Semi-Supervised Treatment Effect Estimation with Unlabeled Covariates for Prediction-Powered Causal Inference

Masahiro Kato *mkato-csecon@g.ecc.u-tokyo.ac.jp*

The University of Tokyo
Osaka Metropolitan University

Abstract

This study investigates treatment effect estimation in the semi-supervised setting, also can be interpreted as prediction-powered inference. In our setting, we can use not only the standard triple of covariates, treatment indicator, and outcome, but also unlabeled auxiliary covariates. For this problem, we develop efficiency bounds and efficient estimators whose asymptotic variance aligns with the efficiency bound. In the analysis, we introduce two different data-generating processes: the one-sample setting and the two-sample setting. The one-sample setting considers the case where we can observe treatment indicators and outcomes for a part of the dataset, which is also called the censoring setting. In contrast, the two-sample setting considers two independent datasets with labeled and unlabeled data, which is also called the case-control setting or the stratified setting. In both settings, we find that by incorporating auxiliary covariates, we can lower the efficiency bound and obtain an estimator with an asymptotic variance smaller than that without such auxiliary covariates. We frame our framework as prediction-powered causal inference.

Keywords: causal inference; prediction-powered inference; double machine learning; Riesz regression; semiparametric efficiency; semi-supervised learning

1 Introduction

A core interest in causal inference is estimating treatment effects, including the average treatment effect (ATE, Imbens & Rubin, 2015). In the standard setup, we estimate such treatment effects from triples of covariates, a treatment indicator, and outcomes. As in other statistical analyses, accuracy depends not only on the statistical method but also on the amount and type of data available. While randomized controlled trials are the gold standard, they are often infeasible. Therefore, in many practical scenarios, we use observational data to perform causal inference. However, observational data are also not necessarily easy to collect. In particular, treatment variables and the corresponding outcomes are often costly, whereas covariates are usually easy to gather.

Under this practical scenario, we consider estimating ATEs more accurately using auxiliary unlabeled covariates, even when treatment variables and outcomes are missing. We also discuss the average treatment effect on the treated, because it is often the causal target in observational studies where treatment participation itself determines the relevant population. This setting corresponds to semi-supervised learning in machine learning, where we utilize both labeled and unlabeled data (Chapelle et al., 2006), which is also referred to as prediction-powered inference (Angelopoulos & Bates, 2021; Demirel et al., 2024).

In many applications, such unlabeled covariates are easy to gather. For example, in the United States, we may aim to estimate the ATE for the effect of a new scholarship. Although we may know the covariates for an enormous number of students, we can assign treatment, scholarship, to only a limited number of them. In such cases, the unlabeled covariates contain information about the population over which the treatment effect is averaged, even though they do not contain treatment indicators or outcomes.

We find that, under appropriate conditions, using unlabeled data allows us to construct an ATE estimator whose asymptotic variance, or equivalently, asymptotic MSE, is smaller than that of an estimator that ignores unlabeled data, as shown by Hahn (1998). To support this finding, we develop an asymptotic efficiency bound, a lower bound on the asymptotic variance, when using unlabeled covariates, propose ATE estimators, and show that the resulting asymptotic variances match the efficiency bound. In the methodological and theoretical arguments, we consider two practical scenarios, called the one-sample and two-sample scenarios. In the one-sample scenario, we interpret the unlabeled covariates as part of a dataset with missing variables, outcomes and the treatment indicator. In the two-sample scenario, we assume that labeled and unlabeled data are two independent datasets. The distinction is important because the two scenarios lead to different tangent spaces, different efficient influence functions, and different ways of using the unlabeled covariates.

Our efficient ATE estimators are developed based on the efficient influence function implied by the efficiency bound. We then extend the same logic to ATT, where the target distribution depends on the propensity score and therefore requires an additional treatment-law correction. This object is also called a Neyman orthogonal score in the debiased machine learning literature (Chernozhukov et al., 2018). The Neyman orthogonal scores include nuisance parameters, regression functions and a Riesz representer, which must be estimated before obtaining the ATE estimators. For the Riesz representer estimation, we employ generalized Riesz regression in Kato (2025b;a), which generalizes the Riesz regression in Chernozhukov et al. (2021). The unified treatment in Kato (2026a) further clarifies how Bregman divergence, loss-link choices, automatic regressor balancing, and automatic Neyman orthogonalization are connected. We extend this generalized Riesz regression perspective to the semi-supervised setting.

We list our contributions as follows:

- We develop efficiency bounds for regular ATE estimators in the one-sample and two-sample scenarios, which also yield the corresponding Neyman orthogonal scores. In the two-sample scenario, the efficient influence functions are centered within each stratum.
- We construct asymptotically efficient estimators using the Neyman orthogonal scores and show that their asymptotic variances match the efficiency bounds. The two-sample target mixture parameter is treated as part of the estimand, not as a data-adaptive tuning parameter.
- We extend generalized Riesz regression for estimating nuisance parameters, including the Riesz representer and regression functions, while explicitly incorporating unlabeled covariates into the Riesz representer objective.
- We derive an explicit efficiency-gain identity showing that, in the same-population case, unlabeled covariates reduce the covariate averaging component but not the residual outcome-noise component.
- We extend the efficient-score construction to ATT. The ATT estimator is a debiased ratio estimator, and the corresponding score contains an additional treatment-law correction because the treated covariate distribution is unknown.

Scope of the contribution. The role of generalized Riesz regression in this paper is different from its role in the general framework of Kato (2026a). That framework studies how to fit Riesz representers under Bregman divergences and how loss-link choices induce automatic regressor balancing and automatic Neyman orthogonalization. In contrast, the present paper fixes a semi-supervised observation scheme and derives the efficient scores, efficiency bounds, and attainable estimators under that scheme. The new content is therefore the interaction between unlabeled covariates, stratum-specific efficiency theory, and ATE or ATT targets. The generalized Riesz regression objectives are used as implementable nuisance-learning devices for the representers implied by these efficient scores.

Related work. The related topics of this study include debiased machine learning, efficiency under the two-sample case (stratified sampling scheme), treatment effect estimation with missing values, density-ratio estimation, and semi-supervised learning.

In treatment effect estimation, we typically aim to attain the $\sqrt{n}$-rate with the smallest asymptotic variance, or equivalently, asymptotic MSE. We provide an efficiency bound, which is a lower bound on the asymptotic variance among regular estimators. As discussed in Uehara et al. (2020), when there are two independent datasets, we cannot apply the usual efficiency bounds developed for a single dataset. To derive efficiency bounds in such settings, existing studies employ the efficiency theory under the stratified sampling scheme (Wooldridge, 2001). Using this scheme, efficiency bounds have been proposed for various settings, including multiple log data, active learning, learning from positive and unlabeled data, and external-validity problems. This study also employs this technique to develop efficiency bounds.

The efficiency bounds are derived from the efficient influence functions. Certain efficient influence functions take forms that allow the removal of bias caused by the estimation errors of the nuisance parameters. Debiased machine learning is a framework for estimating treatment effects by utilizing such properties (Chernozhukov et al., 2018). We refer to efficient influence functions with these properties as Neyman orthogonal scores. Chernozhukov et al. (2022b) reframes this framework by characterizing Neyman orthogonal scores using the Riesz representer. Chernozhukov et al. (2021) proposes Riesz regression, an end-to-end method for estimating the Riesz representer. Kato (2025a) and Kato (2025b) propose generalized Riesz regression by regarding the Riesz representer estimation problem as Bregman divergence minimization (Sugiyama et al., 2011). Kato (2026a) provides a broader formulation that relates Bregman-Riesz fitting to automatic regressor balancing and automatic Neyman orthogonalization. The present paper uses this machinery for a different purpose. We keep the sampling scheme explicit and derive the efficiency bounds induced by semi-supervised covariates, rather than starting from a fixed single-sample Riesz functional. This is why the two-sample result requires separate labeled and unlabeled influence functions, and why the ATT result requires a debiased denominator.

This study generalizes treatment effect estimation under covariate shift (Uehara et al., 2020; Kato et al., 2024) and in the positive-unlabeled (PU) learning setup (Kato et al., 2025). PU learning is a classical problem, originally studied in Imbens & Lancaster (1996), and recently reframed by du Plessis et al. (2015) as a modern statistical machine learning framework. Our sampling scheme arguments are significantly inspired by the works in this literature.

This study is also related to semi-supervised regression (Azriel et al., 2022; Kawakita & Kanamori, 2013) and treatment effect estimation with missing values (Heckman, 1974; Robins et al., 1994; Kennedy, 2020). These studies clarify how auxiliary covariates or missingness mechanisms can improve estimation. Our focus differs because the target is the semiparametric efficiency bound for causal effects when the auxiliary observations contain only covariates.

A further distinction is that the unlabeled observations in this paper affect the target covariate distribution rather than providing additional outcomes or surrogate outcomes. This distinction is important for efficiency. The unlabeled sample can improve the estimation of the covariate averaging component, but it cannot directly reduce the conditional outcome-noise component.

2 Problem Setting

In this section, we formulate our problem setting. We define potential outcomes and observations separately by following the Neyman–Rubin causal model (Neyman, 1923; Rubin, 1974). Then, we define the evaluation covariate density and the two sampling scenarios. .

2.1 Potential Outcomes

There is a binary treatment $d \in \{1, 0\}$. Let us define the corresponding potential outcome by $Y(d)$. Let $X \in \mathcal{X} \subset \mathbb{R}^k$ be a k-dimensional covariate, where $\mathcal{X}$ is the space. For each $d \in \{1, 0\}$, assume that the conditional distribution of $Y(d)$ given X has its density, and let $r_{Y(d),0}(y(d) \mid X)$ be the probability density function.

2.2 Average Treatment Effect

This study focuses on the estimation of the average treatment effect, which is the expected value of $Y(1)-Y(0)$. We take the expectation over a distribution whose covariate probability density is given by

$$\kappa_0(x).$$

We call it the evaluation covariate density. This density function can differ from $p_0(x)$. We make the assumptions for $\kappa_0(x)$ in the following sections.

Under a given covariate density $\kappa_0(x)$, the ATE is defined as follows:

$$\tau_0 \coloneqq \mathbb{E}_{\kappa_0}\big[Y(1)-Y(0)\big] \coloneqq \int y r_{Y(1),0}(y \mid x)\kappa_0(x)\mathrm{d}y\mathrm{d}x - \int y r_{Y(0),0}(y \mid x)\kappa_0(x)\mathrm{d}y\mathrm{d}x,$$

where $\mathbb{E}_{\kappa_0}[\cdot]$ denotes the expectation taken over the distribution whose covariate density is $\kappa_0(x)$.

2.3 Average Treatment Effect on the Treated

In addition to ATE, we consider the average treatment effect on the treated. For a given evaluation covariate density κ_0, define

$$\rho_{0,\kappa} \coloneqq \mathbb{E}_{\kappa_0}\left[e_0(1 \mid X)\right],$$

where $e_0(1 \mid X) = P(D=1 \mid X)$. The ATT target under the evaluation density κ_0 is

$$\tau_{0,\kappa}^{\mathrm{ATT}} \coloneqq \frac{\mathbb{E}_{\kappa_0}\left[e_0(1 \mid X)\tau_0(X)\right]}{\rho_{0,\kappa}}.$$

When $\kappa_0 = p_0$, this target is the usual ATT in the one-sample population. When $\kappa_0 = \kappa_{0,\beta}$ in the two-sample scenario, it is the ATT for the evaluation population determined by the mixture density. This parameter is not obtained by replacing the ATE density with the treated covariate density alone, because the treated covariate density itself depends on the unknown propensity score. This feature is the source of the treatment-law correction in Section B.

2.4 Observation

This section defines the sample, that is, observations of X, D, and Y. To define observations rigorously, we need to consider the censoring setting carefully. To discuss data augmentation within the theory of semiparametric efficiency, we introduce two DGPs. The first DGP is the one-sample scenario, where there is only one dataset, and in this dataset, treatment indicators and outcomes are observed only for a subset of units. We also refer to this setting as the censoring setting. The second DGP is the two-sample scenario, where there are two independent datasets; one of the datasets contains data with covariates, treatment indicator, and outcomes, while the other only contains covariates. We also refer to this setting as the case-control setting or the stratified sampling scheme. We define these two DGPs below.

One-sample scenario. In the one-sample scenario, we observe a single dataset $\mathcal{D}$, defined as follows:

$$\mathcal{D} \coloneqq \left\{\left(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i\right)\right\}_{i=1}^{n} \text{ with } \left(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i\right) \overset{\text{i.i.d}}{\sim} p_0(x, o, \widetilde{d}, \widetilde{y}).$$

where $O_i \in \{1,0\}$ is an observation indicator, $\widetilde{D}_i \in \{1,0,\mathrm{NA}\}$, and $\widetilde{Y}_i$ is the observable treatment indicator and outcome, defined as

$$\widetilde{D}_i \coloneqq \mathbb{1}[O_i=1]D_i + \mathbb{1}[O_i=0]\mathrm{NA},$$
$$\widetilde{Y}_i \coloneqq \mathbb{1}[O_i=1]Y_i + \mathbb{1}[O_i=0]\mathrm{NA},$$

$D_i \in \{1,0\}$ is a treatment indicator, and Y_i is the outcome defined as

$$Y_i \coloneqq \mathbb{1}[D_i=1]Y_i(1) + \mathbb{1}[D_i=0]Y_i(0).$$

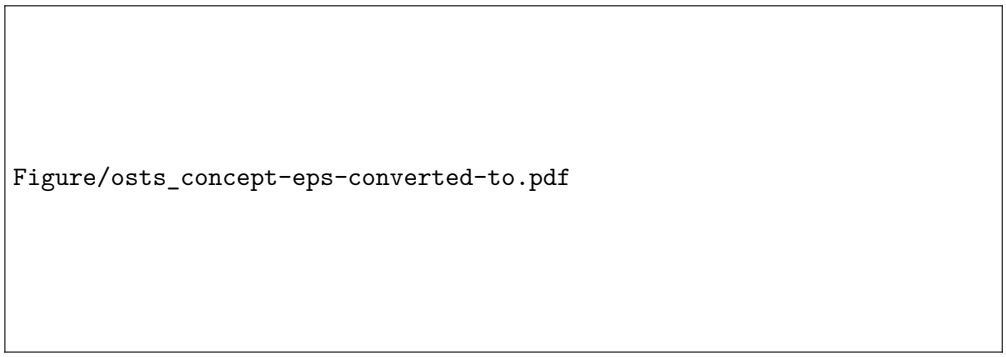

Figure 1: Illustration of the one-sample and two-sample scenarios.

Here, NA denotes a missing value. Equivalently, we can write $\widetilde{Y}_i$ as

$$\widetilde{Y}_i = \mathbb{1}\big[O_i = 1, \widetilde{D}_i = 1\big]Y_i(1) + \mathbb{1}\big[O_i = 1, \widetilde{D}_i = 0\big]Y_i(0) + \mathbb{1}\big[O_i = 0\big]\mathrm{NA}.$$

In this setting, we assume $p_0(x) = \kappa_0(x)$.

Note that $\widetilde{D}$ and $\widetilde{Y}$ are observable, while Y_i and D_i are not observable when $O_i = 0$.

Two-sample scenario In the two-sample scenario, we observe two stratified datasets, $\mathcal{D}_{\mathrm{L}}$ and $\mathcal{D}_{\mathrm{U}}$:

$$\mathcal{D}_{\mathrm{L}} \coloneqq \big\{\big(X_j, D_j, Y_j\big)\big\}_{j=1}^{m} \text{ with } \big(X_j, D_j, Y_j\big) \overset{\text{i.i.d}}{\sim} p_0(x, d, y) \text{ and}$$
$$\mathcal{D}_{\mathrm{U}} \coloneqq \{Z_k\}_{k=1}^{l} \text{ with } \big(Z_k\big) \overset{\text{i.i.d}}{\sim} q_0(x),$$

where m and l are the sample sizes of each dataset, and Y_j is the observed outcome defined as

$$Y_j = \mathbb{1}[D_j = 1]Y_j(1) + \mathbb{1}[D_j = 0]Y_j(0),$$

and $D_j \in \{1, 0\}$ is a treatment indicator.

Difference between the two settings We show an illustration that demonstrates the difference between the one-sample and two-sample scenarios in Figure 1. In both settings, we can identify and estimate the ATE in the standard way if we ignore the unlabeled auxiliary covariates. That is, in the one-sample scenario, we can estimate the ATE only by using $\mathcal{D}$, while in the two-sample scenario, we can estimate the ATE only by using $\mathcal{D}_{\mathrm{L}}$. However, the unlabeled covariates change the information available about the evaluation covariate distribution. We demonstrate that this additional information reduces the asymptotic variance of efficient estimators.

A summary of the differences is provided below:

One-sample scenario: A single dataset is observed, where some observations do not include the treatment and outcome variables (i.e., contain only unlabeled covariates).

Two-sample scenario: Two separate datasets are observed: one consists of fully labeled data, and the other contains only unlabeled covariates.

Remark (PU learning)**.** *Our terminology of the censoring and case-control settings comes from that in PU learning (Niu et al., 2016). In both settings, the goal is to learn a conditional class probability or a classifier using only positive and unlabeled data. In censoring PU learning, we consider one dataset from which labeled data are observed (Elkan & Noto, 2008). In case-control PU learning, we assume that there exist two independent datasets, one labeled and the other unlabeled (du Plessis et al., 2015). The case-control PU learning setting is also studied in Imbens & Lancaster (1996).*

Notations and Assumptions

Throughout this study, let $P(R)$ denote the distribution of a random variable R. For simplicity, we assume that the distribution $P(R)$ of a continuous random variable R has a probability density, whose notation depends on the random variable. For a probability density or mass function p, we denote the expectation over p by $\mathbb{E}_p[\cdot]$. If the dependence is clear from the context, we omit p and simply denote it as $\mathbb{E}[\cdot]$. Similarly, let $\mathrm{Var}(\cdot)$ be the variance operator. Let us denote the true mean and variance of $Y(d)$ conditioned on $X = x \in \mathcal{X}$ by $\mu_0(d,x) = \mathbb{E}[Y(d) \mid X = x]$ and $\sigma_0^2(d,x) = \mathrm{Var}(Y(d) \mid X = x)$, respectively.

We make the following regularity assumption.

Assumption 2.1. *There exist constants $\underline{C}$ and $\overline{C}$ such that $0 < \underline{C} < \overline{C} < \infty$ and for any $x \in \mathcal{X}$, $\left|\mu_0(d,x)\right| < \overline{C}$ and $\underline{C} < \sigma_0^2(d,x) < \overline{C}$ hold.*

3 One-Sample Scenario

First, we consider the one-sample setting for the DGP, which is also referred to as the censoring setting. We redefine the DGP with its notations and assumptions in Section 3.1. Then, for this DGP, we develop an efficiency bound in Section 3.2. We propose our estimator in Section 3.3 and show consistency in Section 3.5 and asymptotic normality in Section 3.6.

3.1 Notation and Assumption

This section introduces and summarizes the notations and assumptions, while recapping the DGP of the one-sample scenario. As defined in Section 2, the DGP of this scenario is

$$\mathcal{D} \coloneqq \left\{\left(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i\right)\right\}_{i=1}^n \text{ with } \left(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i\right) \in \mathcal{X} \times \{1,0\} \times \{1,0,\mathrm{NA}\} \times \{\mathcal{Y} \cup \mathrm{NA}\} \overset{\text{i.i.d}}{\sim} p_0(x, o, \widetilde{d}, \widetilde{y}).$$

Let $\pi_0(o \mid X) = p(O = o \mid X)$ be the probability of observation $O = o$, $e_0(d \mid X) = P(D = d \mid X, O = 1)$ be the propensity score defined in the observed samples, and let $g_0(a \mid X) = P(O = 1, D = a \mid X) = e_0(a \mid X)\pi_0(1 \mid X)$ be the joint probability of $O = 1$ and $D = d$. Under this notation, the probability density $p_0(x, o, \widetilde{d}, \widetilde{y})$ is written as

$$\begin{aligned} &p_0(x, o, \widetilde{d}, \widetilde{y}) = \\ &p_0(x)\Big(\pi_0(0 \mid X)\Big)^{\mathbb{1}[o=0]}\Big(g_0(1 \mid x) r_{Y(1),0}(\widetilde{y} \mid x)\Big)^{\mathbb{1}\left[o=1,\widetilde{d}=1\right]}\Big(g_0(0 \mid x) r_{Y(0),0}(\widetilde{y} \mid x)\Big)^{\mathbb{1}\left[o=1,\widetilde{d}=0\right]}, \end{aligned}$$

For simplicity, we assume that the evaluation density $\kappa_0(x)$ is the marginal density of the covariates.

Assumption 3.1 (Evaluation density in the one-sample scenario)**.** *The evaluation density $\kappa_0(x)$ is given as $\kappa_0(x) = p_0(x)$.*

We also make the following assumptions.

Assumption 3.2 (Unconfoundedness and missing at random (MAR))**.** *It holds that $(Y(1), Y(0)) \perp\!\!\!\perp D \mid X$ (unconfoundedness) and $(Y(1), Y(0)) \perp\!\!\!\perp O \mid X$ (MAR).*

Assumption 3.3 (Common support)**.** *There exists a universal constant $0 < \epsilon < 1/2$ such that for all $d \in \{1,0\}$, $\epsilon < g_0(d \mid X) \leq 1 - \epsilon$ holds almost surely.*

Note that this assumption also implies the existence of a universal constant $0 < \epsilon' < 1/2$ such that $\epsilon' < \pi_0(1 \mid X)$.

3.2 Efficiency Bound

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). In this scenario, the influence function has the usual augmented inverse probability structure, with $g_0(d \mid X)$

replacing the propensity score because both treatment assignment and label observation must occur. The result is stated below, and the proof is provided in Appendix C.

Lemma 3.1. *Suppose that Assumptions 3.1 through 3.3 hold. Then, the efficient influence function is given as*

$$\psi^{OS}\big(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \mu_0, g_0, \tau_0\big),$$

where

$$\begin{aligned}\psi^{OS}\big(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \mu_0, g_0, \tau_0\big) &:= S^{OS}\big(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \mu_0, g_0\big) - \tau_0 \\ S^{OS}\big(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \mu_0, g_0\big) &:= \frac{\mathbb{1}\big[O_i = 1, \widetilde{D}_i = 1\big]\big(\widetilde{Y}_i - \mu_0(1, X_i)\big)}{g_0(1 \mid X)} - \frac{\mathbb{1}\big[O_i = 1, \widetilde{D}_i = 0\big]\big(\widetilde{Y}_i - \mu_0(0, X_i)\big)}{g_0(0 \mid X)} \\ &\quad + \mu_0(1, X_i) - \mu_0(0, X_i).\end{aligned}$$

Recall that $g_0(d \mid X) = \pi_0(1 \mid X)e_0(d \mid X)$. The following proposition is Theorem 25.20 in van der Vaart (1998), which connects the efficient influence function to the efficiency bound.

Proposition 3.2 (Theorem 25.20 in van der Vaart (1998).)**.** *Let R be some random variable, and $\mathcal{M}$ be a model of its DGP. The (semiparametric) efficient influence function $\psi(R)$ is the gradient of θ with respect to the model $\mathcal{M}$, which has the smallest L_2-norm. It satisfies that for any regular estimator $\widehat{\theta}$ of a parameter of interest θ_0 regarding a given parametric submodel, $AMSE\big(\widehat{\theta}\big) \geq Var\big(\psi(R)\big)$, where $AMSE\big(\widehat{\theta}\big)$ is the second moment of the limiting distribution of $\sqrt{n}\left(\widehat{\theta} - \theta_0\right)$.*

Using this proposition, we can derive the following efficiency bound from Lemma 3.1.

Theorem 3.3 (Efficiency bound in the one-sample scenario)**.** *If Assumptions 3.1 through 3.3 hold, then the asymptotic variance of any regular estimator is lower bounded by*

$$V^{OS} := \mathbb{E}\left[\psi^{OS}\left(X, O, \widetilde{D}, \widetilde{Y}; \mu_0, g_0, \tau_0\right)^2\right] = \mathbb{E}\left[\frac{\sigma_0^2(1, X)}{g_0(1 \mid X)} + \frac{\sigma_0^2(0, X)}{g_0(0 \mid X)} + \Big(\tau_0(X) - \tau_0\Big)^2\right],$$

where $\tau_0(X) := \mathbb{E}\big[Y(1) - Y(0) \mid X\big]$ is the conditional ATE.

Here, note that the efficient influence function depends on the unknown μ_0, g_0, which are referred to as nuisance parameters. Since the efficient influence function satisfies the equation $\mathbb{E}\left[\psi^{\mathrm{OS}}(X, O, Y; \mu_0, g_0, \tau_0)\right] = 0$, if the nuisance parameters are known and the exact expectation is computed, we can obtain τ_0 by solving for τ_0 that satisfies this equation. Thus, the efficient influence function provides a direct estimating equation for an efficient estimator. Furthermore, the accuracy of the estimation of the nuisance parameters affects the estimation of τ_0, the parameter of interest.

3.3 ATE Estimator

Based on the efficient influence function, we propose an ATE estimator defined as

$$\widehat{\tau}_n^{\text{OS-eff}} := \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{OS}}\left(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \widehat{\mu}_{n,i}, \widehat{g}_{n,i}\right),$$

where $\widehat{\mu}_{n,i}$ and $\widehat{g}_{n,i}$ are estimators of μ_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). We say that an estimator is efficient if its asymptotic variance aligns with V^{OS}.

For estimating the regression function μ_0, we can employ methods for conditional ATE estimation (Wager & Athey, 2018; Curth & van der Schaar, 2021; Kennedy et al., 2024), as well as standard regression methods using parametric or nonparametric models (Tsybakov, 2008; Schmidt-Hieber, 2020). We can also use targeted maximum likelihood estimation to refine this estimation (van der Laan & Rose, 2011).

For estimating g_0, we can use logistic regression or other advanced methods, such as the covariate balancing propensity score (Imai & Strauss, 2011; Hainmueller, 2012) and Riesz regression (Chernozhukov et al., 2021).

As Zhao (2019), Bruns-Smith et al. (2025), and Kato (2025a) show, Riesz regression and covariate balancing methods are in a dual relationship, and Riesz regression can be interpreted as a special case of density ratio estimation (Kato, 2025d). For details, see Section 3.4 and Appendix A.

3.4 Generalized Riesz Regression

We explain how to construct estimators for g_0. In this study, we employ generalized Riesz regression, also referred to as Bregman-Riesz regression (Kato, 2025a;b; 2026a). In the efficient estimation of causal parameters, Neyman orthogonal scores play an important role and typically correspond to the efficient score. Specifically, asymptotically efficient estimators must be asymptotically linear with respect to the Neyman orthogonal scores. When the parameter of interest is linear in the regression functions, the Neyman orthogonal score can be decomposed into the Riesz representer and regression functions. In our semi-supervised setting, the Riesz representer is the component that determines how labeled residuals and unlabeled covariate averages are combined. In our framework, the Neyman orthogonal score is given by

$$\psi^{\mathrm{OS}}\left(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \mu_0, \alpha_0, \tau_0\right) \coloneqq \alpha_0\left(O_i, \widetilde{D}_i, X_i\right)\left(\widetilde{Y}_i - \mu_0\left(\widetilde{D}_i, X_i\right)\right) + \mu_0(1, X_i) - \mu_0(0, X_i) - \tau_0,$$

where we replace g_0 with α_0 in the original definition of ψ^{OS}, and $\alpha_0\left(O_i, \widetilde{D}_i, X_i\right) \coloneqq \frac{\mathbb{1}[O_i=1, \widetilde{D}_i=1]}{g_0(1|X_i)} - \frac{\mathbb{1}[O_i=1, \widetilde{D}_i=0]}{g_0(0|X_i)}$ is the Riesz representer. Riesz regression, as proposed by Chernozhukov et al. (2021), is a method for estimating the Riesz representer in an end-to-end manner. Kato (2025a) shows that Riesz regression is a specific instance of density ratio estimation and can be generalized via Bregman divergence minimization (Sugiyama et al., 2011). Kato (2025b) further reformulates and extends this approach as direct debiased machine learning (DDML) via generalized Riesz regression. The efficiency results below use high-level product-rate assumptions for the learned nuisance functions. Generalized Riesz regression is one way to construct the representer estimator under these assumptions, while detailed rates for RKHS and neural network classes can be imported from the general theory of Bregman-Riesz fitting.

Generalized Riesz regression. Generalized Riesz regression estimates α_0 by minimizing the Bregman divergence between the true Riesz representer α_0 and its model α. That is, the estimation error of α_0 is measured using the Bregman divergence. The recent unified framework of Kato (2026a) emphasizes that the choice of Bregman divergence and link function determines both the geometry of the fitted representer and the associated balancing interpretation. For a twice differentiable convex function f with bounded derivative, the population objective for Riesz representer estimation is written as

$$\mathrm{BD}_f(\alpha) \coloneqq \tag{1}$$

$$\mathbb{E}\left[\mathbb{1}[O=1]\partial f\left(\alpha\left(O, \widetilde{D}, X\right)\right)\alpha\left(O, \widetilde{D}, X\right) - f\left(\alpha\left(O, \widetilde{D}, X\right)\right) - \left(\partial f\left(\alpha\left(1, 1, X\right)\right) - \partial f\left(\alpha\left(1, 0, X\right)\right)\right)\right]. \tag{2}$$

The empirical counterpart $\widehat{\mathrm{BD}}_f(\alpha)$ replaces expectations with sample averages. Here, we used $\tau_0 = \mathbb{E}\left[\mathbb{E}\left[\widetilde{Y} \mid O=1, \widetilde{D}=1, X\right] - \mathbb{E}\left[\widetilde{Y} \mid O=1, \widetilde{D}=0, X\right]\right]$. Minimizing this objective over a hypothesis class $\mathcal{A}$ yields an estimator of α_0, that is,

$$\widehat{\alpha}^{\mathrm{GRR}} \coloneqq \underset{\alpha \in \mathcal{A}}{\arg\min}\, \widehat{\mathrm{BD}}_f(\alpha),$$

where GRR denotes generalized Riesz regression.

The use of generalized Riesz regression allows us to naturally incorporate unlabeled covariates into the estimation of the Riesz representer. This is because, in equation 1, we can approximate $\mathbb{E}\left[\left(\partial f\left(\alpha\left(1, 1, X\right)\right) - \partial f\left(\alpha\left(1, 0, X\right)\right)\right)\right]$ using unlabeled covariates, whereas $\mathbb{E}\left[\partial f\left(\alpha\left(O, \widetilde{D}, X\right)\right)\alpha\left(O, \widetilde{D}, X\right) - f\left(\alpha\left(O, \widetilde{D}, X\right)\right)\right]$ requires labeled data. That is,

$$\widehat{\mathrm{BD}}_f(\alpha) \coloneqq \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[O_i=1]\partial f\left(\alpha\left(O_i, \widetilde{D}_i, X_i\right)\right)\alpha\left(O_i, \widetilde{D}_i, X_i\right) - f\left(\alpha\left(O_i, \widetilde{D}_i, X_i\right)\right)$$

$$-\frac{1}{n}\sum_{i=1}^{n}\Big(\partial f\left(\alpha\left(1,1,X_i\right)\right)-\partial f\left(\alpha\left(1,0,X_i\right)\right)\Big),$$

where the second term can be evaluated using both labeled and unlabeled data. This is the point at which the semi-supervised structure enters the Riesz representer estimation problem. Note that unlabeled covariates can be utilized even when g_0 is estimated via maximum likelihood. However, the generalized Riesz regression approach is arguably more appropriate in an end-to-end formulation because it directly targets the representer that appears in the Neyman orthogonal score. Also see Kawakita & Kanamori (2013).

Let $\mathcal{A}$ denote the model class for α_0. If we set $f(\alpha)=(\alpha-1)^2$, then

$$\mathrm{BD}_{\mathrm{LSIF}}\big(\alpha\big)\coloneqq\mathbb{E}\left[-2\big(\alpha\left(1,1,X\right)-\alpha\left(1,0,X\right)\big)+\mathbb{1}[O=1]\alpha\left(O,\widetilde{D},X\right)^2\right].$$

This population objective corresponds to Riesz regression as in Chernozhukov et al. (2021). We refer to this objective as least squares Riesz (LS-Riesz) regression.

Now, redefine $\mathcal{A}$ as the set of α such that $\alpha(1,1,\cdot)>1$ and $\alpha(1,0,\cdot)<-1$, a condition that should hold under the common support assumption. For $f(\alpha)=(|\alpha|-1)\log(|\alpha|-1)+|\alpha|$ $(\alpha\in\mathcal{A})$, the corresponding Bregman divergence is

$$\begin{aligned}\mathrm{BD}_{\mathrm{UKL}}\big(\alpha\big)\coloneqq\mathbb{E}\Big[\mathbb{1}[O=1]\left(\log\left(\left|\alpha\left(O,\widetilde{D},X\right)\right|-1\right)+\left|\alpha\left(O,\widetilde{D},X\right)\right|\right)\\-\log\big(\alpha(1,1,X)-1\big)-\log\big(-\alpha(1,0,X)-1\big)\Big].\end{aligned}$$

We refer to this objective as Kullback-Leibler Riesz (KL-Riesz) regression, since the choice of f yields the KL divergence.

By replacing the expectations with the sample mean and minimizing the empirical objective for α, we can estimate α_0.

Interpretation. As Kato (2025b) discusses, LS-Riesz regression corresponds to the stable balancing weights proposed in Zubizarreta (2015), and KL-Riesz corresponds to the entropy balancing weights in Hainmueller (2012). These correspondences were originally shown in the covariate balancing literature, such as in Zhao (2019) and Bruns-Smith et al. (2025). They can be derived from duality relationships. Kato (2026a) further interprets these relationships as automatic regressor balancing under suitable loss-link pairs.

Note that the duality depends on the model class used for α_0, namely $\mathcal{A}$. For the duality between LS-Riesz and stable balancing weights, linear models must be used for $\mathcal{A}$, whereas for the duality between KL-Riesz and entropy balancing weights, logistic models for α_0 are required. Therefore, the choice of $\mathcal{A}$ is not a purely computational choice; it determines which balancing equations are targeted by the Riesz representer estimator.

3.5 Consistency and double robustness

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

Assumption 3.4. *There exist universal constants $\epsilon\in(0,1/2), C\in(0,\infty)$ such that $\widehat{g}_{n,i}(a\mid X)\in(\epsilon,1-\epsilon)$ and $\widehat{\mu}_{n,i}(d\mid X)\in[-C,C]$ hold almost surely. As $n\to\infty$, either of the following holds for all $i\in\{1,2,\ldots,n\}$:*

$$\left\|\widehat{\mu}_{n,i}-\mu_0\right\|_2=o_p(1)\ \text{ or }\ \left\|\widehat{g}_{n,i}-g_0\right\|_2=o_p(1).$$

Then, the following consistency result holds. This result is given as a special case of Theorem 3.5; therefore, we omit the proof.

Theorem 3.4 (Consistency in the one-sample setting)**.** *If Assumptions 3.1 through 3.3, and 3.4 hold, then $\widehat{\tau}_n^{OS\text{-}eff}\xrightarrow{\mathrm{p}}\tau_0$ holds as $n\to\infty$.*

This consistency structure is referred to as double robustness.

Algorithm 1 Cross-fitting in the one-sample scenario

Input: Observations $\mathcal{D} \coloneqq \left\{\left(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i\right)\right\}_{i=1}^{n}$, number of folds L, and estimation methods for μ_0 and g_0. Let $\mathcal{I} = \{1, 2, \ldots, n\}$ be the index set.
Randomly split $\mathcal{I}$ into L roughly equal-sized folds, $(\mathcal{I}^{(b)})_{b\in\mathcal{L}}$. Note that $\bigcup_{b\in\mathcal{L}} \mathcal{I}^{(b)} = \mathcal{I}$.
for $b \in \mathcal{L}$ **do**
  Set the training data as $\mathcal{I}^{(-b)} = \{1, 2, \ldots, n\} \setminus \mathcal{I}^{(b)}$.
  Construct estimators of the nuisance parameters on $\mathcal{I}^{(-b)}$, denoted by $\widehat{\mu}_n^{(b)}$ and $\widehat{g}_n^{(b)}$.
end for
Output: Obtain an ATE estimate $\widehat{\tau}_n^{\text{OS-eff}}$ using $\widehat{\mu}_n^{(b)}$ and $\widehat{g}_n^{(b)}$.

3.6 Asymptotic Normality

Next, we establish the asymptotic normality of our estimator. Unlike consistency, this requires stronger assumptions on the nuisance estimators, especially for the propensity score.

To prove asymptotic normality or $\sqrt{n}$-consistency, we must control the complexity of the nuisance parameter estimators. One simple approach is to assume the Donsker condition; however, it is well known that this condition often fails in high-dimensional regression settings. In such cases, asymptotic normality can still be attained using sample splitting, a common technique in this field (Klaassen, 1987), which has recently been refined by Chernozhukov et al. (2018) as cross-fitting.

Cross-fitting. We estimate μ_0 and g_0 using 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 $b \in \mathcal{L} \coloneqq \{1, 2, \ldots, L\}$, the nuisance parameters are estimated using all other folds. Let the estimators for fold $b \in \mathcal{L}$ be denoted by $\widehat{\mu}_n^{(b)}$ and $\widehat{g}_n^{(b)}$. Let $\mathcal{I}^{(b)}$ be the index set of samples belonging to fold b. This construction separates nuisance estimation from score evaluation and avoids imposing a Donsker condition on the nuisance classes.

Various estimation methods may be used, including neural networks and Lasso, as long as they satisfy the convergence rate conditions in Assumption 3.5. The pseudocode is shown in Algorithm 3.6.

Asymptotic normality. We present results for the case with cross-fitting, but similar results hold under the Donsker condition.

We make the following assumptions:

Assumption 3.5. *For each $b \in \mathcal{L}$, as $n \to \infty$, the following hold:*

- $\left\|\mu_0(a, X) - \widehat{\mu}_{n,i}^{(b)}(a, X)\right\|_2 = o_p(1)$ *and* $\left\|g_0(a \mid X) - \widehat{g}_{n,i}^{(b)}(a \mid X)\right\|_2 = o_p(1)$.
- $\left\|\mu_0(a, X) - \widehat{\mu}_n^{(b)}(a, X)\right\|_2 \cdot \left\|g_0(a \mid X) - \widehat{g}_{n,i}^{(b)}(a \mid X)\right\|_2 = o_p(n^{-1/2})$ *for* $a \in \{1, 0\}$.

We define the estimator as

$$\widehat{\tau}_n^{\text{OS-eff}} \coloneqq \frac{1}{n} \sum_{b\in\mathcal{L}} \sum_{i\in\mathcal{I}^{(b)}} S^{\text{OS}}\left(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right)$$

and show that the asymptotic normality holds as follows:

Theorem 3.5 (Asymptotic normality in the one-sample scenario)**.** *Consider the one-sample scenario. Suppose Assumptions 3.1 through 3.3 and 3.5 hold; that is, $\widehat{\mu}_{n,i} = \widehat{\mu}_n^{(b)}$ and $\widehat{g}_{n,i} = \widehat{g}_n^{(b)}$ are constructed via cross-fitting with suitable convergence rates. Then,*

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

The proof is provided in Appendix D. The asymptotic variance of $\widehat{\tau}_n^{\text{OS-eff}}$ matches the efficiency bound. Therefore, Theorem 3.5 also implies that $\widehat{\tau}_n^{\text{OS-eff}}$ is asymptotically efficient.

For inference, let $\widehat{\psi}_i^{\text{OS}}$ denote the influence function in Lemma 3.1 evaluated at the cross-fitted nuisance estimators and at $\widehat{\tau}_n^{\text{OS-eff}}$. We estimate the asymptotic variance by

$$\widehat{V}^{\text{OS}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\psi}_i^{\text{OS}} - \frac{1}{n}\sum_{i'=1}^{n}\widehat{\psi}_{i'}^{\text{OS}}\Big)^2.$$

The resulting Wald interval uses $\widehat{V}^{\text{OS}}/n$ as the variance of $\widehat{\tau}_n^{\text{OS-eff}}$.

We now discuss alternative ATE estimators.

Remark (Inefficiency of the Inverse Probability Weighting (IPW) estimator)**.** *The IPW estimator is defined as*

$$\widehat{\tau}_n^{OS\text{-}IPW} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}[O_i = 1, \widetilde{D}_i = 1]\widetilde{Y}_i}{\widehat{g}_{n,i}(1 \mid X_i)} - \frac{\mathbb{1}[O_i = 1, \widetilde{D}_i = 0]\widetilde{Y}_i}{\widehat{g}_{n,i}(0 \mid X_i)}\right).$$

Unlike our proposed efficient estimator, this estimator does not use the conditional outcome estimators (Horvitz & Thompson, 1952). When g_0 and π_0 are known, it is unbiased. However, it suffers from a large asymptotic variance:

$$V^{IPW} \coloneqq \mathbb{E}\left[\frac{\mathbb{E}[Y(1)^2 \mid X]}{g_0(1 \mid X)} + \frac{\mathbb{E}[Y(0)^2 \mid X]}{g_0(0 \mid X)}\right].$$

Here, $V^{IPW} \geq V^{OS}$, with equality when $\mu_0(d, x)$ is zero for all x. Thus, the IPW estimator is inefficient relative to $\widehat{\tau}_n^{OS\text{-}eff}$. Moreover, if π_0 is unknown, stronger assumptions are needed to establish asymptotic normality compared to our efficient estimator.

Remark (Regression Adjustment (RA) estimator)**.** *Another alternative is the RA estimator, defined as*

$$\widehat{\tau}_n^{OS\text{-}RA} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\widehat{\mu}_{n,i}(1, X_i) - \widehat{\mu}_{n,i}(0, X_i),$$

also known as the naive plug-in or direct method estimator. Its asymptotic normality heavily depends on the estimators $\widehat{\mu}_{n,i}$.

4 Two-Sample Scenario

Next, we consider the two-sample scenario for the DGP, which is also referred to as the case-control setting and stratified sampling scheme. We reintroduce the notation and assumptions required for our analysis in Section 4.1. Section 4.2 presents the efficiency bound, and Section 4.3 provides an ATE estimator under this setting. We establish consistency in Section 4.5 and asymptotic normality in Section 4.6. Finally, we compare the one-sample and two-sample scenarios in Section 4.7.

4.1 Notation and Assumptions

As introduced in Section 2, the DGP for the two-sample scenario is defined as

$$\mathcal{D}_{\text{L}} \coloneqq \big\{\big(X_j, D_j, Y_j\big)\big\}_{j=1}^{m}, \text{ with } \big(X_j, D_j, Y_j\big) \in \mathcal{X} \times \{1, 0\} \times \mathcal{Y}, \text{ and } \big(X_j, D_j, Y_j\big) \overset{\text{i.i.d}}{\sim} p_0(x, d, y),$$

$$\mathcal{D}_{\text{U}} \coloneqq \{Z_k\}_{k=1}^{l}, \text{ with } Z_k \in \mathcal{X}, \text{ and } Z_k \overset{\text{i.i.d}}{\sim} q_0(x).$$

Let $e_0(d \mid X) = P(D = d \mid X)$ denote the propensity score. Then, the joint density $p_0(x, d, y)$ can be written as

$$p_0(x, d, y) = p_0(x)\Big(e_0(1 \mid x)r_{Y(1),0}(y \mid x)\Big)^{\mathbb{1}[d=1]}\Big(e_0(0 \mid x)r_{Y(0),0}(y \mid x)\Big)^{\mathbb{1}[d=0]}.$$

For the evaluation density, we make the following assumption.

Assumption 4.1 (Evaluation density in the two-sample scenario)**.** *There exists a fixed* $\beta \in [0,1]$ *such that*

$$\kappa_0(x) = \kappa_{0,\beta}(x) = \beta p_0(x) + (1-\beta) q_0(x).$$

The value of β *is part of the estimand and is chosen by the researcher. It is not selected by minimizing an asymptotic variance unless the target density is invariant to* β*.*

The following support condition is needed because all outcome and treatment information is contained in the labeled sample.

Assumption 4.2 (Support)**.** *It holds that* $\kappa_{0,\beta} \ll p_0$*. In addition, there exists a universal constant* $C < \infty$ *such that* $\kappa_{0,\beta}(X)/p_0(X) \leq C$ *almost surely under* p_0*.*

If $\beta < 1$, Assumption 4.2 implies $q_0 \ll p_0$. Without this condition, there may be target covariate regions for which outcomes are never observed in the labeled sample.

We also impose the following assumptions.

Assumption 4.3 (Unconfoundedness)**.** *The potential outcomes satisfy* $(Y(1), Y(0)) \perp\!\!\!\perp D \mid X$*.*

Assumption 4.4 (Common support)**.** *There exists a universal constant* $0 < \epsilon < 1/2$ *such that* $\epsilon < e_0(1 \mid X) < 1 - \epsilon$ *almost surely under* p_0*.*

Define

$$\omega_{0,\beta}(x) \coloneqq \frac{\kappa_{0,\beta}(x)}{p_0(x)},$$

and

$$v_{0,\beta}(d,x) \coloneqq \frac{e_0(d \mid x)}{\omega_{0,\beta}(x)} = \frac{p_0(d,x)}{\kappa_{0,\beta}(x)}.$$

Thus, $1/v_{0,\beta}(d,x)$ is the efficient residual weight for the labeled sample. We also define

$$\tau_{p,0} \coloneqq \mathbb{E}_{p_0}\left[\tau_0(X)\right], \tau_{q,0} \coloneqq \mathbb{E}_{q_0}\left[\tau_0(Z)\right], \tau_0 = \beta\tau_{p,0} + (1-\beta)\tau_{q,0}.$$

4.2 Efficiency Bound

Following Uehara et al. (2020), we derive the efficiency bound using the efficiency arguments under the two-sample scenario. In this scheme, there are two efficient influence functions, one for the labeled stratum and one for the unlabeled stratum. The proof is provided in Appendix E.

Lemma 4.1. *If Assumptions 4.1 through 4.4 hold, then the efficient influence functions for the labeled and unlabeled strata are given by*

$$\begin{aligned}
\psi_L^{TS}(X,D,Y;\mu_0,v_{0,\beta}) &\coloneqq S^{TS}_{(X,D,Y)}(X,D,Y;\mu_0,v_{0,\beta}) + \beta\left(\tau_0(X) - \tau_{p,0}\right), \\
\psi_U^{TS}(Z;\mu_0) &\coloneqq (1-\beta)\left(\tau_0(Z) - \tau_{q,0}\right), \\
S^{TS}_{(X,D,Y)}(X,D,Y;\mu_0,v_{0,\beta}) &\coloneqq \frac{\mathbb{1}\big[D=1\big]\big(Y - \mu_0(1,X)\big)}{v_{0,\beta}(1,X)} - \frac{\mathbb{1}\big[D=0\big]\big(Y - \mu_0(0,X)\big)}{v_{0,\beta}(0,X)}.
\end{aligned}$$

The centering in Lemma 4.1 is stratum specific. The labeled stratum is centered by $\tau_{p,0}$ and the unlabeled stratum is centered by $\tau_{q,0}$. This is necessary because each stratum has its own sampling distribution.

As in the one-sample scenario, the efficient influence functions directly yield the following efficiency bound.

Theorem 4.2 (Efficiency bound in the two-sample scenario)**.** *Let* $N = m+l$*, where* $m = \alpha N$ *and* $l = (1-\alpha)N$ *for some* $\alpha \in (0,1)$*. If Assumptions 4.1 through 4.4 hold, then the asymptotic variance of any regular estimator is lower bounded by*

$$V^{TS}(\beta) \coloneqq \frac{1}{\alpha}\mathbb{E}_{p_0}\left[\psi_L^{TS}(X,D,Y;\mu_0,v_{0,\beta})^2\right] + \frac{1}{1-\alpha}\mathbb{E}_{q_0}\left[\psi_U^{TS}(Z;\mu_0)^2\right]$$

$$= \frac{1}{\alpha}\mathbb{E}_{p_0}\left[\left(\frac{\sigma_0^2(1,X)}{e_0(1\mid X)} + \frac{\sigma_0^2(0,X)}{e_0(0\mid X)}\right)\left(\frac{\kappa_{0,\beta}(X)}{p_0(X)}\right)^2\right]$$
$$+ \frac{\beta^2}{\alpha}\mathbb{E}_{p_0}\left[\left(\tau_0(X) - \tau_{p,0}\right)^2\right] + \frac{(1-\beta)^2}{1-\alpha}\mathbb{E}_{q_0}\left[\left(\tau_0(Z) - \tau_{q,0}\right)^2\right].$$

The parameter β is fixed throughout the theorem. If $p_0 = q_0$, all values of β define the same evaluation density. Only in such cases can β be selected to improve precision without changing the estimand.

4.3 ATE Estimator

Based on the efficient influence functions, we define the estimator as

$$\widehat{\tau}_n^{\text{TS-eff}} \coloneqq \frac{1}{m}\sum_{j=1}^{m} S^{\text{TS}}_{(X,D,Y)}\big(X_j, D_j, Y_j; \widehat{\mu}, \widehat{v}_\beta\big) + \beta\frac{1}{m}\sum_{j=1}^{m} S^{\text{TS}}_{(X)}\big(X_j; \widehat{\mu}\big) + (1-\beta)\,\frac{1}{l}\sum_{k=1}^{l} S^{\text{TS}}_{(X)}\big(Z_k; \widehat{\mu}\big),$$

where $S^{\text{TS}}_{(X)}(x;\mu) \coloneqq \mu(1,x) - \mu(0,x)$. Here, $\widehat{\mu}$ and $\widehat{v}_\beta$ denote estimators of μ_0 and $v_{0,\beta}$. Unlike in the one-sample scenario, we do not use the observation indicator O, since it is deterministically known whether a unit belongs to the labeled or unlabeled stratum. This distinction leads to theoretical differences from the one-sample scenario.

4.4 Generalized Riesz Regression in the Two-Sample Scenario

The two-sample efficient score also yields a semi-supervised generalized Riesz regression objective. The Riesz representer for the labeled residual part is

$$\alpha_{0,\beta}(D,X) \coloneqq \frac{\mathbb{1}\big[D=1\big]}{v_{0,\beta}(1,X)} - \frac{\mathbb{1}\big[D=0\big]}{v_{0,\beta}(0,X)}.$$

It represents the linear functional $h \mapsto \mathbb{E}_{\kappa_{0,\beta}}\left[h(1,X) - h(0,X)\right]$ through expectations under the labeled distribution. For a twice differentiable convex function f, define

$$\begin{aligned}\mathrm{BD}^{\text{TS}}_{f,\beta}(\alpha) \coloneqq\ & \mathbb{E}_{p_0}\left[\partial f\left(\alpha(D,X)\right)\alpha(D,X) - f\left(\alpha(D,X)\right)\right] \\ & - \beta\mathbb{E}_{p_0}\left[\partial f\left(\alpha(1,X)\right) - \partial f\left(\alpha(0,X)\right)\right] \\ & - (1-\beta)\mathbb{E}_{q_0}\left[\partial f\left(\alpha(1,Z)\right) - \partial f\left(\alpha(0,Z)\right)\right].\end{aligned}$$

The empirical counterpart is

$$\begin{aligned}\widehat{\mathrm{BD}}^{\text{TS}}_{f,\beta}(\alpha) \coloneqq\ & \frac{1}{m}\sum_{j=1}^{m}\Big(\partial f\left(\alpha(D_j,X_j)\right)\alpha(D_j,X_j) - f\left(\alpha(D_j,X_j)\right)\Big) \\ & - \beta\frac{1}{m}\sum_{j=1}^{m}\Big(\partial f\left(\alpha(1,X_j)\right) - \partial f\left(\alpha(0,X_j)\right)\Big) \\ & - (1-\beta)\frac{1}{l}\sum_{k=1}^{l}\Big(\partial f\left(\alpha(1,Z_k)\right) - \partial f\left(\alpha(0,Z_k)\right)\Big).\end{aligned}$$

Then, we estimate $\alpha_{0,\beta}$ by

$$\widehat{\alpha}^{\text{TS-GRR}} \in \operatorname*{arg\,min}_{\alpha\in\mathcal{A}}\left\{\widehat{\mathrm{BD}}^{\text{TS}}_{f,\beta}(\alpha) + \lambda J(\alpha)\right\}.$$

This objective makes explicit how the unlabeled covariates enter generalized Riesz regression: the labeled sample controls the residual representer geometry, while both labeled and unlabeled covariates define the target linear functional. This connection follows the generalized Riesz regression perspective, where Bregman divergence and loss-link choices determine both representer fitting and automatic balancing behavior (Kato, 2026a).

4.5 Consistency

We impose the following assumption.

Assumption 4.5. *As $m, l \to \infty$, it holds that $\left\|\widehat{\mu} - \mu_0\right\|_2 = o_p(1)$ or $\left\|\widehat{v}_\beta - v_{0,\beta}\right\|_2 = o_p(1)$.*

Then, the following consistency result holds.

Theorem 4.3 (Consistency in the two-sample scenario)**.** *If Assumptions 4.1 through 4.5 hold, then $\widehat{\tau}_n^{TS\text{-}eff} \xrightarrow{\mathrm{p}} \tau_0$ as $N \to \infty$.*

4.6 Asymptotic Normality

Next, we establish the asymptotic normality of the estimator.

Assumption 4.6. *For each $b \in \mathcal{L}$, as $m, l \to \infty$, the following hold: for each $d \in \{1, 0\}$,*

$$\left\|\mu_0(d, X) - \widehat{\mu}^{(b)}(d, X)\right\|_2 = o_p(1), \left\|v_{0,\beta}(d \mid X) - \widehat{v}_\beta^{(b)}(d \mid X)\right\|_2 = o_p(1),$$
$$\left\|\mu_0(d, X) - \widehat{\mu}^{(b)}(d, X)\right\|_2 \left\|v_{0,\beta}(d \mid X) - \widehat{v}_\beta^{(b)}(d \mid X)\right\|_2 = o_p(1/\sqrt{\min\{m, l\}}).$$

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

Theorem 4.4 (Asymptotic normality in the two-sample scenario)**.** *Let $N = m+l$. Fix $\alpha \in (0, 1)$. Consider the two-sample scenario with sample sizes m, l such that $m = \alpha N$ and $l = (1-\alpha)N$. Suppose that Assumptions 4.1 through 4.5 and 4.6 hold. Also assume that nuisance estimators are constructed via cross-fitting. Then,*

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

Thus, the proposed estimator is efficient with respect to the efficiency bound derived in Theorem 4.2.

For inference, let $\widehat{\psi}_{\mathrm{L},j}^{\mathrm{TS}}$ and $\widehat{\psi}_{\mathrm{U},k}^{\mathrm{TS}}$ denote the two stratum influence functions in Lemma 4.1 evaluated at the cross-fitted nuisance estimators and at $\widehat{\tau}_n^{\text{TS-eff}}$. We estimate the scaled variance by

$$\widehat{V}^{\mathrm{TS}}(\beta) \coloneqq \frac{N}{m}\frac{1}{m}\sum_{j=1}^{m}\left(\widehat{\psi}_{\mathrm{L},j}^{\mathrm{TS}} - \frac{1}{m}\sum_{j'=1}^{m}\widehat{\psi}_{\mathrm{L},j'}^{\mathrm{TS}}\right)^2 + \frac{N}{l}\frac{1}{l}\sum_{k=1}^{l}\left(\widehat{\psi}_{\mathrm{U},k}^{\mathrm{TS}} - \frac{1}{l}\sum_{k'=1}^{l}\widehat{\psi}_{\mathrm{U},k'}^{\mathrm{TS}}\right)^2.$$

The resulting Wald interval uses $\widehat{V}^{\mathrm{TS}}(\beta)/N$ as the variance of $\widehat{\tau}_n^{\text{TS-eff}}$.

Corollary 4.5. *If $p_0(X) = q_0(X)$ almost surely, then every $\beta \in [0, 1]$ defines the same evaluation density. In this case, $V^{TS}(\beta)$ is minimized at $\beta = \alpha$, and*

$$V^{TS}(\alpha) = \frac{1}{\alpha}\mathbb{E}_{p_0}\left[\frac{\sigma_0^2(1, X)}{e_0(1 \mid X)} + \frac{\sigma_0^2(0, X)}{e_0(0 \mid X)}\right] + \mathbb{E}_{p_0}\left[\left(\tau_0(X) - \tau_0\right)^2\right].$$

4.7 Comparison with the One-Sample Scenario

The difference between the one-sample and two-sample scenarios appears in the formulation of ATE estimators, the setup of Riesz regression, and the corresponding efficiency arguments. In the one-sample scenario, the observation indicator is random within a single superpopulation, and the efficient influence function uses $g_0(d \mid X) = P(O = 1, D = d \mid X)$. In the two-sample scenario, the sample membership is fixed by design, and the efficient influence function has separate labeled and unlabeled stratum components. This distinction is the reason why the two-sample bound in Theorem 4.2 uses stratum-specific centering.

5 Many Unlabeled Data

In many applications, we often have access to many more unlabeled data points than fully labeled ones, as unlabeled covariates are less costly to collect. The main results above use the fixed-ratio regime $m/N \to \alpha$. This section records the corresponding sequential implication when the unlabeled sample is asymptotically much larger than the labeled sample.

5.1 Two-Sample Scenario

In the two-sample scenario, let $l/m \to \infty$ and normalize by $\sqrt{m}$. Then, the unlabeled stratum average is negligible, but the labeled stratum still contains both the residual component and the p_0-covariate averaging component when $\beta > 0$.

Corollary 5.1. *Assume the same conditions as in Theorem 4.4. Let $l/m \to \infty$. Then,*

$$\sqrt{m}\left(\widehat{\tau}_n^{TS\text{-}eff} - \tau_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \widetilde{V}^{TS}(\beta)\right),$$

where

$$\widetilde{V}^{TS}(\beta) := \mathbb{E}_{p_0}\left[\left(\frac{\sigma_0^2(1,X)}{e_0(1 \mid X)} + \frac{\sigma_0^2(0,X)}{e_0(0 \mid X)}\right)\left(\frac{\kappa_{0,\beta}(X)}{p_0(X)}\right)^2\right] + \beta^2 \mathbb{E}_{p_0}\left[\left(\tau_0(X) - \tau_{p,0}\right)^2\right].$$

In the target-domain case $\beta = 0$, only the residual component remains under this normalization.

5.2 One-Sample Scenario

The one-sample analogue requires a triangular-array formulation if the probability of observing labels changes with the sample size. To avoid conflating this regime with the fixed-DGP efficiency theory in Section 2, we state the qualitative implication only. If an external covariate sample makes the empirical distribution of X negligible relative to the labeled sample size, then the covariate averaging component in the one-sample efficiency bound is estimated with negligible additional noise, while the conditional outcome-noise component remains governed by the labeled observations. A fully formal triangular-array statement can be added separately if this regime is the focus of the application.

6 Efficiency Gain

By using auxiliary unlabeled covariates, we can reduce the asymptotic variance of ATE estimators. As Hahn (1998) shows, for a labeled dataset $\left\{\left(X_i, D_i, Y_i\right)\right\}_{i=1}^{n^\dagger}$, the efficiency bound of ATE estimators $\widehat{\tau}$ is given as $V^\dagger := \mathbb{E}\left[\frac{\sigma_0^2(1,X)}{P(D=1|X)} + \frac{\sigma_0^2(0,X)}{P(D=0|X)}\right] + \mathbb{E}\left[\left(\tau_0(X) - \tau_0\right)^2\right]$, and an efficient ATE estimator satisfies

$$\sqrt{n^\dagger}\left(\widehat{\tau} - \tau_0\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, V^\dagger\right).$$

The corrected two-sample bound makes the efficiency gain transparent in the same-population case $p_0 = q_0$. If $m/N \to \alpha$ and $\beta = \alpha$, then the semi-supervised bound is

$$V^{\mathrm{TS}}(\alpha) = \frac{1}{\alpha}\mathbb{E}_{p_0}\left[\frac{\sigma_0^2(1,X)}{e_0(1 \mid X)} + \frac{\sigma_0^2(0,X)}{e_0(0 \mid X)}\right] + \mathbb{E}_{p_0}\left[\left(\tau_0(X) - \tau_0\right)^2\right].$$

By contrast, using only the m labeled observations and scaling by $\sqrt{N}$ gives

$$V^{\mathrm{sup}} = \frac{1}{\alpha}\mathbb{E}_{p_0}\left[\frac{\sigma_0^2(1,X)}{e_0(1 \mid X)} + \frac{\sigma_0^2(0,X)}{e_0(0 \mid X)} + \left(\tau_0(X) - \tau_0\right)^2\right].$$

Therefore,

$$V^{\mathrm{sup}} - V^{\mathrm{TS}}(\alpha) = \left(\frac{1}{\alpha} - 1\right)\mathbb{E}_{p_0}\left[\left(\tau_0(X) - \tau_0\right)^2\right].$$

Table 1: Simulation summary. The same-population setting verifies the efficiency gain for ATE and ATT. The covariate-shift setting shows that the unlabeled target covariates remove source-target averaging bias.

Setting	Estimator	Bias	SD	RMSE	N-scaled variance
Same population	Supervised ATE	-0.001	0.067	0.067	18.217
Same population	Semi-supervised ATE	-0.001	0.066	0.066	17.611
Same population	Supervised ATT	-0.001	0.073	0.073	21.050
Same population	Semi-supervised ATT	-0.000	0.071	0.071	20.371
Covariate shift	Source ATE	-0.333	0.067	0.339	18.217
Covariate shift	Target semi-supervised ATE	-0.003	0.130	0.130	67.523
Covariate shift	Source ATT	-0.324	0.073	0.332	21.050
Covariate shift	Target semi-supervised ATT	-0.003	0.150	0.150	89.807

This identity is the main variance-reduction message. Unlabeled covariates reduce the covariate averaging component but do not reduce the residual outcome-noise component.

The same message holds for ATT. In the same-population case $p_0 = q_0$ with $\beta = \alpha$, let $\rho_0 = \mathbb{E}_{p_0}[e_0(X)]$ and $\Delta_0(X) = \tau_0(X) - \tau_0^{\mathrm{ATT}}$. The labeled-only ATT bound under $\sqrt{N}$ scaling contains $\frac{1}{\alpha\rho_0^2}\mathbb{E}_{p_0}\left[e_0(X)^2\Delta_0(X)^2\right]$ as the treated-covariate averaging component. The semi-supervised ATT bound contains the same component without the factor $1/\alpha$. Therefore,

$$V^{\text{ATT-sup}} - V^{\text{ATT-TS}}(\alpha) = \left(\frac{1}{\alpha} - 1\right)\frac{\mathbb{E}_{p_0}\left[e_0(X)^2\Delta_0(X)^2\right]}{\rho_0^2}.$$

Thus, for ATT, unlabeled covariates reduce the treated-covariate averaging component weighted by the squared propensity score.

7 Covariate Shift Adaptation

Uehara et al. (2020) investigates ATE estimation, equivalently, off-policy evaluation, under covariate shift. Our formulation includes their ATE estimation approach as a special case. When $\beta = 0$ in the two-sample scenario, the evaluation density is determined by the unlabeled covariate distribution q_0. In this case, $\kappa_{0,0} = q_0$ and the residual weights are density-ratio weighted through $q_0(X)/p_0(X)$. The fixed-ratio efficiency bound contains the labeled residual component and the unlabeled target-covariate averaging component. If $l/m \to \infty$, the latter vanishes under $\sqrt{m}$ normalization, which yields the usual covariate-shift form.

8 Numerical Illustration

We add a small simulation to verify the efficiency-gain identity and to check the behavior under covariate shift. The numerical illustration uses oracle nuisance quantities so that the experiment isolates the variance-decomposition mechanism. The accompanying code is written in the style of generalized Riesz regression, with the residual weights treated as Riesz representers, but it does not import the genriesz package. The purpose of this section is therefore diagnostic rather than a full benchmark of nuisance-learning algorithms. The experiment is included to verify the main variance decomposition before adding additional implementation-specific variation from first-step learning. A full implementation can replace the oracle representers with estimators obtained from the empirical Bregman objectives in Sections 4.4 and B, following the loss-link workflow of generalized Riesz regression (Kato, 2026a).

In the same-population experiment, the semi-supervised estimators have smaller N-scaled variance than the supervised estimators for both ATE and ATT. In the covariate-shift experiment, the supervised estimators are biased because they average over the source covariate distribution, while the semi-supervised estimators target the evaluation distribution using the unlabeled covariates.

9 Discussion and Limitations

The results rely on several restrictions that are useful to state explicitly. First, all causal targets are identified under unconfoundedness, and the theory does not address unobserved confounding. Second, the two-sample scenario requires support of the evaluation density inside the labeled covariate distribution, because outcomes are observed only under p_0. Third, the mixture parameter β is part of the estimand. Except in special cases such as $p_0 = q_0$, changing β changes the target population rather than merely improving precision. Fourth, the main asymptotic theory uses fixed stratum proportions. The many-unlabeled discussion records the limiting implication of that theory, while a one-sample regime with a label probability converging to zero requires a separate triangular-array formulation. Finally, the numerical illustration isolates the efficiency mechanism using oracle nuisances. It is intended to verify the variance decomposition, while full empirical evaluation of generalized Riesz regression requires replacing the oracle representers with the empirical Bregman objectives.

These limitations do not change the main message. Unlabeled covariates are useful when the causal target averages conditional effects over a covariate law that can be learned more accurately from the auxiliary sample. They do not create outcome information in regions without labeled support, and they do not remove the need for accurate residual correction. This distinction explains both the support condition and the efficiency-gain identities.

10 Conclusion

This study investigates semiparametric efficient estimation of ATE and ATT when auxiliary unlabeled covariates are accessible. We consider both one-sample and two-sample scenarios, and derive semiparametric efficiency bounds for each. Based on the corresponding efficient influence functions, we construct asymptotically efficient estimators via Neyman orthogonal scores. Our approach leverages generalized Riesz regression for estimating nuisance parameters, allowing flexible incorporation of unlabeled covariates. The main implication is that unlabeled covariates reduce the variance associated with averaging the conditional treatment effect over the evaluation covariate density, while the conditional outcome-noise component remains governed by the labeled data. The two-sample analysis also clarifies that the evaluation-density mixture parameter is part of the estimand, and that the efficient influence functions must be centered within each stratum. The ATT extension shows that the same principle continues to hold for debiased ratio targets, but the efficient score must also correct the treatment law and the treatment mass. The proposed framework performs prediction-powered causal inference and extends existing methods for treatment effect estimation under covariate shift, missing labels, and semi-supervised settings.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. arXiv:2107.07511.

David Azriel, Lawrence D. Brown, Michael Sklar, Richard Berk, Andreas Buja, and Linda Zhao. Semi-supervised linear regression. *Journal of the American Statistical Association*, 117(540):2238–2251, 2022.

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

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.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

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, Victor Quintas-Martinez, and Vasilis Syrgkanis. Automatic debiased machine learning via riesz regression, 2021. arXiv:2104.14737.

Victor Chernozhukov, Whitney Newey, Víctor M Quintas-Martínez, and Vasilis Syrgkanis. RieszNet and ForestRiesz: Automatic debiased machine learning with neural nets and random forests. In *International Conference on Machine Learning (ICML)*, 2022a.

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

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.

Ilker Demirel, Ahmed Alaa, Anthony Philippakis, and David Sontag. Prediction-powered generalization of causal inferences. In *International Conference on Machine Learning (ICML)*, 2024.

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.

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

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.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J. Smola. Correcting sample selection bias by unlabeled data. In *NeurIPS*, pp. 601–608. MIT Press, 2007.

Kosuke Imai and Aaron Strauss. Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis*, 19(1):1–19, 2011.

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.

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

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

Masahiro Kato. Direct debiased machine learning via bregman divergence minimization, 2025b. aXiv: 2510.23534.

Masahiro Kato. Nearest neighbor matching as least squares density ratio estimation and riesz regression, 2025c. arXiv: 2510.24433.

Masahiro Kato. A unified theory for causal inference: Direct debiased machine learning via bregman-riesz regression, 2025d.

Masahiro Kato. A unified framework for debiased machine learning: Riesz representer fitting under bregman divergence, 2026a. arXiv: 2601.07752.

Masahiro Kato. Scorematchingriesz: Score matching for debiased machine learning and policy path estimation. In *International Conference on Machine Learning (ICML)*, 2026b.

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, 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.

Masahiro Kato, Fumiaki Kozai, and Ryo Inokuchi. Puate: Semiparametric efficient average treatment effect estimation from treated (positive) and unlabeled units, 2025. arXiv:2501.19345.

Masanori Kawakita and Takafumi Kanamori. Semi-supervised learning with density-ratio estimation. *Machine Learning*, 91(2):189–209, 2013.

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

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.

Kaitlyn J. Lee and Alejandro Schuler. Rieszboost: Gradient boosting for riesz regression, 2025. arXiv: 2501.04871.

Zhexiao Lin, Peng Ding, and Fang Han. Estimation based on nearest neighbor matching: from density ratio to average treatment effect. *Econometrica*, 91(6):2187–2217, 2023.

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

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.

Benjamin Rhodes, Kai Xu, and Michael U. Gutmann. Telescoping density-ratio estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

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.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density ratio matching under the bregman divergence: A unified framework of density ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64, 10 2011.

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

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.

M.J. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Series in Statistics. Springer New York, 2011.

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

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.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24. Curran Associates, Inc., 2011.

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

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.

A DDML

This section explains the DDML framework proposed in Kato (2025b), which refines the arguments about Riesz regression and direct density ratio estimation discussed in Kato (2025a). The core of debiased machine learning is to construct an estimator using the Neyman orthogonal scores (Chernozhukov et al., 2018). For this problem, Kato (2025a;b; 2026a;b) establish the DDML framework, which consists of targeted Neyman estimation and generalized Riesz regression.

A.1 Targeted Neyman estimation

Targeted Neyman estimation formulates the nuisance parameters estimation problem as minimizing the discrepancy between the true Neyman orthogonal scores and their model-based counterparts. Since the Neyman orthogonal score is zero in expectation, we only need to estimate the nuisance parameters so that the sample mean of the Neyman orthogonal score with plug-in parameters is zero. In our setting, we estimate the nuisance parameters μ_0 and g_0, aiming for $\frac{1}{n}\sum_{i=1}^n \psi^{\mathrm{OS}}\big(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \widehat{\mu}, \widehat{\alpha}, \widehat{\tau}\big)$ to be zero, where $\widehat{\mu}$, $\widehat{\alpha}$, and $\widehat{\tau}$ are the estimators of μ_0, α_0, and τ_0. Note that we need to ensure that $\widehat{\tau}$ is asymptotically linear for $\frac{1}{n}\sum_{i=1}^n \psi^{\mathrm{OS}}\big(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \mu_0, \alpha_0, \tau_0\big)$. As discussed in Kato (2025b), the term is decomposed as

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \psi^{\mathrm{OS}}\big(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \widehat{\mu}, \widehat{\alpha}, \widehat{\tau}\big) = \frac{1}{n}\sum_{i=1}^n \Big(\widehat{\alpha}\left(O_i, \widetilde{D}_i, X_i\right)\left(\widetilde{Y}_i - \widehat{\mu}\left(\widetilde{D}_i, X_i\right)\right) + \widehat{\mu}(1, X_i) - \widehat{\mu}(0, X_i) - \widehat{\tau}\Big) \\
&= \frac{1}{n}\sum_{i=1}^n \Big(\left(\widehat{\alpha}\left(O_i, \widetilde{D}_i, X_i\right) - \alpha_0\left(O_i, \widetilde{D}_i, X_i\right)\right)\left(\widetilde{Y}_i - \mu_0\left(\widetilde{D}_i, X_i\right)\right) + \underbrace{\alpha_0\left(O_i, \widetilde{D}_i, X_i\right)\left(\widetilde{Y}_i - \widehat{\mu}\left(\widetilde{D}_i, X_i\right)\right)}_{=(\star)} \\
&+ \underbrace{\widehat{\mu}(1, X_i) - \widehat{\mu}(0, X_i) - \widehat{\tau}}_{=(\star\star)}\Big).
\end{aligned}
$$

A.2 Iterative Procedure for Regression Function and Riesz Representer Estimation

This section explains an approach for estimating the regression function and the Riesz representer using the iterative procedure proposed in Kato (2025b). We do not adopt this approach in the main text, as it complicates the arguments, but we recommend its use.

Targeted maximum likelihood (TMLE) We can make the term $(\star)$ zero in expectation, and we can make the term $(\star\star)$ zero using the TMLE-based ATE estimator. Assume that α_0 is known. If we set $\widehat{\tau}$ as

$$
\widehat{\tau}^{\mathrm{TMLE}} \coloneqq \frac{1}{n}\sum_{i=1}^n \Big(\widehat{\mu}^{\mathrm{TMLE}}(1, X_i) - \widehat{\mu}^{\mathrm{TMLE}}(0, X_i)\Big),
$$

where

$$
\widehat{\mu}^{\mathrm{TMLE}}(d, x) \coloneqq \widehat{\mu}^{(0)}(d, x) + \frac{\sum_{i=1}^n \alpha_0(\widetilde{D}_i, X_i)\big(\widetilde{Y}_i - \widehat{\mu}^{(0)}(\widetilde{D}_i, X_i)\big)}{\sum_{i=1}^n \widehat{\alpha}(\widetilde{D}_i, X_i)^2}\widehat{\alpha}(d, x),
$$

and $\widehat{\mu}^{(0)}(d, x)$ is an initial estimate of $\mu_0(d, x)$. If we set $\widehat{\mu} = \widehat{\mu}^{\mathrm{TMLE}}$ and $\widehat{\tau} = \widehat{\tau}^{\mathrm{TMLE}}$, the terms $(\star)$ and $(\star\star)$ are automatically zero.

Iterative Algorithm As explained above, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n \psi^{\mathrm{OS}}\big(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \widehat{\mu}^{\mathrm{TMLE}}, \widehat{\alpha}, \widehat{\tau}^{\mathrm{TMLE}}\big) \\
&= \frac{1}{n}\sum_{i=1}^n \Big(\left(\widehat{\alpha}\left(O_i, \widetilde{D}_i, X_i\right) - \alpha_0\left(O_i, \widetilde{D}_i, X_i\right)\right)\left(\widetilde{Y}_i - \mu_0\left(\widetilde{D}_i, X_i\right)\right) \Big).
\end{aligned}
$$

Therefore, our target is to minimize

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\Big(\widehat{\alpha}\left(O_i,\widetilde{D}_i,X_i\right)-\alpha_0\left(O_i,\widetilde{D}_i,X_i\right)\Big)\left(\widetilde{Y}_i-\mu_0\left(\widetilde{D}_i,X_i\right)\right)\Big).$$

Here, there are two problems. In minimizing

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\Big(\widehat{\alpha}\left(O_i,\widetilde{D}_i,X_i\right)-\alpha_0\left(O_i,\widetilde{D}_i,X_i\right)\Big)\left(\widetilde{Y}_i-\mu_0\left(\widetilde{D}_i,X_i\right)\right)\Big),$$

we do not know μ_0. In addition, in the TMLE part, we do not know α_0.

If we know μ_0, we can estimate α_0 using the weighted version of generalized Riesz regression proposed in Kato (2025b). In the context of this study, we weight the loss for α_0 by $(\widetilde{Y}_i-\mu_0(\widetilde{D}_i,X_i))$. We omit the details here and refer to Kato (2025b) and Kato (2025d).

Finally, we suggest the following iterative procedure for T steps, following Kato (2025b):

- Obtain an initial estimate of μ_0 and denote it by $\widehat{\mu}^{(0)}$.
- For each $t=1,2,\ldots,T$,
  - Estimate $\widehat{\alpha}^{(t)}$ using weighted generalized Riesz regression with weight $(\widetilde{Y}_i-\widehat{\mu}^{(t-1)}(\widetilde{D}_i,X_i))$.
  - Estimate $\widehat{\mu}^{(t)}$ by the TMLE procedure with $\widehat{\mu}^{(t-1)}$ and $\alpha^{(t)}$ as

$$\widehat{\mu}^{(t)}(d,x)\coloneqq\widehat{\mu}^{(t-1)}(d,x)+\frac{\sum_{i=1}^{n}\widehat{\alpha}^{(t)}(\widetilde{D}_i,X_i)\big(\widetilde{Y}_i-\widehat{\mu}^{(t-1)}(\widetilde{D}_i,X_i)\big)}{\sum_{i=1}^{n}\widehat{\alpha}^{(t)}(\widetilde{D}_i,X_i)^2}\widehat{\alpha}^{(t)}(d,x),$$

A.3 Riesz Regression as Density Ratio Estimation

Riesz regression can be interpreted as a special case of direct density ratio estimation algorithms (Sugiyama et al., 2012; Huang et al., 2007; Kanamori et al., 2009). Therefore, we can employ various estimation techniques as in Yamada et al. (2011), Kiryo et al. (2017), Rhodes et al. (2020), and Kato & Teshima (2021), as well as the methods proposed for Riesz regression (Chernozhukov et al., 2022a) and Lee & Schuler (2025). From this perspective, we can also interpret the nearest neighbor matching ATE estimator as a special case of Riesz regression. These arguments are based on Lin et al. (2023), which finds that the nearest neighbor matching ATE estimator can be interpreted as a density ratio estimation method. For details, see Kato (2025c).

B Average Treatment Effect on Treated Units

This section extends the preceding construction from ATE to ATT. The extension is useful because ATT is often the target when the treated population is the scientific or policy-relevant population. It is also theoretically informative because ATT depends on the treatment law through the target distribution, and the efficient influence function therefore contains a treatment-law correction.

B.1 One-Sample Scenario

In the one-sample scenario, define $s_0(X)\coloneqq\pi_0(1\mid X)$ and $e_0(X)\coloneqq e_0(1\mid X)$. For ATT, we strengthen the missingness condition so that the full labeled variables are missing at random.

Assumption B.1 (Treatment missing at random)**.** *It holds that* $(D,Y(1),Y(0))\perp\!\!\!\perp O\mid X$.

Under Assumption B.1, the propensity score in the full population is identified from the observed labeled part. Let

$$\rho_0\coloneqq\mathbb{E}\left[e_0(X)\right],\Delta_0(X)\coloneqq\tau_0(X)-\tau_0^{\mathrm{ATT}},$$

where $\tau_0^{\mathrm{ATT}}\coloneqq\mathbb{E}\left[e_0(X)\tau_0(X)\right]/\rho_0$.

Lemma B.1 (Efficient influence function for ATT in the one-sample scenario)**.** *Suppose that Assumptions 3.1 through 3.3 and B.1 hold. Then, the efficient influence function for τ_0^{ATT} is*

$$\psi^{ATT\text{-}OS}(X,O,\widetilde{D},\widetilde{Y}) \coloneqq \frac{1}{\rho_0}\left[\frac{O}{s_0(X)}\Big(\widetilde{D}\Big(\widetilde{Y}-\mu_0(1,X)\Big)-\frac{e_0(X)(1-\widetilde{D})}{1-e_0(X)}\Big(\widetilde{Y}-\mu_0(0,X)\Big)+\Big(\widetilde{D}-e_0(X)\Big)\Delta_0(X)\Big)+e_0(X)\Delta_0(\right.$$

The proof is provided in Appendix G.

The corresponding efficiency bound is $V^{\text{ATT-OS}} \coloneqq \mathbb{E}\left[\psi^{\text{ATT-OS}}(X,O,\widetilde{D},\widetilde{Y})^2\right]$. Equivalently, if

$$C_0^{\text{ATT}}(X) \coloneqq e_0(X)\sigma_0^2(1,X)+\frac{e_0(X)^2}{1-e_0(X)}\sigma_0^2(0,X)+e_0(X)(1-e_0(X))\Delta_0(X)^2,$$

then

$$V^{\text{ATT-OS}} = \frac{1}{\rho_0^2}\mathbb{E}\left[\frac{C_0^{\text{ATT}}(X)}{s_0(X)}+e_0(X)^2\Delta_0(X)^2\right].$$

This decomposition has the same interpretation as the ATE bound. The residual outcome-noise part is controlled by labeled observations, while the covariate averaging part can be improved by using unlabeled covariates.

The efficient ATT estimator is a ratio estimator. Let $\widehat{s}$, $\widehat{e}$, and $\widehat{\mu}$ be cross-fitted nuisance estimators and set $\widehat{\tau}(x) \coloneqq \widehat{\mu}(1,x)-\widehat{\mu}(0,x)$. Define

$$\widehat{A}^{\text{ATT-OS}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left[\frac{O_i}{\widehat{s}(X_i)}\Big(\widetilde{D}_i\Big(\widetilde{Y}_i-\widehat{\mu}(1,X_i)\Big)-\frac{\widehat{e}(X_i)(1-\widetilde{D}_i)}{1-\widehat{e}(X_i)}\Big(\widetilde{Y}_i-\widehat{\mu}(0,X_i)\Big)+\Big(\widetilde{D}_i-\widehat{e}(X_i)\Big)\widehat{\tau}(X_i)\Big)+\widehat{e}(X_i)\widehat{\tau}(X_i)\right],$$

$$\widehat{B}^{\text{ATT-OS}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left[\frac{O_i}{\widehat{s}(X_i)}\Big(\widetilde{D}_i-\widehat{e}(X_i)\Big)+\widehat{e}(X_i)\right].$$

Then, we estimate ATT by

$$\widehat{\tau}_n^{\text{ATT-OS}} \coloneqq \frac{\widehat{A}^{\text{ATT-OS}}}{\widehat{B}^{\text{ATT-OS}}}.$$

The denominator is also debiased. This is important for efficiency because the treatment mass ρ_0 is unknown and must be estimated from partially labeled data. A simple sufficient high-level condition for the remainder terms is that, for $a \in \{1,0\}$,

$$\left\|\widehat{\mu}(a,X)-\mu_0(a,X)\right\|_2+\left\|\widehat{e}(X)-e_0(X)\right\|_2+\left\|\widehat{s}(X)-s_0(X)\right\|_2 = o_p(n^{-1/4}),$$

with uniform boundedness away from the overlap and observation-probability boundaries. This condition is stronger than necessary but makes the ratio expansion checkable.

For the asymptotic result, we use the following high-level expansion, which can be verified from the usual cross-fitting and product-rate arguments. Let $A_0^{\text{ATT-OS}} = \mathbb{E}\left[e_0(X)\tau_0(X)\right]$ and $B_0^{\text{ATT-OS}} = \rho_0$. We assume that

$$\widehat{A}^{\text{ATT-OS}}-A_0^{\text{ATT-OS}} = \frac{1}{n}\sum_{i=1}^{n}\left(A_i^{\text{ATT-OS}}-A_0^{\text{ATT-OS}}\right)+o_p(n^{-1/2}),$$

$$\widehat{B}^{\text{ATT-OS}}-B_0^{\text{ATT-OS}} = \frac{1}{n}\sum_{i=1}^{n}\left(B_i^{\text{ATT-OS}}-B_0^{\text{ATT-OS}}\right)+o_p(n^{-1/2}),$$

where $A_i^{\text{ATT-OS}}$ and $B_i^{\text{ATT-OS}}$ are the corresponding efficient one-step signals obtained by replacing the estimated nuisances in $\widehat{A}^{\text{ATT-OS}}$ and $\widehat{B}^{\text{ATT-OS}}$ with the true nuisances. This condition makes explicit that the ratio theorem only requires first-order expansions for the numerator and denominator.

Theorem B.2 (Asymptotic normality for ATT in the one-sample scenario)**.** *Suppose that Assumptions 3.1 through 3.3 and B.1 hold. Also assume that the nuisance estimators $\widehat{s}$, $\widehat{e}$, and $\widehat{\mu}$ are constructed via cross-fitting, are uniformly bounded away from the boundary, and satisfy the numerator and denominator expansions stated above. Then,*

$$\sqrt{n}\left(\widehat{\tau}_n^{ATT\text{-}OS}-\tau_0^{ATT}\right) \xrightarrow{\mathrm{d}} \mathcal{N}\left(0,V^{ATT\text{-}OS}\right).$$

B.2 Two-Sample Scenario

In the two-sample scenario, ATT is defined with respect to $\kappa_{0,\beta}$. We assume that the treatment assignment law is invariant across the labeled and unlabeled covariate populations.

Assumption B.2 (Treatment-law invariance for ATT)**.** *There exists a common propensity score $e_0(X) = P(D = 1 \mid X)$ that is shared by the labeled population and the evaluation population induced by $\kappa_{0,\beta}$.*

Define

$$\rho_{0,\beta} \coloneqq \mathbb{E}_{\kappa_{0,\beta}}\left[e_0(X)\right], \tau_{0,\beta}^{\mathrm{ATT}} \coloneqq \frac{\mathbb{E}_{\kappa_{0,\beta}}\left[e_0(X)\tau_0(X)\right]}{\rho_{0,\beta}}, \Delta_{0,\beta}(X) \coloneqq \tau_0(X) - \tau_{0,\beta}^{\mathrm{ATT}}.$$

Let

$$\bar{\Delta}_{p,\beta} \coloneqq \mathbb{E}_{p_0}\left[e_0(X)\Delta_{0,\beta}(X)\right], \bar{\Delta}_{q,\beta} \coloneqq \mathbb{E}_{q_0}\left[e_0(Z)\Delta_{0,\beta}(Z)\right].$$

Then, $\beta\bar{\Delta}_{p,\beta} + (1-\beta)\bar{\Delta}_{q,\beta} = 0$.

Lemma B.3 (Efficient influence functions for ATT in the two-sample scenario)**.** *Suppose that Assumptions 4.1 through 4.4 and B.2 hold. The efficient influence functions for the labeled and unlabeled strata are*

$$\psi_L^{ATT\text{-}TS}(X,D,Y) \coloneqq \frac{1}{\rho_{0,\beta}}\Bigg[\omega_{0,\beta}(X)\Big(D\Big(Y-\mu_0(1,X)\Big) - \frac{e_0(X)(1-D)}{1-e_0(X)}\Big(Y-\mu_0(0,X)\Big) + \Big(D-e_0(X)\Big)\Delta_{0,\beta}(X)\Big) + \beta\Big(e_0(X$$

$$\psi_U^{ATT\text{-}TS}(Z) \coloneqq \frac{1-\beta}{\rho_{0,\beta}}\Big(e_0(Z)\Delta_{0,\beta}(Z) - \bar{\Delta}_{q,\beta}\Big).$$

The proof is provided in Appendix I.

Let

$$C_{0,\beta}^{\mathrm{ATT}}(X) \coloneqq e_0(X)\sigma_0^2(1,X) + \frac{e_0(X)^2}{1-e_0(X)}\sigma_0^2(0,X) + e_0(X)(1-e_0(X))\Delta_{0,\beta}(X)^2.$$

Then the two-sample ATT efficiency bound is

$$\begin{aligned} V^{\text{ATT-TS}}(\beta) &\coloneqq \frac{1}{\alpha}\mathbb{E}_{p_0}\left[\psi_{\mathrm{L}}^{\text{ATT-TS}}(X,D,Y)^2\right] + \frac{1}{1-\alpha}\mathbb{E}_{q_0}\left[\psi_{\mathrm{U}}^{\text{ATT-TS}}(Z)^2\right] \\ &= \frac{1}{\alpha\rho_{0,\beta}^2}\mathbb{E}_{p_0}\left[\omega_{0,\beta}(X)^2 C_{0,\beta}^{\mathrm{ATT}}(X) + \beta^2\Big(e_0(X)\Delta_{0,\beta}(X) - \bar{\Delta}_{p,\beta}\Big)^2\right] \\ &\quad + \frac{(1-\beta)^2}{(1-\alpha)\rho_{0,\beta}^2}\mathbb{E}_{q_0}\left[\Big(e_0(Z)\Delta_{0,\beta}(Z) - \bar{\Delta}_{q,\beta}\Big)^2\right]. \end{aligned}$$

The ATT bound is parallel to the ATE bound but contains the additional treatment-law term $e_0(X)$ and the treatment residual $D - e_0(X)$.

The corresponding estimator is

$$\widehat{\tau}_n^{\text{ATT-TS}} \coloneqq \frac{\widehat{A}^{\text{ATT-TS}}}{\widehat{B}^{\text{ATT-TS}}},$$

where

$$\begin{aligned} \widehat{A}^{\text{ATT-TS}} &\coloneqq \frac{1}{m}\sum_{j=1}^{m}\widehat{\omega}_\beta(X_j)\Big(D_j\Big(Y_j - \widehat{\mu}(1,X_j)\Big) - \frac{\widehat{e}(X_j)(1-D_j)}{1-\widehat{e}(X_j)}\Big(Y_j - \widehat{\mu}(0,X_j)\Big) + \Big(D_j - \widehat{e}(X_j)\Big)\widehat{\tau}(X_j)\Big) \\ &\quad + \beta\frac{1}{m}\sum_{j=1}^{m}\widehat{e}(X_j)\widehat{\tau}(X_j) + (1-\beta)\frac{1}{l}\sum_{k=1}^{l}\widehat{e}(Z_k)\widehat{\tau}(Z_k), \\ \widehat{B}^{\text{ATT-TS}} &\coloneqq \frac{1}{m}\sum_{j=1}^{m}\widehat{\omega}_\beta(X_j)\Big(D_j - \widehat{e}(X_j)\Big) + \beta\frac{1}{m}\sum_{j=1}^{m}\widehat{e}(X_j) + (1-\beta)\frac{1}{l}\sum_{k=1}^{l}\widehat{e}(Z_k). \end{aligned}$$

Under the same cross-fitting and product-rate conditions as in Theorem 4.4, with the corresponding conditions for $\widehat{e}$, the estimator is asymptotically linear with the influence functions in Lemma B.3. More explicitly, it is enough that the numerator and denominator admit first-order two-stratum expansions with remainders $o_p(N^{-1/2})$, namely,

$$\widehat{A}^{\text{ATT-TS}} - A_0^{\text{ATT-TS}} = \frac{1}{m}\sum_{j=1}^{m}\Big(A_{\text{L},j}^{\text{ATT-TS}} - A_{\text{L},0}^{\text{ATT-TS}}\Big) + \frac{1}{l}\sum_{k=1}^{l}\Big(A_{\text{U},k}^{\text{ATT-TS}} - A_{\text{U},0}^{\text{ATT-TS}}\Big) + o_p(N^{-1/2}),$$

$$\widehat{B}^{\text{ATT-TS}} - B_0^{\text{ATT-TS}} = \frac{1}{m}\sum_{j=1}^{m}\Big(B_{\text{L},j}^{\text{ATT-TS}} - B_{\text{L},0}^{\text{ATT-TS}}\Big) + \frac{1}{l}\sum_{k=1}^{l}\Big(B_{\text{U},k}^{\text{ATT-TS}} - B_{\text{U},0}^{\text{ATT-TS}}\Big) + o_p(N^{-1/2}),$$

where the signals are evaluated at the true nuisances. A convenient sufficient condition is that, for $a \in \{1,0\}$,

$$\big\|\widehat{\mu}(a,X) - \mu_0(a,X)\big\|_{p_0,2} + \big\|\widehat{\omega}_\beta(X) - \omega_{0,\beta}(X)\big\|_{p_0,2} + \big\|\widehat{e}(X) - e_0(X)\big\|_{p_0,2} + \big\|\widehat{e}(Z) - e_0(Z)\big\|_{q_0,2} = o_p(N^{-1/4}),$$

where $\|\cdot\|_{p_0,2}$ and $\|\cdot\|_{q_0,2}$ denote L_2 norms under the labeled and unlabeled covariate laws. As in the ATE theorem, these sufficient rates can be weakened to product-rate conditions.

Theorem B.4 (Asymptotic normality for ATT in the two-sample scenario)**.** *Let $N = m + l$, $m/N \to \alpha$, and $l/N \to 1-\alpha$ for some $\alpha \in (0,1)$. Suppose that Assumptions 4.1 through 4.4 and B.2 hold. Also assume that the nuisance estimators $\widehat{\omega}_\beta$, $\widehat{e}$, and $\widehat{\mu}$ are constructed via cross-fitting, are uniformly bounded away from the boundary, and satisfy the numerator and denominator expansions stated above. Then,*

$$\sqrt{N}\left(\widehat{\tau}_n^{ATT\text{-}TS} - \tau_{0,\beta}^{ATT}\right) \xrightarrow{\text{d}} \mathcal{N}\left(0, V^{ATT\text{-}TS}(\beta)\right).$$

B.3 Variance Estimation for ATT

For inference, we use plug-in empirical variances of the estimated influence functions. In the one-sample scenario, let $\widehat{\psi}_i^{\text{ATT-OS}}$ denote the influence function in Lemma B.1 evaluated at the cross-fitted nuisance estimators and at $\widehat{\tau}_n^{\text{ATT-OS}}$. We estimate the variance by

$$\widehat{V}^{\text{ATT-OS}} \coloneqq \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\psi}_i^{\text{ATT-OS}} - \frac{1}{n}\sum_{i'=1}^{n}\widehat{\psi}_{i'}^{\text{ATT-OS}}\Big)^2.$$

In the two-sample scenario, let $\widehat{\psi}_{\text{L},j}^{\text{ATT-TS}}$ and $\widehat{\psi}_{\text{U},k}^{\text{ATT-TS}}$ denote the estimated labeled and unlabeled stratum influence functions. We estimate the scaled variance by

$$\begin{aligned}\widehat{V}^{\text{ATT-TS}}(\beta) \coloneqq{}& \frac{N}{m}\frac{1}{m}\sum_{j=1}^{m}\Big(\widehat{\psi}_{\text{L},j}^{\text{ATT-TS}} - \frac{1}{m}\sum_{j'=1}^{m}\widehat{\psi}_{\text{L},j'}^{\text{ATT-TS}}\Big)^2 \\ &+ \frac{N}{l}\frac{1}{l}\sum_{k=1}^{l}\Big(\widehat{\psi}_{\text{U},k}^{\text{ATT-TS}} - \frac{1}{l}\sum_{k'=1}^{l}\widehat{\psi}_{\text{U},k'}^{\text{ATT-TS}}\Big)^2.\end{aligned}$$

The same construction applies to ATE by replacing the ATT influence functions with the corresponding ATE influence functions in Lemma 3.1 and Lemma 4.1.

B.4 Semi-Supervised Generalized Riesz Regression for ATT

The ATT numerator has a treatment-weighted linear functional. In the one-sample scenario, the residual Riesz representer for this numerator is

$$\alpha_0^{\text{ATT-OS}}(O,\widetilde{D},X) \coloneqq \frac{O}{s_0(X)}\Big(\widetilde{D} - \frac{e_0(X)(1-\widetilde{D})}{1-e_0(X)}\Big).$$

It satisfies

$$\mathbb{E}\Big[\alpha_0^{\text{ATT-OS}}(O,\widetilde{D},X)h(\widetilde{D},X)\Big] = \mathbb{E}\Big[e_0(X)\Big(h(1,X) - h(0,X)\Big)\Big].$$

Thus, the one-sample ATT generalized Riesz regression objective is obtained from equation 1 by replacing the ATE target functional with the treatment-weighted functional above. The unlabeled covariates enter the target functional through the average over X, while the labeled observations identify the residual inner product through $O/s_0(X)$.

In the two-sample scenario, the residual Riesz representer is

$$\alpha_{0,\beta}^{\mathrm{ATT}}(D,X) \coloneqq \omega_{0,\beta}(X)\Big(D - \frac{e_0(X)(1-D)}{1-e_0(X)}\Big).$$

It satisfies

$$\mathbb{E}_{p_0}\left[\alpha_{0,\beta}^{\mathrm{ATT}}(D,X)h(D,X)\right] = \mathbb{E}_{\kappa_{0,\beta}}\left[e_0(X)\Big(h(1,X)-h(0,X)\Big)\right].$$

Therefore, the ATT version of generalized Riesz regression can be obtained by replacing the target linear functional in Section 4.4 with its treatment-weighted counterpart. For a convex function f, define

$$\begin{aligned}\mathrm{BD}_{f,\beta}^{\text{ATT-TS}}(\alpha) \coloneqq{}& \mathbb{E}_{p_0}\left[\partial f\left(\alpha(D,X)\right)\alpha(D,X) - f\left(\alpha(D,X)\right)\right] \\ &- \beta\mathbb{E}_{p_0}\left[e_0(X)\Big(\partial f\left(\alpha(1,X)\right) - \partial f\left(\alpha(0,X)\right)\Big)\right] \\ &- (1-\beta)\mathbb{E}_{q_0}\left[e_0(Z)\Big(\partial f\left(\alpha(1,Z)\right) - \partial f\left(\alpha(0,Z)\right)\Big)\right].\end{aligned}$$

This objective shows why ATT is not a purely cosmetic extension of ATE. The target functional itself contains e_0, so the Riesz objective for ATT combines direct representer fitting with propensity-score learning. This connection is consistent with the general Riesz-representer formulation in Kato (2026a), where ATT appears as one of the causal functionals that can be represented by a problem-specific Riesz representer.

In implementation, one may first estimate e_0 on the labeled sample and then fit $\alpha_{0,\beta}^{\mathrm{ATT}}$ by the empirical version of $\mathrm{BD}_{f,\beta}^{\text{ATT-TS}}$. Alternatively, one may use a structured model that parameterizes e_0 and $\alpha_{0,\beta}^{\mathrm{ATT}}$ jointly. The first approach separates the treatment-law nuisance from the residual representer, while the second approach is closer to the loss-link construction of generalized Riesz regression. In both cases, the final ATT estimator should use the debiased ratio form above, because direct minimization of the Riesz objective alone does not debias the treatment mass $\rho_{0,\beta}$.

C Proof for Lemma 3.1: Efficient Influence Function in the One-Sample Scenario

We provide the proof of Lemma 3.1. Our proof strategy is inspired by the approaches in Hahn (1998) and Kato et al. (2025).

Proof procedure Their proof considers a nonparametric model for the distribution of potential outcomes and defines regular parametric submodels. The procedure involves the following steps: (i) characterizing the tangent set for all regular parametric submodels, (ii) verifying that the parameter of interest is pathwise differentiable, (iii) confirming that the proposed semiparametric efficient influence function lies within the tangent set, and (iv) calculating the expectation of the squared influence function.

Proof. In Section 2, we defined the probability density function for $(X, O, \widetilde{D}, \widetilde{Y})$ as

$$\begin{aligned}&p_0\left(x,o,\widetilde{d},\widetilde{y}\right) = \\ &p_0(x)\pi_0(0\mid X)^{\mathbb{1}[o=0]}\Big(g_0(1\mid x)r_{Y(1),0}\left(\widetilde{y}\mid x\right)\Big)^{\mathbb{1}\left[o=1,\widetilde{d}=1\right]}\Big(g_0(0\mid x)r_{Y(0),0}\left(\widetilde{y}\mid \widetilde{d}=0\right)\Big)^{\mathbb{1}\left[o=1,\widetilde{d}=0\right]},\end{aligned}$$

where $r_{Y(1),0}(y\mid x)$ and $r_{Y(0),0}(y\mid x)$ are the conditional densities of $Y(1)$ and $Y(0)$. Recall that for each $a\in\{1,0\}$, we have

$$g_0(a\mid x) = P(D=a, O=1\mid X) = \pi_0(1\mid X)e_0(a\mid X)$$

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 probability density function:

$$\begin{aligned} p\left(x, o, \widetilde{d}, \widetilde{y}; \theta\right) &= p(x;\theta)\Big(\pi(1 \mid x;\theta)\Big)^{\mathbb{1}[o=0]} \\ &\cdot \Big(g(1 \mid x;\theta) r_{Y(1)}(y \mid x;\theta)\Big)^{\mathbb{1}\left[o=1,\widetilde{d}=1\right]} \Big(g(0 \mid x;\theta) r_{Y(0)}(y \mid x;\theta)\Big)^{\mathbb{1}\left[o=1,\widetilde{d}=0\right]}. \end{aligned}$$

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

$$p(x, o, \widetilde{d}, \widetilde{y}; \theta_0) = p_0(x, o, \widetilde{d}, \widetilde{y}).$$

We can define such a parametric submodel, as shown in van der Vaart (1998).

Then, we define score functions (the derivative of the log likelihood function) as follows:

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

where

$$\begin{aligned} S_X(x;\theta) &\coloneqq \frac{\partial}{\partial \theta} \log p(x;\theta), \\ S_{Y(d)}(y \mid x;\theta) &\coloneqq \frac{\partial}{\partial \theta} \log r_{Y(d)}(y \mid x;\theta), \text{ for } d \in \{1,0\}, \\ \dot{\pi}(o \mid x;\theta) &\coloneqq \frac{\partial}{\partial \theta} \pi(o \mid x;\theta), \text{ for } d \in \{1,0\}, \\ \dot{g}(a \mid x;\theta) &\coloneqq \frac{\partial}{\partial \theta} g(a \mid x;\theta), \text{ for } a \in \{1,0\}. \end{aligned}$$

Using the parametric submodels and their score functions, we denote the tangent space as $\mathcal{T} \coloneqq \{S(x, o, y; \theta)\}$.

Under the parametric submodels, We redefine the ATE as a function of θ as

$$\tau(\theta) \coloneqq \iint y(1) r_{Y(1)}(y(1) \mid x;\theta) p(x;\theta) \mathrm{d}y(1) \mathrm{d}x - \iint y(0) r_{Y(0)}(y(0) \mid x;\theta) p(x;\theta) \mathrm{d}y(0) \mathrm{d}x.$$

Them, the derivative of the ATE function is given as

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

where

$$\tau(X;\theta) \coloneqq \mu(1, X;\theta) - \mu(0, X;\theta),$$

and $\mu(d, X;\theta) \coloneqq \int y(d) r_{Y(d)}(y(d) \mid x;\theta) p(x;\theta) \mathrm{d}y(d)$.

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

$$\frac{\partial \tau(\theta)}{\partial \theta}\Big|_{\theta=\theta_0} = \mathbb{E}\big[\psi(X, O, \widetilde{D}, \widetilde{Y}) S(X, O, \widetilde{D}, \widetilde{Y}; \theta_0)\big]. \tag{3}$$

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

$$\begin{aligned}
&\psi^{\mathrm{OS}}(X, O, \widetilde{D}, \widetilde{Y}; \mu_0, g_0, \tau_0) \\
&= S^{\mathrm{OS}}(X, O, Y; \mu_0, g_0) - \tau_0, \\
&= \frac{\mathbb{1}\left[O = 1, \widetilde{D} = 1\right]\left(\widetilde{Y} - \mu_0(1, X)\right)}{g_0(1 \mid X)} - \frac{\mathbb{1}\left[O = 1, \widetilde{D} = 0\right]\left(\widetilde{Y} - \mu_0(0, X)\right)}{g_0(0 \mid X)} \\
&+ \mu_0(1, X) - \mu_0(0, X) - \tau_0.
\end{aligned}$$

We prove that $\psi^{\mathrm{OS}}(X, O, \widetilde{D}, \widetilde{Y}; \mu_0, g_0, \tau_0)$ is actually the unique efficient influence function by verifying that ψ^{OS} satisfies equation 3 and $\psi^{\mathrm{OS}} \in \mathcal{T}$.

Proof of equation 3: First, we confirm that ψ^{OS} satisfies equation 3. We have

$$\begin{aligned}
&\mathbb{E}\left[\psi^{\mathrm{OS}}(X, O, Y; \mu_0, g_0) S(X, O, \widetilde{D}, \widetilde{Y}; \tau_0)\right] \\
&= \mathbb{E}\Bigg[\psi^{\mathrm{OS}}(X, O, Y; \mu_0, g_0) \\
&\cdot \Big(S_X(X; \theta) + \mathbb{1}[O = 0] \frac{\dot{\pi}(1 \mid X; \theta)}{\pi(1 \mid X; \theta)} \\
&+ \mathbb{1}\left[O = 1, \widetilde{D} = 1\right]\left(S_{Y(1)}(Y \mid X; \theta_0) + \frac{\dot{g}(1 \mid X; \theta)}{g(1 \mid X; \theta_0)}\right) + \mathbb{1}\left[O = 1, \widetilde{D} = 0\right]\left(S_{Y(0)}(Y \mid X; \theta) + \frac{\dot{g}(0 \mid X; \theta_0)}{g(0 \mid X; \theta_0)}\right)\Big)\Bigg] \\
&= \mathbb{E}\Bigg[\left(\frac{\mathbb{1}\left[O = 1, \widetilde{D} = 1\right]\left(Y - \mu_0(1, X)\right)}{g_0(1 \mid X)} - \frac{\mathbb{1}\left[O = 1, \widetilde{D} = 0\right]\left(Y - \mu_0(0, X)\right)}{g_0(0 \mid X)} + \mu_0(1, X) - \mu_0(0, X) - \tau_0\right) \\
&\cdot \Bigg(S_X(X; \theta) + \mathbb{1}[O = 0] \frac{\dot{\pi}(1 \mid X; \theta)}{\pi(1 \mid X; \theta)} \\
&+ \mathbb{1}\left[O = 1, \widetilde{D} = 1\right]\left(S_{Y(1)}(\widetilde{Y} \mid X; \theta_0) + \frac{\dot{g}(1 \mid X; \theta)}{g(1 \mid X; \theta_0)}\right) + \mathbb{1}\left[O = 1, \widetilde{D} = 0\right]\left(S_{Y(0)}(\widetilde{Y} \mid X; \theta) + \frac{\dot{g}(0 \mid X; \theta_0)}{g(0 \mid X; \theta_0)}\right)\Bigg)\Bigg] \\
&= \mathbb{E}\Bigg[\left(\mu_0(1, X) - \mu_0(0, X) - \tau_0\right)\left(S_X(X; \theta_0) + \mathbb{1}[O = 0] \frac{\dot{\pi}(1 \mid X; \theta)}{\pi(1 \mid X; \theta)}\right) \\
&+ \left(\frac{\mathbb{1}\left[O = 1, \widetilde{D} = 1\right]\left(\widetilde{Y} - \mu_0(1, X)\right)}{g_0(1 \mid X)} + \mu_0(1, X) - \mu_0(0, X) - \tau_0\right) \\
&\cdot \mathbb{1}\left[O = 1, \widetilde{D} = 1\right]\left(S_{Y(1)}(Y \mid X; \theta_0) + \frac{\dot{g}(1 \mid X; \theta_0)}{g(1 \mid X; \theta_0)}\right) \\
&- \left(\frac{\mathbb{1}\left[O = 1, \widetilde{D} = 0\right]\left(\widetilde{Y} - \mu_0(0, X)\right)}{g_0(0 \mid X)} + \mu_0(1, X) - \mu_0(0, X) - \tau_0\right) \\
&\cdot \mathbb{1}\left[O = 1, \widetilde{D} = 0\right]\left(S_{Y(0)}(Y \mid X; \theta) + \frac{\dot{g}(0 \mid X; \theta_0)}{g(0 \mid X; \theta_0)}\right)\Bigg],
\end{aligned}$$

where we used $\mathbb{1}\left[O = 1, \widetilde{D} = 1\right]\mathbb{1}\left[O = 1, \widetilde{D} = 0\right] = 0$, $\mathbb{1}[O = 1, \widetilde{D} = d]\mathbb{1}[O = 0] = 0$, and

$$\mathbb{E}\left[\frac{\mathbb{1}\left[O = 1, \widetilde{D} = 1\right]\left(\widetilde{Y} - \mu_0(1, X)\right)}{g_0(1 \mid X)}\right] = \mathbb{E}\left[\frac{\mathbb{1}\left[O = 1, \widetilde{D} = 1\right]\left(Y(1) - \mu_0(1, X)\right)}{g_0(1 \mid X)}\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\frac{g_0(1 \mid X)\Big(\mu_0(1,X) - \mu_0(1,X)\Big)}{g_0(1 \mid X)}\right] = 0, \\
&\mathbb{E}\left[\frac{\mathbb{1}\left[O=1,\widetilde{D}=0\right]\Big(Y - \mu_0(0,X)\Big)}{g_0(0 \mid X)}\right] = \mathbb{E}\left[\frac{\mathbb{1}\left[O=1,\widetilde{D}=0\right]\Big(\widetilde{Y} - \mu_0(0,X)\Big)}{g_0(0 \mid X)}\right] \\
&= \mathbb{E}\left[\frac{g_0(0 \mid X)\Big(\mu_0(0,X) - \mu_0(0,X)\Big)}{g_0(0 \mid X)}\right] = 0.
\end{aligned}
$$

We have

$$
\begin{aligned}
&\mathbb{E}\Bigg[\Big(\mu_0(1,X) - \mu_0(0,X) - \tau_0\Big)\left(S_X(X;\theta_0) + \mathbb{1}[O=0]\frac{\dot{\pi}(1 \mid X;\theta)}{\pi(1 \mid X;\theta)}\right) \\
&+ \left(\frac{\mathbb{1}\left[O=1,\widetilde{D}=1\right]\Big(\widetilde{Y} - \mu_0(1,X)\Big)}{g_0(1 \mid X)} + \mu_0(1,X) - \mu_0(0,X) - \tau_0\right) \\
&\cdot \mathbb{1}\left[O=1,\widetilde{D}=1\right]\left(S_{Y(1)}(Y \mid X;\theta_0) + \frac{\dot{g}(1 \mid X;\theta_0)}{g(1 \mid X;\theta_0)}\right) \\
&- \left(\frac{\mathbb{1}\left[O=1,\widetilde{D}=0\right]\Big(\widetilde{Y} - \mu_0(0,X)\Big)}{g_0(0 \mid X)} + \mu_0(1,X) - \mu_0(0,X) - \tau_0\right) \\
&\cdot \mathbb{1}\left[O=1,\widetilde{D}=0\right]\left(S_{Y(0)}(Y \mid X;\theta) + \frac{\dot{g}(0 \mid X;\theta_0)}{g(0 \mid X;\theta_0)}\right)\Bigg] \\
&= \mathbb{E}\Bigg[\Big(\mu_0(1,X) - \mu_0(0,X)\Big)S_X(X;\theta_0) \\
&+ \frac{\mathbb{1}\left[O=1,\widetilde{D}=1\right]\Big(Y - \mu_0(1,X)\Big)}{g_0(1 \mid X)}S_{Y(1)}(Y \mid X;\theta_0) \\
&- \frac{\mathbb{1}\left[O=1,\widetilde{D}=0\right]\Big(Y - \mu_0(0,X)\Big)}{g_0(0 \mid X)}S_{Y(0)}(Y \mid X;\theta)\Bigg] \\
&= \mathbb{E}\Bigg[\Big(\mu_0(1,X) - \mu_0(0,X)\Big)S_X(X;\theta_0) \\
&+ \frac{\mathbb{1}\left[O=1,\widetilde{D}=1\right]Y(1)}{g_0(1 \mid X)}S_{Y(1)}(Y(1) \mid X;\theta) - \frac{\mathbb{1}\left[O=1,\widetilde{D}=0\right]\widetilde{Y}}{g_0(0 \mid X)}S_{Y(0)}(Y(0) \mid X;\theta_0)\Bigg],
\end{aligned}
$$

where we used

$$
\begin{aligned}
&\mathbb{E}\Big[\tau_0 S_X(X;\theta)\Big] = 0 \\
&\mathbb{E}\left[\Big(\mu_0(1,X) - \mu_0(0,X) - \tau_0\Big)\mathbb{1}[O=1]\left(S_{Y(1)}(Y \mid X;\theta_0) + \frac{\dot{g}(1 \mid X;\theta_0)}{g_0(1 \mid X;\theta_0)}\right)\right] = 0.
\end{aligned}
$$

Finally, we have

$$
\mathbb{E}\Bigg[\Big(\mu_0(1,X) - \mu_0(0,X)\Big)S_X(X;\theta_0)
$$

$$
\begin{aligned}
&+ \frac{\mathbb{1}[O=1]\Big(Y-\mu_0(1,X)\Big)}{g_0(1\mid X)} S_{Y(1)}(Y\mid X;\theta_0) \\
&- \frac{\mathbb{1}[O=0]\Big(Y-\mu_0(0,X)\Big)}{g_0(0\mid X)\pi_0(0\mid X)} S_{Y(0)}(Y\mid X;\theta)\Bigg] \\
&= \mathbb{E}\Bigg[\Big(\mu_0(1,X)-\mu_0(0,X)\Big) S_X(X;\theta_0) \\
&+ \frac{\mathbb{1}\Big[O=1,\widetilde{D}=1\Big]Y(1)}{g_0(1\mid X)} S_{Y(1)}(Y(1)\mid X;\theta) - \frac{\mathbb{1}\Big[O=1,\widetilde{D}=0\Big]\widetilde{Y}}{g_0(0\mid X)} S_{Y(0)}(Y(0)\mid X;\theta_0)\Bigg] \\
&= \mathbb{E}\Big[Y(1)S_{Y(1)}(Y(1)\mid X;\theta_0)\Big] - \mathbb{E}\Big[Y(0)S_{Y(0)}(\widetilde{Y}\mid X;\theta_0)\Big] \\
&+ \mathbb{E}_{\theta_0}\Big[\tau(X;\theta)S_X(X;\theta_0)\Big] \\
&= \frac{\partial\tau(\theta)}{\partial\theta}\Big|_{\theta=\theta_0}
\end{aligned}
$$

Proof of $\psi^{\mathbf{OS}} \in \mathcal{T}$: Set

$$
\begin{aligned}
S_{Y(d)}(y\mid x) &= \frac{y-\mu_0(d\mid x)}{g_0(d\mid x)}, \\
S_X(X;\theta) &= \mu_0(1,X)-\mu_0(0,X)-\tau_0.
\end{aligned}
$$

Then, $\psi^{\mathrm{OS}} \in \mathcal{T}$ holds. □

D Proof of Theorem 3.5: Efficient ATE Estimator under the One-Sample Scenario

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 $b\in\{1,2\}$, we denote the subset of the dataset in cross-fitting as

$$
\mathcal{D}^{(b)} \coloneqq \left\{\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}\right)\right\}_{i=1}^{\overline{n}}.
$$

We defined the estimator as

$$
\widehat{\tau}_n^{\text{OS-eff}} \coloneqq \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{OS}}\left(X_i, O_i, \widetilde{D}_i, \widetilde{Y}_i; \widehat{\mu}_{n,i}, \widehat{g}_{n,i}\right),
$$

where recall that

$$
\begin{aligned}
&S^{\mathrm{OS}}\left(X, O, \widetilde{D}, \widetilde{Y}; \widehat{\mu}_{n,i}, \widehat{g}_{n,i}\right) \\
&= \frac{\mathbb{1}\Big[O=1,\widetilde{D}=1\Big]\Big(\widetilde{Y}-\widehat{\mu}_{n,i}(1,X)\Big)}{\widehat{g}_{n,i}(1\mid X)} - \frac{\mathbb{1}\Big[O=1,\widetilde{D}=0\Big]\Big(\widetilde{Y}-\widehat{\mu}_{n,i}(0,X)\Big)}{\widehat{g}_{n,i}(0\mid X)} \\
&+ \widehat{\mu}_{n,i}(1,X)-\widehat{\mu}_{n,i}(0,X).
\end{aligned}
$$

We have

$$
\begin{aligned}
\widehat{\tau}_n^{\text{OS-eff}} &= \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{OS}}\left(X_i, O_i, Y_i; \widehat{\mu}_{n,i}, \widehat{g}_{n,i}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{OS}}(X_i, O_i, Y_i; \mu_0, g_0) - \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{OS}}(X_i, O_i, Y_i; \mu_0, g_0) + \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{OS}}(X_i, O_i, Y_i; \widehat{\mu}_{n,i}, \widehat{g}_{n,i}).
\end{aligned}
$$

Here, if it holds that

$$\frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{OS}}(X_i, O_i, Y_i; \mu_0, g_0) - \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{OS}}(X_i, O_i, Y_i; \widehat{\mu}_{n,i}, \widehat{g}_{n,i}) = o_p(1/\sqrt{n}) \tag{4}$$

then we have

$$\begin{aligned}\sqrt{n}\Big(\widehat{\tau}_n^{\text{OS-eff}} - \tau_0\Big) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} S^{\mathrm{OS}}(X_i, O_i, Y_i; \mu_0, g_0) + o_p(1) \\ &\xrightarrow{\mathrm{d}} \mathcal{N}(0, V^{\mathrm{OS}}),\end{aligned}$$

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

Therefore, we prove Theorem 3.5 by showing equation 4. We decompose the LHS of equation 4 as

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{OS}}(X_i, O_i, Y_i; \mu_0, g_0) - \frac{1}{n}\sum_{i=1}^{n} S^{\mathrm{OS}}(X_i, O_i, Y_i; \widehat{\mu}_{n,i}, \widehat{g}_{n,i}) \\ &= \frac{\overline{n}}{n}\sum_{b\in\{1,2\}}\left(\frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - \frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right)\right).\end{aligned}$$

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

$$\begin{aligned}&\frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - \frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \\ &= \frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - \frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \\ &- \Bigg(\mathbb{E}\Big[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\Big] \\ &- \mathbb{E}\Big[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\Big]\Bigg) \\ &+ \Bigg(\mathbb{E}\Big[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\Big] \\ &- \mathbb{E}\Big[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\Big]\Bigg).\end{aligned}$$

To show equation 4, we show the following two inequalities separately:

$$\begin{aligned}&\frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - \frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \\ &- \left(\mathbb{E}\Big[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\Big] - \mathbb{E}\Big[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\Big]\right) \\ &= o_p(1/\sqrt{n}), \qquad (5) \\ &\mathbb{E}\Big[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\Big] - \mathbb{E}\Big[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\Big] \\ &= o_p(1/\sqrt{n}). \qquad (6)\end{aligned}$$

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.

D.1 Proof of equation 5

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

$$\lim_{n\to\infty} \Pr\left(\sqrt{\overline{n}}\left|\frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - \frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right)\right.\right.$$
$$\left.\left. - \left(\mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\right] - \mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\right]\right)\right| > \varepsilon\right)$$
$$= 0. \tag{7}$$

We show equation 7 by showing that for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr\left(\sqrt{\overline{n}}\left|\frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - \frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right)\right.\right.$$
$$\left.\left. - \left(\mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\right] - \mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\right]\right)\right| \geq \varepsilon \mid \mathcal{D}^{(b)}\right)$$
$$= 0. \tag{8}$$

If equation 8 holds, then equation 7 also holds from dominated convergence theorem.

We prove equation 8 using Chebychev's inequality. From Chebychev's inequality we have

$$\Pr\left(\sqrt{\overline{n}}\left|\frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - \frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right)\right.\right.$$
$$\left.\left. - \left(\mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\right] - \mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\right]\right)\right| \geq \varepsilon \mid \mathcal{D}^{(b)}\right)$$
$$\leq \frac{\overline{n}}{\varepsilon}\mathrm{Var}\left(\frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - \frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right)\right.$$
$$\left. - \left(\mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\right] - \mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\right]\right) \mid \mathcal{D}^{(b)}\right).$$

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

$$m\mathrm{Var}\left(\frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - \frac{1}{\overline{n}}\sum_{i=1}^{\overline{n}} S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right)\right.$$
$$\left. - \left(\mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\right] - \mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\right]\right) \mid \mathcal{D}^{(b)}\right)$$
$$= \mathrm{Var}\left(S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right)\right.$$
$$\left. - \left(\mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\right] - \mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\right]\right) \mid \mathcal{D}^{(b)}\right)$$
$$= \mathbb{E}\left[\left(S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) - S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right)\right.\right. \tag{9}$$
$$\left.\left. - \left(\mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0\right) \mid \mathcal{D}^{(b)}\right] - \mathbb{E}\left[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)}\right) \mid \mathcal{D}^{(b)}\right]\right)\right)^2 \mid \mathcal{D}^{(b)}\right].$$

The term equation 9 converges to zero in probability as $n \to \infty$ if

$$\left\|\mu_0 - \widehat{\mu}_n^{(b)}\right\|_2 = o_p(1), , \left\|g_0 - \widehat{g}_n^{(b)}\right\|_2 = o_p(1)$$

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 equation 9, respectively. Then, we have

$$\textit{equation } 9 = \mathbb{E}\left[\left(A - B - \mathbb{E}\left[A - B \mid \mathcal{D}^{(b)}\right]\right)^2 \mid \mathcal{D}^{(b)}\right].$$

Here, we have

$$\textit{equation } 9 = \mathbb{E}\left[(A-B)^2 \mid \mathcal{D}^{(b)}\right] - \left(\mathbb{E}\left[A - B \mid \mathcal{D}^{(b)}\right]\right)^2 \leq \mathbb{E}\left[(A-B)^2 \mid \mathcal{D}^{(b)}\right].$$

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

$$\begin{aligned}
&S^{\mathrm{OS}}\left(X, O, \widetilde{D}, \widetilde{Y}; \mu, g\right) \\
&= \frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right]\left(\widetilde{Y} - \mu_0(1, X)\right)}{g(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right]\left(\widetilde{Y} - \mu(0, X)\right)}{g(0 \mid X)} + \mu(1, X) - \mu(0, X).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&A - B \\
&= \frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right]\left(\widetilde{Y} - \mu_0(1, X)\right)}{g_0(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right]\left(\widetilde{Y} - \mu_0(0, X)\right)}{g_0(0 \mid X)} + \mu_0(1, X) - \mu_0(0, X) \\
&- \left(\frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right]\left(\widetilde{Y} - \widehat{\mu}_n^{(b)}(1, X)\right)}{\widehat{g}_n^{(b)}(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right]\left(\widetilde{Y} - \widehat{\mu}_n^{(b)}(0, X)\right)}{\widehat{g}_n^{(b)}(0 \mid X)} + \widehat{\mu}_n^{(b)}(1, X) - \widehat{\mu}_n^{(b)}(0, X)\right)
\end{aligned}$$

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:

$$\left(\mu_0(1, X) - \mu_0(0, X)\right) - \left(\widehat{\mu}_0^{(b)}(1, X) - \widehat{\mu}_0^{(b)}(0, X)\right).$$

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

$$\begin{aligned}
(\star) = &\frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right]\left(\widetilde{Y} - \mu_0(1, X)\right)}{g_0(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right]\left(\widetilde{Y} - \mu_0(0, X)\right)}{g_0(0 \mid X)} \\
&- \left(\frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right]\left(\widetilde{Y} - \widehat{\mu}_n^{(b)}(1, X)\right)}{\widehat{g}_n^{(b)}(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right]\left(\widetilde{Y} - \widehat{\mu}_n^{(b)}(0, X)\right)}{\widehat{g}_n^{(b)}(0 \mid X)}\right).
\end{aligned}$$

Next, we have

$$(\star) = \frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right]\left(\widetilde{Y} - \mu_0(1, X)\right)}{g_0(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right]\left(\widetilde{Y} - \mu_0(0, X)\right)}{g_0(0 \mid X)}$$

$$
\begin{aligned}
&- \left(\frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right](\widetilde{Y} - \mu_0(1,X))}{\widehat{g}_n^{(b)}(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right](\widetilde{Y} - \mu_0(0,X))}{\widehat{g}_n^{(b)}(0 \mid X)} \right) \\
&+ \left(\frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right](\widetilde{Y} - \mu_0(1,X))}{\widehat{g}_n^{(b)}(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right](\widetilde{Y} - \mu_0(0,X))}{\widehat{g}_n^{(b)}(0 \mid X)} \right) \\
&- \left(\frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right](\widetilde{Y} - \widehat{\mu}_n^{(b)}(1,X))}{\widehat{g}_n^{(b)}(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right](\widetilde{Y} - \widehat{\mu}_n^{(b)}(0,X))}{\widehat{g}_n^{(b)}(0 \mid X)} \right).
\end{aligned}
$$

Then, from the parallelogram law, we have

$$
\begin{aligned}
(\star)^2 \leq\ & 2 \left(\frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right](\widetilde{Y} - \mu_0(1,X))}{g_0(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right](\widetilde{Y} - \mu_0(1,X))}{\widehat{g}_n^{(b)}(1 \mid X)} \right)^2 \\
&+ 2 \left(\frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right](\widetilde{Y} - \mu_0(0,X))}{\widehat{g}_{n,i}(0 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right](\widetilde{Y} - \mu_0(0,X))}{\widehat{g}_n^{(b)}(0 \mid X)} \right)^2 \\
&+ \cdots \\
&+ 2 \left(\frac{g_0(1 \mid X)\mathbb{1}\left[O=1, \widetilde{D}=1\right](\widetilde{Y} - \mu_0(1,X))}{g_0(0 \mid X)\widehat{g}_n^{(b)}(1 \mid X)} - \frac{g_0(1 \mid X)\mathbb{1}\left[O=1, \widetilde{D}=1\right](\widetilde{Y} - \widehat{\mu}_n^{(b)}(1,X))}{g_0(0 \mid X)\widehat{g}_n^{(b)}(1 \mid X)} \right)^2.
\end{aligned}
$$

Here, we can bound

$$
2\mathbb{E}\left[\left(\frac{g_0(1 \mid X)\mathbb{1}\left[O=1, \widetilde{D}=1\right](\widetilde{Y} - \mu_0(1,X))}{\widehat{g}_n^{(b)}(1 \mid X)} - \frac{g_0(1 \mid X)\mathbb{1}\left[O=1, \widetilde{D}=1\right](\widetilde{Y} - \widehat{\mu}_n^{(b)}(1,X))}{\widehat{g}_n^{(b)}(0 \mid X)} \right)^2 \mid \mathcal{D}^{(b)} \right]
$$

by

$$
C\mathbb{E}\left[\left(\mu_0(1,X) - \widehat{\mu}_n^{(b)}(1,X) \right)^2 \right],
$$

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. □

D.2 Proof of equation 6

Proof. We have

$$
\begin{aligned}
&\mathbb{E}\Big[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \mu_0, g_0 \right) \mid \mathcal{D}^{(b)} \Big] - \mathbb{E}\Big[S^{\mathrm{OS}}\left(X_i^{(b)}, O_i^{(b)}, \widetilde{D}_i^{(b)}, \widetilde{Y}_i^{(b)}; \widehat{\mu}_n^{(b)}, \widehat{g}_n^{(b)} \right) \mid \mathcal{D}^{(b)} \Big] \\
&= \mathbb{E}\left[\frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right]\left(Y - \mu_0(1,X)\right)}{g_0(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right]\left(Y - \mu_0(0,X)\right)}{g_0(0 \mid X)} + \mu_0(1,X) - \mu_0(0,X) \right] \\
&- \mathbb{E}\left[\frac{\mathbb{1}\left[O=1, \widetilde{D}=1\right]\left(Y - \widehat{\mu}_n^{(b)}(1,X)\right)}{\widehat{g}_n^{(b)}(1 \mid X)} - \frac{\mathbb{1}\left[O=1, \widetilde{D}=0\right]\left(Y - \widehat{\mu}_n^{(b)}(0,X)\right)}{\widehat{g}_n^{(b)}(0 \mid X)} + \widehat{\mu}_n^{(b)}(1,X) - \widehat{\mu}_n^{(b)}(0,X) \right] \\
&= \mathbb{E}\Big[\mu_0(1,X) - \mu_0(0,X) \Big] \\
&- \mathbb{E}\left[\frac{g_0(1 \mid X)\Big(\mu_0(1,X) - \widehat{\mu}_n^{(b)}(1,X)\Big)}{\widehat{g}_n^{(b)}(1 \mid X)} - \frac{g_0(0 \mid X)\Big(\mu_0(0,X) - \widehat{\mu}_n^{(b)}(0,X)\Big)}{\widehat{g}_n^{(b)}(0 \mid X)} + \widehat{\mu}_n^{(b)}(1,X) - \widehat{\mu}_n^{(b)}(0,X) \right]
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}\left[\left(1 - \frac{g_0(1 \mid X)}{\widehat{g}_n^{(b)}(1 \mid X)}\right)\left(\mu_0(1,X) - \widehat{\mu}_n^{(b)}(1,X)\right)\right] + \mathbb{E}\left[\left(1 - \frac{g_0(0 \mid X)}{\widehat{g}_n^{(b)}(0 \mid X)}\right)\left(\mu_0(0,X) - \widehat{\mu}_n^{(b)}(0,X)\right)\right] \\
&\leq C \sum_{d \in \{1,0\}} \sqrt{\mathbb{E}\left[\left(\widehat{g}_n^{(b)}(d \mid X) - g_0(d \mid X)\right)^2\right]\mathbb{E}\left[\left(\mu_0(d,X) - \widehat{\mu}_n^{(b)}(d,X)\right)^2\right]} \\
&= o_p(1/\sqrt{n}),
\end{aligned}
$$

where we used Cauchy-Schwarz inequality. □

E Proof of Lemma 4.1: Efficient Influence Function in the Two-Sample Scenario

This section provides the proof of Lemma 4.1. The argument follows the efficiency theory for stratified sampling schemes used in Uehara et al. (2020). We present the proof in a form that keeps the two strata separate.

Consider regular parametric submodels $p(x, d, y; \theta)$ and $q(z; \theta)$ through $p_0(x, d, y)$ and $q_0(z)$. Let $S_p(X, D, Y; \theta)$ and $S_q(Z; \theta)$ denote the corresponding scores. Under the fixed evaluation density

$$
\kappa_\beta(x; \theta) = \beta p(x; \theta) + (1 - \beta) q(x; \theta),
$$

the target parameter is

$$
\tau(\theta) = \beta \mathbb{E}_{p(\theta)}\left[\tau(X; \theta)\right] + (1 - \beta)\mathbb{E}_{q(\theta)}\left[\tau(Z; \theta)\right],
$$

where $\tau(x; \theta) = \mu(1, x; \theta) - \mu(0, x; \theta)$. Its derivative at θ_0 is

$$
\begin{aligned}
\left.\frac{\partial \tau(\theta)}{\partial \theta}\right|_{\theta=\theta_0} &= \mathbb{E}_{p_0}\left[\omega_{0,\beta}(X)\Big(Y(1)S_{Y(1)}(Y(1) \mid X) - Y(0)S_{Y(0)}(Y(0) \mid X)\Big)\right] \\
&\quad + \beta \mathbb{E}_{p_0}\left[\left(\tau_0(X) - \tau_{p,0}\right) S_X(X)\right] + (1 - \beta)\mathbb{E}_{q_0}\left[\left(\tau_0(Z) - \tau_{q,0}\right) S_Z(Z)\right].
\end{aligned}
$$

Here, $S_{Y(d)}(Y(d) \mid X)$ is the conditional outcome score, $S_X(X)$ is the covariate score under p_0, and $S_Z(Z)$ is the covariate score under q_0.

The first term is represented using the labeled data by

$$
\begin{aligned}
&\mathbb{E}_{p_0}\left[S^{\mathrm{TS}}_{(X,D,Y)}(X, D, Y; \mu_0, v_{0,\beta}) S_p(X, D, Y)\right] \\
&= \mathbb{E}_{p_0}\left[\omega_{0,\beta}(X)\Big(Y(1)S_{Y(1)}(Y(1) \mid X) - Y(0)S_{Y(0)}(Y(0) \mid X)\Big)\right].
\end{aligned}
$$

Moreover, because

$$
\mathbb{E}_{p_0}\left[S^{\mathrm{TS}}_{(X,D,Y)}(X, D, Y; \mu_0, v_{0,\beta}) \mid X\right] = 0,
$$

the residual term is orthogonal to every function of X. Hence the labeled stratum gradient is

$$
\psi^{\mathrm{TS}}_{\mathrm{L}}(X, D, Y; \mu_0, v_{0,\beta}) = S^{\mathrm{TS}}_{(X,D,Y)}(X, D, Y; \mu_0, v_{0,\beta}) + \beta\left(\tau_0(X) - \tau_{p,0}\right).
$$

The unlabeled stratum gradient is

$$
\psi^{\mathrm{TS}}_{\mathrm{U}}(Z; \mu_0) = (1 - \beta)\left(\tau_0(Z) - \tau_{q,0}\right).
$$

These functions have mean zero under their respective strata and reproduce the pathwise derivative for all regular parametric submodels. They are therefore the efficient influence functions for the two strata.

The variance calculation uses the independence of the two samples and the conditional mean zero property of the residual term. We have

$$
\mathbb{E}_{p_0}\left[S^{\mathrm{TS}}_{(X,D,Y)}(X, D, Y; \mu_0, v_{0,\beta})^2\right] = \mathbb{E}_{p_0}\left[\left(\frac{\sigma_0^2(1, X)}{e_0(1 \mid X)} + \frac{\sigma_0^2(0, X)}{e_0(0 \mid X)}\right)\left(\frac{\kappa_{0,\beta}(X)}{p_0(X)}\right)^2\right],
$$

and

$$\mathbb{E}_{p_0}\left[S^{\mathrm{TS}}_{(X,D,Y)}(X,D,Y;\mu_0,v_{0,\beta})\,(\tau_0(X)-\tau_{p,0})\right]=0.$$

Therefore, for $N=m+l$, $m/N\to\alpha$, and $l/N\to 1-\alpha$, the scaled efficiency bound is

$$\begin{aligned}V^{\mathrm{TS}}(\beta)&=\frac{1}{\alpha}\mathbb{E}_{p_0}\left[\psi^{\mathrm{TS}}_{\mathrm{L}}(X,D,Y;\mu_0,v_{0,\beta})^2\right]+\frac{1}{1-\alpha}\mathbb{E}_{q_0}\left[\psi^{\mathrm{TS}}_{\mathrm{U}}(Z;\mu_0)^2\right]\\&=\frac{1}{\alpha}\mathbb{E}_{p_0}\left[\left(\frac{\sigma_0^2(1,X)}{e_0(1\mid X)}+\frac{\sigma_0^2(0,X)}{e_0(0\mid X)}\right)\left(\frac{\kappa_{0,\beta}(X)}{p_0(X)}\right)^2\right]\\&+\frac{\beta^2}{\alpha}\mathbb{E}_{p_0}\left[\Big(\tau_0(X)-\tau_{p,0}\Big)^2\right]+\frac{(1-\beta)^2}{1-\alpha}\mathbb{E}_{q_0}\left[\Big(\tau_0(Z)-\tau_{q,0}\Big)^2\right].\end{aligned}$$

This proves Lemma 4.1 and Theorem 4.2.

F Proof of Theorem 4.4: Efficient ATE estimator under the Two-Sample Scenario

Recall that the two-sample estimator is

$$\begin{aligned}\widehat{\tau}^{\text{TS-eff}}_n&=\frac{1}{m}\sum_{j=1}^{m}S^{\mathrm{TS}}_{(X,D,Y)}(X_j,D_j,Y_j;\widehat{\mu}^{(b)},\widehat{v}^{(b)}_{\beta})\\&+\beta\frac{1}{m}\sum_{j=1}^{m}S^{\mathrm{TS}}_{(X)}(X_j;\widehat{\mu}^{(b)})+(1-\beta)\frac{1}{l}\sum_{k=1}^{l}S^{\mathrm{TS}}_{(X)}(Z_k;\widehat{\mu}^{(b)}).\end{aligned}$$

Using the same cross-fitting argument as in the proof of Theorem 3.5, the nuisance estimation remainder is second order. Assumption 4.6 implies

$$\begin{aligned}\widehat{\tau}^{\text{TS-eff}}_n-\tau_0&=\frac{1}{m}\sum_{j=1}^{m}\psi^{\mathrm{TS}}_{\mathrm{L}}(X_j,D_j,Y_j;\mu_0,v_{0,\beta})\\&+\frac{1}{l}\sum_{k=1}^{l}\psi^{\mathrm{TS}}_{\mathrm{U}}(Z_k;\mu_0)+o_p(N^{-1/2}).\end{aligned}$$

The two leading sums are independent. Therefore,

$$\begin{aligned}\sqrt{N}\left(\widehat{\tau}^{\text{TS-eff}}_n-\tau_0\right)&=\frac{1}{\sqrt{\alpha m}}\sum_{j=1}^{m}\psi^{\mathrm{TS}}_{\mathrm{L}}(X_j,D_j,Y_j;\mu_0,v_{0,\beta})\\&+\frac{1}{\sqrt{(1-\alpha)l}}\sum_{k=1}^{l}\psi^{\mathrm{TS}}_{\mathrm{U}}(Z_k;\mu_0)+o_p(1).\end{aligned}$$

By the central limit theorem for independent triangular arrays with fixed stratum proportions,

$$\sqrt{N}\left(\widehat{\tau}^{\text{TS-eff}}_n-\tau_0\right)\xrightarrow{\mathrm{d}}\mathcal{N}\left(0,V^{\mathrm{TS}}(\beta)\right).$$

This completes the proof.

G Proof of Lemma B.1: Efficient Influence Function for ATT in the One-Sample Scenario

We prove Lemma B.1 by first deriving the full-data influence function and then applying the missing-label transformation. Define

$$R_0^{\mathrm{ATT}}(D,Y,X)\coloneqq D\Big(Y-\mu_0(1,X)\Big)-\frac{e_0(X)(1-D)}{1-e_0(X)}\Big(Y-\mu_0(0,X)\Big).$$

The numerator of ATT is $A_0 \coloneqq \mathbb{E}\left[e_0(X)\tau_0(X)\right]$ and the denominator is $\rho_0 \coloneqq \mathbb{E}\left[e_0(X)\right]$. The full-data influence functions for A_0 and ρ_0 are

$$\begin{aligned}\phi_A(D,Y,X) &\coloneqq R_0^{\mathrm{ATT}}(D,Y,X) + \Big(D - e_0(X)\Big)\tau_0(X) + e_0(X)\tau_0(X) - A_0,\\ \phi_\rho(D,X) &\coloneqq D - \rho_0.\end{aligned}$$

Hence, by the quotient rule, the full-data influence function for $\tau_0^{\mathrm{ATT}} = A_0/\rho_0$ is

$$\begin{aligned}\phi_{\mathrm{full}}^{\mathrm{ATT}}(D,Y,X) &\coloneqq \frac{1}{\rho_0}\Big(\phi_A(D,Y,X) - \tau_0^{\mathrm{ATT}}\phi_\rho(D,X)\Big)\\ &= \frac{1}{\rho_0}\left[R_0^{\mathrm{ATT}}(D,Y,X) + \Big(D - e_0(X)\Big)\Delta_0(X) + e_0(X)\Delta_0(X)\right].\end{aligned}$$

The conditional mean of this full-data influence function given X is

$$\mathbb{E}\left[\phi_{\mathrm{full}}^{\mathrm{ATT}}(D,Y,X) \mid X\right] = \frac{e_0(X)\Delta_0(X)}{\rho_0},$$

because $\mathbb{E}\left[R_0^{\mathrm{ATT}}(D,Y,X) \mid X\right] = 0$ and $\mathbb{E}\left[D - e_0(X) \mid X\right] = 0$. Under Assumption B.1, the observed-data efficient influence function is obtained by the standard MAR projection formula,

$$\psi^{\text{ATT-OS}}(X,O,\widetilde{D},\widetilde{Y}) = \frac{O}{s_0(X)}\Big(\phi_{\mathrm{full}}^{\mathrm{ATT}}(D,Y,X) - \mathbb{E}\left[\phi_{\mathrm{full}}^{\mathrm{ATT}}(D,Y,X) \mid X\right]\Big) + \mathbb{E}\left[\phi_{\mathrm{full}}^{\mathrm{ATT}}(D,Y,X) \mid X\right].$$

Substituting the expressions above gives the formula in Lemma B.1. The variance decomposition follows from the conditional mean-zero identities. In particular,

$$\begin{aligned}\mathbb{E}\left[\Big(R_0^{\mathrm{ATT}}(D,Y,X) + \Big(D - e_0(X)\Big)\Delta_0(X)\Big)^2 \mid X\right] &= e_0(X)\sigma_0^2(1,X) + \frac{e_0(X)^2}{1-e_0(X)}\sigma_0^2(0,X)\\ &\quad + e_0(X)(1-e_0(X))\Delta_0(X)^2.\end{aligned}$$

This proves the stated bound.

H Proof of Theorem B.2: Efficient ATT Estimator in the One-Sample Scenario

Let $A_0 \coloneqq \mathbb{E}\left[e_0(X)\tau_0(X)\right]$ and $\rho_0 \coloneqq \mathbb{E}\left[e_0(X)\right]$. Under the cross-fitting and product-rate conditions in Theorem B.2, the debiased numerator and denominator satisfy

$$\begin{aligned}\widehat{A}^{\text{ATT-OS}} - A_0 &= \frac{1}{n}\sum_{i=1}^{n}\phi_A(O_i,\widetilde{D}_i,\widetilde{Y}_i,X_i) + o_p(n^{-1/2}),\\ \widehat{B}^{\text{ATT-OS}} - \rho_0 &= \frac{1}{n}\sum_{i=1}^{n}\phi_\rho(O_i,\widetilde{D}_i,X_i) + o_p(n^{-1/2}),\end{aligned}$$

where the observed-data versions of ϕ_A and ϕ_ρ are obtained by the same MAR transformation as in Appendix G. Applying the delta method to a/b at (A_0,ρ_0) yields

$$\begin{aligned}\widehat{\tau}_n^{\text{ATT-OS}} - \tau_0^{\mathrm{ATT}} &= \frac{1}{\rho_0}\Big(\widehat{A}^{\text{ATT-OS}} - A_0 - \tau_0^{\mathrm{ATT}}\Big(\widehat{B}^{\text{ATT-OS}} - \rho_0\Big)\Big) + o_p(n^{-1/2})\\ &= \frac{1}{n}\sum_{i=1}^{n}\psi^{\text{ATT-OS}}(X_i,O_i,\widetilde{D}_i,\widetilde{Y}_i) + o_p(n^{-1/2}).\end{aligned}$$

The central limit theorem gives the desired asymptotic normality.

I Proof of Lemma B.3: Efficient Influence Functions for ATT in the Two-Sample Scenario

We derive the two-sample ATT influence functions by applying the quotient rule to the numerator and denominator under the stratified sampling scheme. Let

$$A_{0,\beta} \coloneqq \mathbb{E}_{\kappa_{0,\beta}}\left[e_0(X)\tau_0(X)\right], \rho_{0,\beta} \coloneqq \mathbb{E}_{\kappa_{0,\beta}}\left[e_0(X)\right], \tau_{0,\beta}^{\mathrm{ATT}} = A_{0,\beta}/\rho_{0,\beta}.$$

For the numerator, the labeled stratum gradient is

$$\begin{aligned}\phi_{\mathrm{L}}^{A}(X,D,Y) \coloneqq \omega_{0,\beta}(X)\Big(D\Big(Y-\mu_0(1,X)\Big) - \frac{e_0(X)(1-D)}{1-e_0(X)}\Big(Y-\mu_0(0,X)\Big) + \Big(D-e_0(X)\Big)\tau_0(X)\Big) \\ + \beta\Big(e_0(X)\tau_0(X) - \mathbb{E}_{p_0}\left[e_0(X)\tau_0(X)\right]\Big).\end{aligned}$$

The unlabeled stratum gradient for the numerator is

$$\phi_{\mathrm{U}}^{A}(Z) \coloneqq (1-\beta)\Big(e_0(Z)\tau_0(Z) - \mathbb{E}_{q_0}\left[e_0(Z)\tau_0(Z)\right]\Big).$$

For the denominator, the labeled and unlabeled stratum gradients are

$$\begin{aligned}\phi_{\mathrm{L}}^{\rho}(X,D) \coloneqq \omega_{0,\beta}(X)\Big(D-e_0(X)\Big) + \beta\Big(e_0(X) - \mathbb{E}_{p_0}\left[e_0(X)\right]\Big), \\ \phi_{\mathrm{U}}^{\rho}(Z) \coloneqq (1-\beta)\Big(e_0(Z) - \mathbb{E}_{q_0}\left[e_0(Z)\right]\Big).\end{aligned}$$

Therefore, the quotient rule gives

$$\psi_{\mathrm{L}}^{\text{ATT-TS}}(X,D,Y) = \frac{1}{\rho_{0,\beta}}\Big(\phi_{\mathrm{L}}^{A}(X,D,Y) - \tau_{0,\beta}^{\mathrm{ATT}}\phi_{\mathrm{L}}^{\rho}(X,D)\Big),$$

and

$$\psi_{\mathrm{U}}^{\text{ATT-TS}}(Z) = \frac{1}{\rho_{0,\beta}}\Big(\phi_{\mathrm{U}}^{A}(Z) - \tau_{0,\beta}^{\mathrm{ATT}}\phi_{\mathrm{U}}^{\rho}(Z)\Big).$$

Substituting $\Delta_{0,\beta}(X) = \tau_0(X) - \tau_{0,\beta}^{\mathrm{ATT}}$ yields the two functions in Lemma B.3. The centered terms appear because the labeled and unlabeled strata have their own sampling distributions.

It remains to compute the variance. The residual part has conditional mean zero given X, and it is orthogonal to the covariate-score component. Thus,

$$\begin{aligned}&\mathbb{E}_{p_0}\left[\Big(\omega_{0,\beta}(X)\Big(D\Big(Y-\mu_0(1,X)\Big) - \frac{e_0(X)(1-D)}{1-e_0(X)}\Big(Y-\mu_0(0,X)\Big) + \Big(D-e_0(X)\Big)\Delta_{0,\beta}(X)\Big)\Big)^2\right] \\ &= \mathbb{E}_{p_0}\left[\omega_{0,\beta}(X)^2 C_{0,\beta}^{\mathrm{ATT}}(X)\right].\end{aligned}$$

Combining this equality with the independence of the two strata gives the bound displayed in Lemma B.3.

J Proof of Theorem B.4: Efficient ATT Estimator in the Two-Sample Scenario

Under the stated cross-fitting and product-rate conditions, the debiased numerator and denominator satisfy the following two-stratum expansions:

$$\begin{aligned}\widehat{A}^{\text{ATT-TS}} - A_{0,\beta} &= \frac{1}{m}\sum_{j=1}^{m}\phi_{\mathrm{L}}^{A}(X_j,D_j,Y_j) + \frac{1}{l}\sum_{k=1}^{l}\phi_{\mathrm{U}}^{A}(Z_k) + o_p(N^{-1/2}), \\ \widehat{B}^{\text{ATT-TS}} - \rho_{0,\beta} &= \frac{1}{m}\sum_{j=1}^{m}\phi_{\mathrm{L}}^{\rho}(X_j,D_j) + \frac{1}{l}\sum_{k=1}^{l}\phi_{\mathrm{U}}^{\rho}(Z_k) + o_p(N^{-1/2}).\end{aligned}$$

Applying the delta method to a/b at $(A_{0,\beta}, \rho_{0,\beta})$ gives

$$\widehat{\tau}_n^{\text{ATT-TS}} - \tau_{0,\beta}^{\text{ATT}} = \frac{1}{m}\sum_{j=1}^{m} \psi_{\text{L}}^{\text{ATT-TS}}(X_j, D_j, Y_j) + \frac{1}{l}\sum_{k=1}^{l} \psi_{\text{U}}^{\text{ATT-TS}}(Z_k) + o_p(N^{-1/2}).$$

The two sums are independent. Hence,

$$\begin{aligned}\sqrt{N}\left(\widehat{\tau}_n^{\text{ATT-TS}} - \tau_{0,\beta}^{\text{ATT}}\right) = &\frac{1}{\sqrt{\alpha m}}\sum_{j=1}^{m} \psi_{\text{L}}^{\text{ATT-TS}}(X_j, D_j, Y_j) \\ &+ \frac{1}{\sqrt{(1-\alpha) l}}\sum_{k=1}^{l} \psi_{\text{U}}^{\text{ATT-TS}}(Z_k) + o_p(1).\end{aligned}$$

The central limit theorem for independent strata yields the stated limiting distribution with variance $V^{\text{ATT-TS}}(\beta)$.