated from $p =$ 3-dimensional normal distributions: $X_{\mathrm{T}} \sim \zeta_{\mathrm{T},0}(x)$ and $X \sim \zeta_0(x) = e_0(1)\zeta_{\mathrm{T},0}(x) + e_0(0)\zeta_{\mathrm{C}}(x)$, where $\zeta_{\mathrm{T},0}(x)$ and $\zeta_{\mathrm{C}}(x)$ are the densities of the normal distributions $\mathcal{N}(\mu_p \mathbf{1}_p, I_p)$ and $\mathcal{N}(\mu_n \mathbf{1}_p, I_p)$, respectively. We set $\mu_p = 0.5$, $\mu_n = 0$, and $\mathbf{1}_p = (1\ 1\ \cdots\ 1)^\top$. The class prior is set as $e_0(1) = 0.3$.

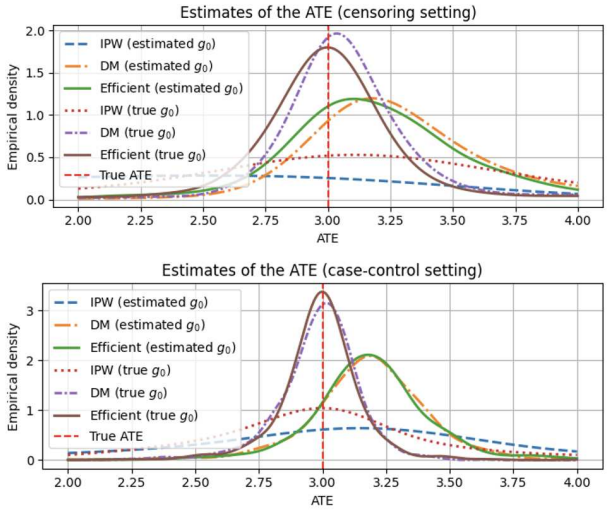

Figure 3: Empirical distributions of ATE estimates.

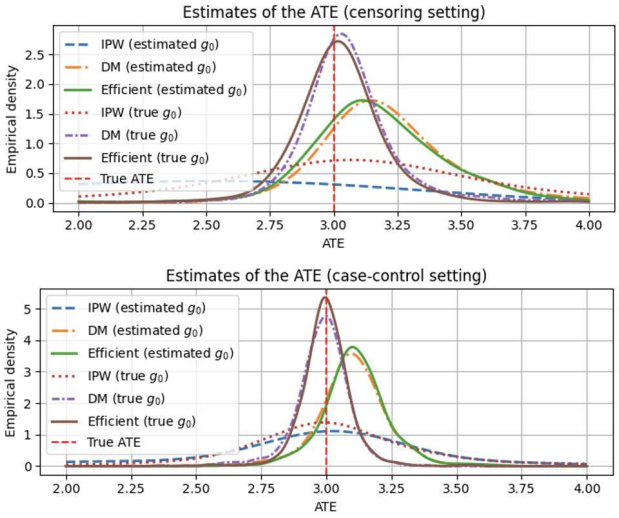

Figure 4: Empirical distributions of ATE estimates.

Censoring	IPW	DM	Efficient
MSE	327.49	4.15	1.14
Bias	-17.52	-1.58	-0.28
Cov. ratio	0.00	0.00	0.01

Case-control	IPW	DM	Efficient
MSE	46.15	3.34	3.77
Bias	2.66	0.41	0.93
Cov. ratio	0.42	0.21	0.43

Table 5: Response surface B. Left: censoring setting; Right: case-control setting.

By definition, the propensity score $e_0(d \mid x)$ is given by $e_0(1 \mid x) = e_0(1)\zeta_{\mathrm{T},0}(x)/\zeta_0(x)$. The outcome is generated in the same manner as in the censoring setting: $Y = (X^\top \beta)^2 + 1.1 + \tau_0 D + \varepsilon$, where $\tau_0 = 3$.

Since we use neural networks, we estimate the propensity score g_0 using the non-negative PU learning method proposed by Kiryo et al. (2017), which is designed to mitigate overfitting when neural networks are applied. For simplicity, we assume that the class prior $e_0(1)$ is known.

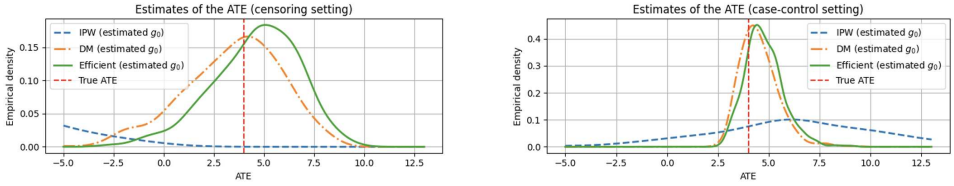

Figure 5: Response surface A. Left: censoring setting; Right: case-control setting.

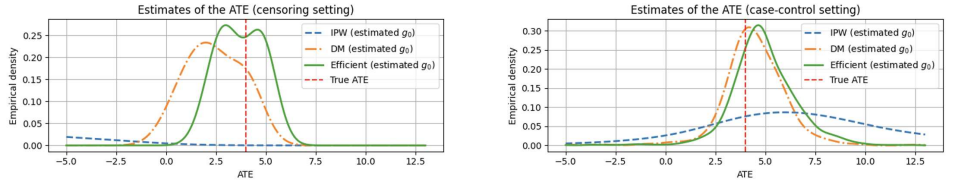

Figure 6: Response surface B. Left: censoring setting; Right: case-control setting.

We consider two cases: $(m, l) = (1000, 2000)$ and $(2000, 3000)$. We conduct 5000 trials and report the empirical mean squared errors (MSEs) and biases for the true ATE, along with the coverage ratio computed from the confidence intervals in Table 2 for $(m, l) = (1000, 2000)$ and Table 3 for $(m, l) = (2000, 3000)$. We also present the empirical distributions of the ATE estimates in Figure 3 for $(m, l) = (1000, 2000)$ and Figure 4 for $(m, l) = (2000, 3000)$.

M Empirical analysis using semi-synthetic data

In this section, we investigate the empirical performance of our estimators using the Infant Health and Development Program (IHDP) dataset. The dataset contains simulated outcomes paired with covariates observed in the real world (Hill, 2011).

M.1 Dataset.

The sample size is 747, and the covariates consist of 6 continuous variables and 19 binary variables.

Hill (2011) considers two scenarios for the outcome models: response surface A and response surface B. Response surface A generates the potential outcomes $Y_t(1)$ and $Y_t(0)$ according to the following model:

$$Y_t(0) \sim \mathcal{N}(X_t^\top \boldsymbol{\gamma}_A, 1),$$
$$Y_t(1) \sim \mathcal{N}(X_t^\top \boldsymbol{\gamma}_A + 4, 1),$$

where each element of $\boldsymbol{\gamma}_A \in \mathbb{R}^{25}$ is randomly drawn from $\{0, 1, 2, 3, 4\}$ with probabilities $(0.5, 0.2, 0.15, 0.1, 0.05)$.

In contrast, response surface B generates the potential outcomes $Y_t(1)$ and $Y_t(0)$ as follows:

$$Y_t(0) \sim \mathcal{N}\left(\exp\left((X_t + W)^\top \boldsymbol{\gamma}_B\right), 1\right),$$
$$Y_t(1) \sim \mathcal{N}(X_t^\top \boldsymbol{\gamma}_B - q, 1),$$

where W is an offset matrix of the same dimension as X_t with all elements equal to 0.5, q is a constant chosen to normalize the average treatment effect conditional on $d = 1$ to be 4, and each element of $\boldsymbol{\gamma}_B \in \mathbb{R}^{25}$ is randomly drawn from $\{0, 0.1, 0.2, 0.3, 0.4\}$ with probabilities $(0.6, 0.1, 0.1, 0.1, 0.1)$.

M.2 Censoring setting

We first investigate the censoring setting. The other experimental setups are identical to those in Section 6. Given $\{(X_i, D_i, Y_i)\}$ from the IHDP dataset, we generate the observation indicator O

from a Bernoulli distribution with probability 0.1 if $D_i = 1$, and set $O_i = 0$ if $D_i = 0$. The nuisance parameters are estimated using linear regression and (linear) logistic regression.

We compare our proposed estimator, $\widehat{\tau}_n^{\text{cens-eff}}$, with two other candidates: the IPW estimator $\widehat{\tau}_n^{\text{cens-IPW}}$ and the DM estimator $\widehat{\tau}_n^{\text{cens-DM}}$, as defined in Remarks 4.4 and 4.4, respectively. All of these estimators are proposed by us. Our aim is not to demonstrate that $\widehat{\tau}_n^{\text{cens-eff}}$ strictly outperforms the others, although we recommend it in practice. Unlike in Section 6, we only consider the case in which the propensity score is estimated using the method proposed by Elkan & Noto (2008).

For each outcome model (response surface A and B), we conduct 1000 trials and report the empirical mean squared errors (MSEs), biases for the true ATE, and the coverage ratio (Cov. ratio) computed from the confidence intervals in Tables 4 and 5. We also present the empirical distributions of the ATE estimates in Figures 5 and 6.

M.3 Case-control setting

In the case-control setting, we randomly split the dataset $\mathcal{D}$ into two subsets. One is used as an unlabeled dataset, and the other is used as a positive dataset by selecting only the treated units from it. The class prior is set as $e_0(1) = 0.1$.

For each outcome model (response surface A and B), we conduct 1000 trials and report the empirical mean squared errors (MSEs), biases for the true ATE, and the coverage ratio (Cov. ratio) computed from the confidence intervals in Tables 4 and 5. We also present the empirical distributions of the ATE estimates in Figures 5 and 6.