Conditional Front-door Adjustment for Heterogeneous Treatment Assignment Effect Estimation Under Non-adherence

Winston Chen   CHENWT@UMICH.COM
Trenton Chang   CTRENTON@UMICH.COM
Jenna Wiens   WIENSJ@UMICH.COM
University of Michigan, Ann Arbor, MI, USA

Abstract

Estimates of heterogeneous treatment assignment effects can inform treatment decisions. Under the presence of non-adherence (*e.g.*, patients do not adhere to their assigned treatment), both the standard backdoor adjustment (SBD) and the conditional front-door adjustment (CFD) can recover unbiased estimates of the treatment assignment effects. However, the estimation variance of these approaches may vary widely across settings, which remains underexplored in the literature. In this work, we demonstrate theoretically and empirically that CFD yields lower-variance estimates than SBD when the true effect of treatment assignment is small (*i.e.*, assigning an intervention leads to small changes in patients' future outcome). Additionally, since CFD requires estimating multiple nuisance parameters, we introduce LobsterNet, a multi-task neural network that implements CFD with joint modeling of the nuisance parameters. Empirically, LobsterNet reduces estimation error across several semi-synthetic and real-world datasets compared to baselines even when the true effect of treatment assignment is large. Our findings suggest CFD with shared nuisance parameter modeling can improve treatment assignment effect estimation under non-adherence.

Data and Code Availability This paper uses IHDP Hill (2011) and AMR-UTI Oberst et al. (2020) datasets. Our training and evaluation codes are available in https://github.com/MLD3/LobsterNet.

Institutional Review Board (IRB) This work is not regulated as human subjects research since data are de-identified and publicly available.

1. Introduction

We study the estimation of heterogeneous treatment assignment effects under treatment non-adherence from observational data. Treatment non-adherence arises when some individuals' treatment intake does not match their treatment assignment (*e.g.*, patients fail to take their prescribed medication). This can occur in many real-world applications, such as public health (Tressler and Bhandari, 2019) and online advertising (Gordon et al., 2019).

In many settings, decision-makers (*e.g.*, clinicians) can only intervene by assigning treatment and cannot enforce adherence. To facilitate decisions about individual treatment assignments, decision-makers must estimate the heterogeneous treatment *assignment* effect, which accounts for both the adherence probability and standard treatment effect. We refer to this quantity as the *conditional average treatment effect of the assignment (CATEA)*, which corresponds to the effect of treatment *assignment* on the outcome of interest (*e.g.*, disease progression)[1].

To estimate CATEAs, one can use the standard backdoor adjustment (SBA) (Montori and Guyatt, 2001; Gupta, 2011), which estimates potential outcomes while controlling for confounders. However, SBD ignores treatment intake information, which is often available. We propose to leverage conditional front-door adjustment (CFD) (Pearl, 2009; Xu et al., 2024) as an alternative for estimating the CATEA that leverages treatment intake. CFD works by combining estimates of three *nuisance parameters:* (1) the probability of treatment assignment, (2) the probability of treatment intake given assignment, and (3) the outcome given treatment assignment and intake via an adjustment formula to form CATEA estimates.

---

1. Our estimand differs from the standard conditional average treatment effect (CATE), *i.e.*, the effect of adhering to the assigned treatment.
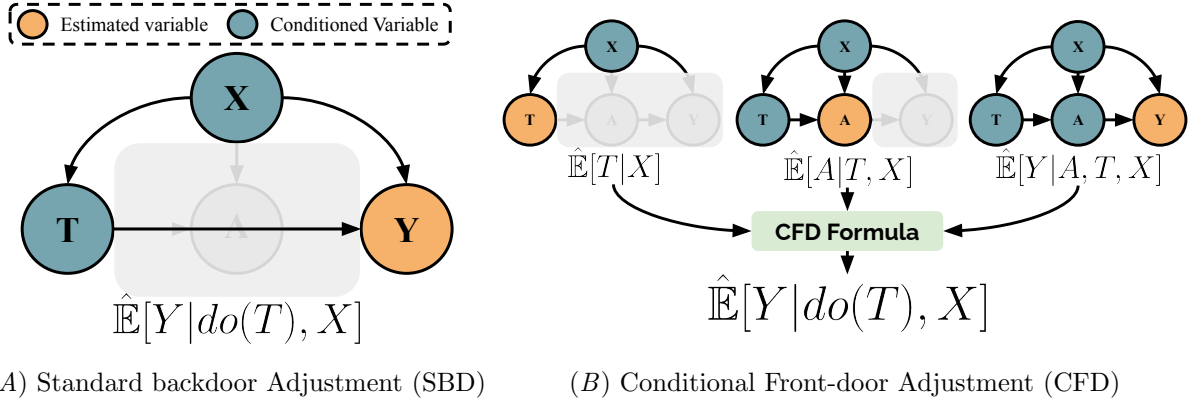
(A) Standard backdoor Adjustment (SBD) (B) Conditional Front-door Adjustment (CFD)

Figure 1: Overview of SBD and CFD in causal graphs. Variables X, T, A, and Y indicate confounders, treatment assignment, intake, and outcome. (A): SBD directly estimates the potential outcomes. (B): CFD first estimates three types of nuisance parameters and recovers the potential outcomes by combining the estimates via the CFD formula.

The error associated with CATEA estimates decomposes into bias and variance (Domingos, 2000). In this work, we first prove that given sufficient identification assumptions, both CFD and SBD produce unbiased CATEA estimates (Section 3). Thus, one should choose the estimator with the lower variance. However, whether CFD or SBD has lower variance is underexplored in the literature (see Section 2). Intuitively, using treatment intake information in CFD should reduce variance. However, since CFD estimates more nuisance parameters, variance from each estimate could inflate the overall variance. Due to this tension, it is unclear when CFD will yield lower-variance estimates than SBD.

Thus, we compare the asymptotic variance of CFD to that of SBD and show that CFD has lower variance than SBD when the true CATEA is small. Small CATEA is common in real-world treatment assignments (Ioannidis, 2005; Gu et al., 2016; Vivalt, 2020), highlighting the practical benefits of CFD.

Building on our theoretical analysis, we propose LobsterNet, a multi-task neural network architecture that jointly estimates nuisance parameters required by CFD. In particular, LobsterNet leverages shared representations between nuisance parameters to improve CFD's performance in finite sample settings. In summary, our contributions are as follows:

- to the best of our knowledge, we are the first to propose CFD for CATEA estimation and prove its unbiasedness under sufficient assumptions,

- we compare the asymptotic variance of CFD to SBD in CATEA estimation, showing that CFD improves variance under small CATEA,

- we propose LobsterNet, a novel CFD implementation with joint nuisance parameter modeling, improving the empirical performance of CFD compared to existing implementations.

2. Related Work

Under non-adherence, the effect of treatment intake on the outcome, known as the conditional average treatment effect (CATE), is often estimated to understand the impact of treatment intake (*e.g.*, a drug's effectiveness) (Marcus and Gibbons, 2001; Nagelkerke et al., 2000; Bargagli-Stoffi et al., 2022; Miladinovic et al., 2011). While CATE estimation is important, estimating the treatment assignment effect (*i.e.*, CATEA) could be independently useful in scenarios where decision-makers can only assign treatments but cannot enforce treatment intake (*e.g.*, encouragement designs (Hirano et al., 2000)).

In this work, we are interested in using observational data to estimate CATEAs (Rahier et al., 2021), which account for both the individual's adherence probability and the effect of the treatment. Knowledge of CATEA can help optimize decisions when assigning treatments. For example, if an individual is unlikely to adhere to a prescription, the overall effect of *assigning* the drug would remain small even for a highly effective drug.

To estimate the CATEA, prior work has applied SBD (Montori and Guyatt, 2001; Gupta, 2011). How-

ever, such approaches ignore the treatment intake, an intermediate variable that can improve the precision of effect estimates. Such variables have been used outside the treatment non-adherence setting (Fisher, 1928; Cox, 1960). Most recently, Rahier et al. (2021) demonstrated that combining CATE estimates with estimated intake probability results in better CATEA estimates than SBD. However, their approach does not leverage CFD and only applies to one-sided non-adherence (defined in Section 3.2) with randomized control trial (RCT) data. In contrast, we leverage the CFD formula to enable CATEA estimation with both one-sided and two-sided non-adherence using observational data.

The CFD formula provides a flexible approach to combine an intermediate variable in causal effect estimation (Pearl, 2009; Xu et al., 2024). It has previously been proposed as an alternative to SBD in scenarios outside of treatment non-adherence (Hill et al., 2015; Bellemare et al., 2020). However, the relative performance of SBD and CFD is poorly understood. Glynn and Kashin (2018) studied their estimation bias subject to unobserved confounding and demonstrated that CFD has a lower bias in a broad class of applications than SBD. Similar to our work, Gupta et al. (2021) and Ramsahai (2012) compared SBD and CFD when both approaches are unbiased. Their findings suggest that under linear Gaussian data generation processes (DGPs), either approach could outperform the other in estimation variance. In contrast, we compare the estimation variance of SBD and CFD without assuming a specific DGP and characterize scenarios where CFD outperforms SBD.

Beyond ignoring the treatment intake, SBD is also limited by sample inefficiency and sensitivity to model misspecification. As a result, alternative methods such as the DR-learner (Kennedy, 2023), targeted maximum likelihood estimation (TMLE) (Van Der Laan and Rubin, 2006), and double machine learning (DML) (Chernozhukov et al., 2017) have been proposed to augment SBD with more robust and efficient estimators. However, we focus on comparing CFD and SBD implemented with standard supervised learning. Efficient estimators could potentially improve both CFD and SBD, but we aim to first understand how the two approaches perform with standard estimators.

3. Background

Here, we discuss preliminaries for CATEA estimation under treatment non-adherence.

3.1. Problem setup

Consider an observational dataset of i.i.d. samples $\{(x_i, t_i, a_i, y_i)\}_{i=1}^n$ drawn from a joint distribution $\mathcal{P}(X, T, A, Y)$ described by the causal model in Figure 1. X, T, A, and Y denote confounders, binary treatment assignment, binary treatment intake, and the outcome of interest, respectively.

For every individual i, we define $A(t_i, x_i)$ as their potential treatment intake under different treatment assignments, $t_i \in \{0, 1\}$, and $Y(a_i, x_i)$ as their potential outcome under different treatment, intake $a_i \in \{0, 1\}$. The fundamental problem of estimating the treatment assignment effect under non-adherence is that we only observe the outcome under one of the treatment assignment and intake combinations for each individual. The observed outcome can be expressed as $y_i = Y(A(t_i, x_i), x_i)$. We aim to estimate CATEA, $\Phi(x_i)$, defined as:

$$\Phi(x_i) := \mathbb{E}[Y|do(T = 1), x_i] - \mathbb{E}[Y|do(T = 0), x_i]$$

3.2. Non-adherence settings

We consider two standard non-adherence settings (Gordon et al., 2019): one-sided and two-sided non-adherence. Under one-sided non-adherence, only individuals assigned the treatment ($t_i = 1$) can be non-adherent ($a_i \neq t_i$). Individuals not assigned the treatment ($t_i = 0$) always adherer; *i.e.*, $A(t_i = 0, x_i) = 0 \; \forall \; i$. Under two-sided non-adherence, treatment intake (A) may deviate arbitrarily from the treatment assignment (T). In this paper, we use $\Phi_{\text{os}}(\cdot)$ and $\Phi_{\text{ts}}(\cdot)$ to denote CATEA estimates under one-sided and two-sided non-adherence, respectively.

3.3. SBD-based approach

The SBD-based approach ignores individuals' treatment intake (A) and only considers treatment assignment (T), outcome (Y), and confounders (X). Its CATEA estimations under both one-sided and two-sided non-adherence are defined as follows:

$$\hat{\Phi}^{\text{SBD}}(x_i) = \hat{\mathbb{E}}[Y|T = 1, x_i] - \hat{\mathbb{E}}[Y|T = 0, x_i]$$

$\hat{\mathbb{E}}[\cdot]$ indicates the empirical estimate of the expectation. In practice, SBD is often implemented with meta-learners such as T-learner (Künzel et al., 2019), which estimates $\hat{\mathbb{E}}[Y|T = 0, x_i]$ and $\hat{\mathbb{E}}[Y|T = 1, x_i]$ with two separate models. When all the potential confounders between T, A, and Y are included in X, SBD provides unbiased CATEA estimation because it removes any

effect from confounders by conditioning them. (Proof of SBD's unbiasedness is in Appendix B.1)

3.4. CFD-based approach

Our proposed CFD-based approach first estimates three sets of nuisance parameters from the data:

$$\hat{\pi}(x_i) = \hat{\mathbb{E}}[T|x_i]$$
$$\hat{A}(t_i, x_i) = \hat{\mathbb{E}}[A|T = t_i, x_i]$$
$$\hat{Y}(a_i, t_i, x_i) = \hat{\mathbb{E}}[Y|A = a_i, T = t_i, x_i]$$

Each of these nuisance parameters can be estimated via standard supervised learning techniques. CATEA is then recovered by combining these estimates with the CFD formula (Pearl, 2009; Xu et al., 2024). Under one-sided non-adherence, the formula is defined as:

$$\hat{\Phi}_{os}^{CFD} = \left[\left(\hat{Y}(1,1,x_i) - \hat{Y}(0,0,x_i) \right) (1 - \hat{\pi}(x_i)) + \right.$$
$$\left. \left(\hat{Y}(1,1,x_i) - \hat{Y}(0,1,x_i) \right) \hat{\pi}(x_i) \right] \hat{A}(1, x_i)$$

The CFD formula under two-sided non-adherence is defined as follows:

$$\hat{\Phi}_{ts}^{CFD}(x_i) = \left[\left(\hat{Y}(1,0,x_i) - \hat{Y}(0,0,x_i) \right) (1 - \hat{\pi}(x_i)) + \right.$$
$$\left. \left(\hat{Y}(1,1,x_i) - \hat{Y}(0,1,x_i) \right) \hat{\pi}(x_i) \right] \cdot \left[\hat{A}(1, x_i) - \hat{A}(0, x_i) \right]$$

Leveraging Pearl (2009)'s do calculus, we prove that CFD leads to unbiased CATEA estimates under both one-sided and two-sided non-adherence (details in Appendix B.2 and B.3). Intuitively, these formulas compute the effect of intake (a_i) on the outcome (y_i) under different assignments (t_i) and then combine them via a weighted sum, where the weights are defined by the probability of treatment assignment ($\hat{\pi}(\cdot)$) and the effect of treatment assignment on intake. CFD only uses treatment intake to estimate nuisance parameters; at test time, both SBD and CFD only use confounders (X). The derivations for the above formulas are in Appendix C.

3.5. Identifiability assumptions

To make the CATEA identifiable under our problem setup, we assume consistency and positivity, as well as no unobserved confounders between any pair of T, A, and Y (ignorability), which are standard assumptions in causal effect estimation (Pearl, 2009).

Additionally, we assume full mediation, a standard assumption under treatment non-adherence (Rahier et al., 2021; Gordon et al., 2019). See Appendix A for detailed definitions of these assumptions.

4. Theoretical Analysis

In this section, we explore when the CFD-based approach is guaranteed to have lower estimation variances than the SBD-based approach.

4.1. Setup

We compare the asymptotic variances of $\hat{\Phi}^{SBD}$ and $\hat{\Phi}^{CFD}$ in the single-stratum setting, which focuses on estimation for a single value $x_0 \in X$, for which we assume to observe n samples $\{(x_0, t_i, a_i, y_i)\}_{1 \le i \le n}$. We simplify our notation as follows:

$$A_{t=0} \sim P(a_i|t_i = 0, x_0) \quad A_{t=1} \sim P(a_i|t_i = 1, x_0)$$
$$Y_{a=0} \sim P(y_i|a_i = 0, x_0) \quad Y_{a=1} \sim P(y_i|a_i = 1, x_0)$$

$\mathbb{E}(\cdot)$ and $\mathbb{V}(\cdot)$ denote the expectation and variance of the corresponding variable. We also define Δ_A and Δ_Y as the effect of assignment on intake and the effect of intake on the outcome. They are decompositions of the CATEA (i.e., $\Phi(x_o) = \Delta_A \Delta_Y$):

$$\Delta_A = \mathbb{E}(A_{t=1}) - \mathbb{E}(A_{t=0})$$
$$\Delta_Y = \mathbb{E}(Y_{a=1}) - \mathbb{E}(Y_{a=0})$$

Because CFD requires estimating outcomes conditioned on each assignment and intake combination, we define ρ as the minimum probability that the individual experiences one of the possible combinations. Formally, ρ is defined as below:

$$\rho = \begin{cases} \min_{\substack{t',a' \in \{0,1\} \\ (a',t') \neq (1,0)}} \{P(a', t' \mid x_0)\} & \text{if one-sided} \\ & \text{non-adherence} \\ \min_{t',a' \in \{0,1\}} \{P(a', t' \mid x_0)\} & \text{if two-sided} \\ & \text{non-adherence} \end{cases}$$

Later, ρ will help simplify the bound on the variance of CFD by representing the fewest observations among all possible treatment assignment/intake combinations.

4.2. Asymptotic Variance Bounds

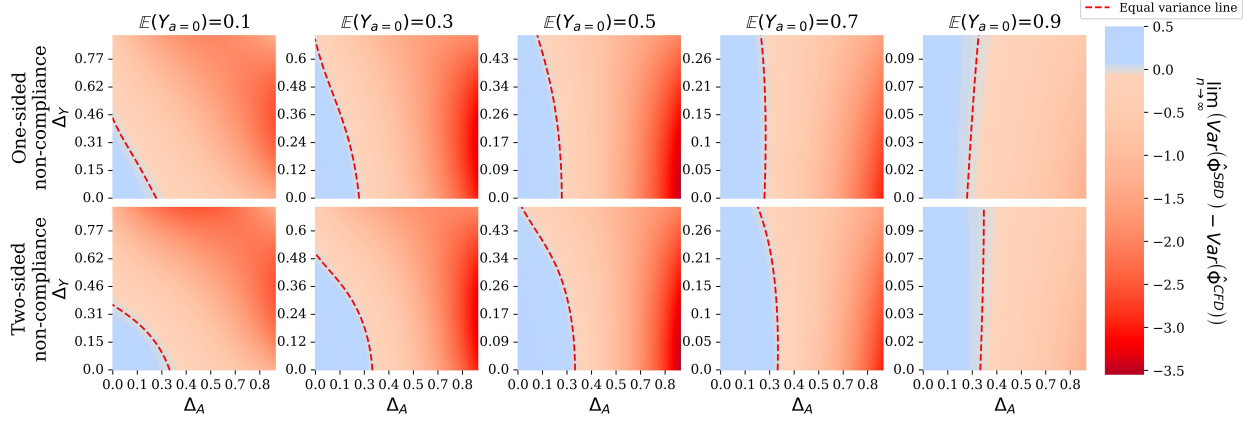Here, we present asymptotic variance *lower* bounds for SBD-based CATEA estimation and *upper* bounds

Figure 2: Numerical visualization of lower bounds on CFD's asymptotic variance reduction over SBD. More blue colors indicate greater variance reduction by CFD. The red dotted line indicates contours where the reduction equals zero. CFD is guaranteed to reduce variance when the effect size (either Δ_A or Δ_Y) is small.

for CFD-based CATEA estimation. Our derivation follows the analysis framework proposed by Rahier et al. (2021), and are described in Appendix D.3.

Proposition 1 *Under one-sided and two-sided non-adherence, the asymptotic variance of SBD-based CATEA estimation is lower-bounded by:*

$$\lim_{n \to \infty} n\mathrm{Var}\left(\hat{\Phi}_{\mathrm{os}}^{\mathrm{SBD}}\right) > \mathbb{V}\left(Y_{a=0}\right)\left(2 - \mathbb{E}\left(A_{t=1}\right)\right) +$$

$$\mathbb{V}\left(Y_{a=1}\right)\mathbb{E}\left(A_{t=1}\right) + \mathbb{V}\left(A_{t=1}\right)\left(\Delta_Y\right)^2$$

$$\lim_{n \to \infty} n\mathbb{V}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{SBD}}) > \mathbb{V}\left(Y_{a=0}\right)\left(2 - \mathbb{E}\left(A_{t=0}\right) - \mathbb{E}\left(A_{t=1}\right)\right)$$

$$+ \mathbb{V}\left(Y_{a=1}\right)\left(\mathbb{E}\left(A_{t=0}\right) + \mathbb{E}\left(A_{t=1}\right)\right)$$

$$+ \left(\mathbb{V}\left(A_{t=0}\right) + \mathbb{V}\left(A_{t=1}\right)\right)\left(\Delta_Y\right)^2.$$

Proposition 2 *Under one-sided and two-sided non-adherence, the asymptotic variance of CFD-based CATEA estimation is upper-bounded by:*

$$\lim_{n \to \infty} n\mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{CFD}}) < \frac{1}{\rho}\Bigg(\left(\mathbb{V}\left(Y_{a=0}\right) + \mathbb{V}\left(Y_{a=1}\right)\right)\Delta_A^2$$

$$+ \mathbb{V}\left(A_{t=1}\right)\Delta_Y^2\Bigg)$$

$$\lim_{n \to \infty} n\mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{CFD}}) < \frac{1}{\rho}\Bigg(\left(\mathbb{V}\left(Y_{a=0}\right) + \mathbb{V}\left(Y_{a=1}\right)\right)\Delta_A^2$$

$$+ \left(\mathbb{V}\left(A_{t=0}\right) + \mathbb{V}\left(A_{t=1}\right)\right)\Delta_Y^2\Bigg).$$

4.3. Analytical Comparison

From the upper bounds on CFD's variance and lower bounds on SBD's variance we can obtain a *lower* bound on CFD's variance reduction over SBD (*i.e.,* $\lim_{n \to \infty}\left(n\mathrm{Var}\left(\Phi^{\mathrm{SBD}}\right) - n\mathrm{Var}\left(\Phi^{\mathrm{CFD}}\right)\right)$).

To simplify this comparison, we assume that each potential outcome or intake shares the same variance, which we denote with $\mathrm{V_Y}$ and $\mathrm{V_A}$ respectively:

$$\mathbb{V}\left(Y_{a=0}\right) = \mathbb{V}\left(Y_{a=1}\right) = \mathrm{V_Y},$$

$$\mathbb{V}\left(A_{t=0}\right) = \mathbb{V}\left(A_{t=1}\right) = \mathrm{V_A}.$$

This is reasonable because real-world potential outcomes often share the same generation processes, leading to similar variances (Curth and Van der Schaar, 2021). Then, we obtain the following simplified variance reduction lower-bound (details in Appendix D.4):

Proposition 3 *CFD's asymptotic variance reduction over SBD is lower-bounded by:*

$$\lim_{n \to \infty} n\left(\mathrm{Var}(\hat{\Phi}^{\mathrm{SBD}}) - \mathrm{Var}(\hat{\Phi}^{\mathrm{CFD}})\right) >$$

$$\mathcal{O}\left(V_Y - \frac{V_Y\Delta_A^2}{\rho} - \frac{(1-\rho)V_A\Delta_Y^2}{\rho}\right).$$

This variance reduction bound applies to both one-sided and two-sided non-adherence. Intuitively, **decreasing either Δ_A^2 and Δ_Y^2 will increase the overall bound, giving CFD a greater advantage over SBD**. Additionally, ρ also influences the bound. Because $0 < \rho < 1$, **decreasing ρ will reduce the overall bound, making CFD less advantageous**.

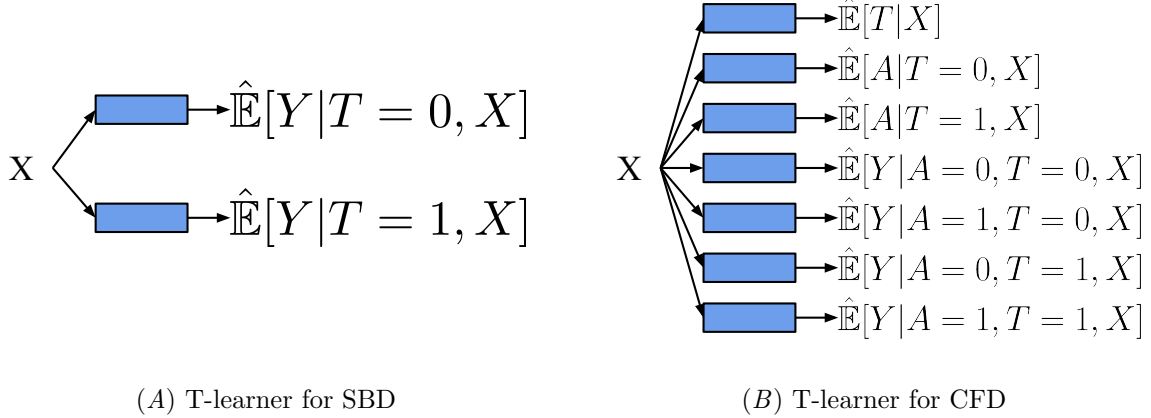(A) T-learner for SBD

(B) T-learner for CFD

Figure 3: Illustrations of the T-learner method for implementing both SBD and CFD. Because CFD requires estimating more nuisance parameters than SBD, we extend the T-learner by introducing additional independent models.

4.4. Numerical Comparison

In Figure 2, we visualize the variance reduction lower bound in settings where variance may not be equal between potential outcomes and inputs. Without loss of generality, we consider that $\Delta_A, \Delta_Y \geq 0$ and set $\rho = 0.1$. $\mathbb{E}(Y_{a=0})$ is set to five different values between 0 and 1. Under one-sided non-adherence, $\mathbb{E}(A_{a=0})$ is zero by definition, and under two-sided non-adherence, $\mathbb{E}(A_{a=0})$ is set to 0.1.

The smaller the reduction lower bound (more blue in Figure 2), the larger is the gap between the CFD variance versus the SBD variance in favor of CFD. Across all values of $\mathbb{E}(Y_{a=0})$, CFD-based estimation has increasing advantage as Δ_A and Δ_Y decrease. Intuitively, CFD enjoys an advantage over SBD when the true CATEA is small, *i.e.*, when either Δ_A or Δ_Y is small, supporting our analytical finding. As $\mathbb{E}(Y_{a=0})$ increases, the area where CFD is guaranteed to achieve lower variances also increases. This is because the feasible value of Δ_Y decreases as $\mathbb{E}(Y_{a=0})$ increases, reducing CATEA and giving more CFD advantages. More details about the numerical visualization can be found in Appendix E.

4.5. Limitations of CFD

Our analysis also highlights regions where CFD does not guarantee a reduced variance compared to SBD. For example, the advantage of CFD over SBD in terms of asymptotic variance diminishes when ρ (the minimum probability for individuals to experience one of the possible assignment and intake combina-

tions) is small, as shown via an analytical comparison (Section 4.3). Furthermore, Figure 2 assumes a conservative ρ value (0.1), making it hard to guarantee CFD's advantage under many settings. While lower CFD variance can be achieved with greater ρ, this requires individuals having similar probabilities for experiencing different assignment and intake combinations, which may not hold in practice (Khan et al., 2024).

To mitigate CFD's need for having sufficient samples across all assignment and intake combinations, we draw inspiration from the hypothesis that the true potential outcome functions may have shared structures in practice (Curth and Van der Schaar, 2021). In Section 5, we introduce a multi-task neural network architecture designed to leverage shared representations when learning CFD's nuisance parameters. When nuisance parameters are learned via shared representations, ρ is virtually increased because every sample contributes to estimating nuisance parameters conditioned on each assignment and intake combination.

5. Implementing CFD in Practice

We first discuss implementing CFD by extending the T-learner, an existing method for implementing SBD. Then, we introduce LobsterNet, a multi-task neural network architecture tailored for CFD-based CATEA estimation.
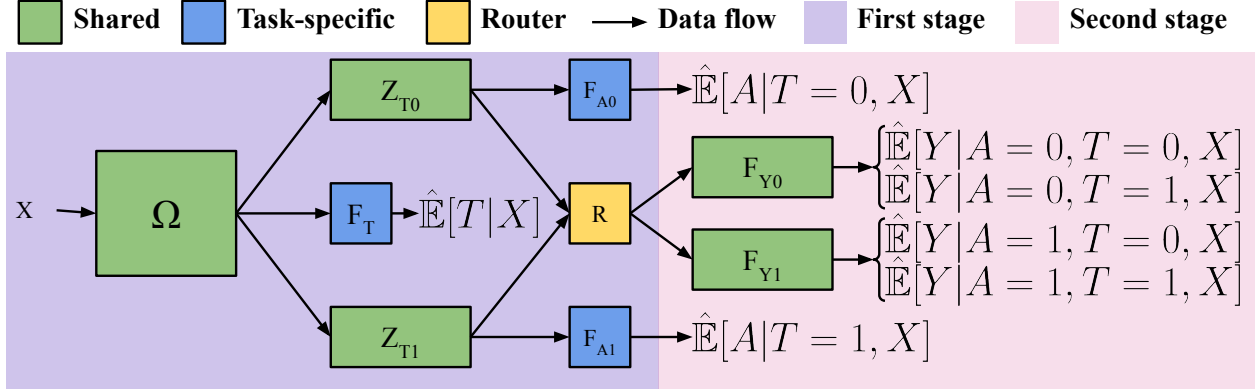
Figure 4: Illustration of our proposed LobsterNet architecture,which jointly models all required nuisance parameters

5.1. CFD with T-learner

T-learner (short for Two-model learner) is a meta-algorithm commonly used to implement SBD (Künzel et al., 2019). It leverages two independent models to estimate potential outcomes required by SBD (Figure 3(A)). We implement CFD by extending the T-learner. Because CFD requires estimating more than two nuisance parameters, we introduce additional models to independently estimate all nuisance parameters (Figure 3(B)). We provide implementation details of CFD with T-learner in Appendix Section F.1. Despite its popularity, the T-learner may be sample inefficient because it does not leverage potential similarities between estimates. When applied to implement CFD, the T-learner approach can have high variance when certain nuisance parameters have a small number of observed samples. Thus, we propose a new approach to overcome this challenge by estimating the nuisance parameters jointly.

5.2. CFD with LobsterNet

We propose LobsterNet to simultaneously estimate all nuisance parameters required for CFD via shared representations and a joint training objective. LobsterNet extends DragonNet (Shi et al., 2019) by introducing a two-stage structure: stage one models treatment assignment and intake, and stage two leverages the representations learned in stage one to predict potential outcomes (Figure 4). However, in contrast to DragonNet, our proposed approach enables sharing across all CFD-required estimation tasks.

The first stage starts with a backbone network, $\Omega(\cdot)$, which produces a shared representation to predict treatment assignment and conditional treatment intake via networks: $F_T(\cdot)$, $F_{A0}(Z_{T0}(\cdot))$, and $F_{A1}(Z_{T1}(\cdot))$. In the second stage, LobsterNet predicts outcomes conditioned on both treatment assignment and intake. The conditioning on intake is achieved by running predictions with separate networks, $F_{Y0}(\cdot)$ and $F_{Y1}(\cdot)$. To condition on different assignments, LobsterNet leverages a router module, $R(\cdot)$, which selects output from either $Z_{T0}(\cdot)$ or $Z_{T1}(\cdot)$ as the input to $F_{Y0}(\cdot)$ and $F_{Y1}(\cdot)$.

Let $\hat{T}(x)$, $\hat{A}(t,x)$ and $\hat{Y}(t,a,x)$ denote LobsterNet's estimates of treatment assignment, intake, and outcome, respectively. We optimize it by minimizing:

$$\mathcal{L}(\hat{Y}, \hat{A}, \hat{T}) = \frac{1}{N} \sum_i^N [\mathcal{L}_Y(\hat{Y}(t_i, a_i, x_i), y_i) +$$

$$\alpha \, \mathcal{L}_{\mathrm{CE}}(\hat{A}(t_i, x_i), a_i) + \beta \, \mathcal{L}_{\mathrm{CE}}(\hat{T}(x_i), t_i)]$$

where $\alpha, \beta \in \mathbb{R}^+$ are hyperparameters balancing each loss component. $\mathcal{L}_{\mathrm{CE}}$ is cross entropy loss, and $\mathcal{L}_Y$ is the outcome (y_i) prediction loss (e.g., mean squared error). During inference, LobsterNet outputs $\hat{T}(x)$, $\hat{A}(t,x)$ and $\hat{Y}(t,a,x)$, then plugs them into the CFD formula to estimate CATEA (see Section 3.4). We provide additional implementation details about LobsterNet in Appendix Section F.2.

Through joint optimization with shared neural network representations, LobsterNet leverages shared components between each nuisance parameter. We hypothesize that this sharing can lead to improved CFD-based CATEA estimation. In Section 6, we empirically validate this hypothesis using semi-synthetic and real-world datasets.

6. Experiments & Results

In this section, we empirically validate our theoretical insights that CFD reduces estimation error compared

to SBD when Δ_A and Δ_Y are small in finite sample settings. Additionally, we benchmark LobsterNet against T-learner-based CFD estimators.

6.1. Datasets

Building on our theoretical analysis, we design two synthetic datasets. **Synthetic dataset A** evaluates CATEA estimation under different Δ_A:

$$
\begin{aligned}
x_i &\sim \mathcal{N}(0, \mathbf{I}_{30\times30}), \quad t_i|x_i = \text{Bern}\left(\sigma\left(\frac{wx_i}{30}\right)\right) \\
y_i|a_i, x_i &= \text{Bern}\left(\sigma\left((1-a_i)\frac{w_{a0}x_i}{30} + a_i\frac{w_{a1}x_i}{30}\right)\right) \\
p_i^{\text{nc}}|t_i, x_i &= \frac{\sigma((1-t_i)w_{t0}x_i + t_iw_{t1}x_i)}{\sum_{i=1}^n \sigma((1-t_i)w_{t0}x_i + t_iw_{t1}x_i)} \\
&\{w, w_{t0}, w_{t1}, w_{a0}, w_{a1}\} \in \mathbb{R}^{30} \sim \mathcal{U}(-10, 10)
\end{aligned}
$$

We sample 1000 data points following the above. To simulate different Δ_A, we generate $a_i|t_i, x_i$ with various non-adherence rate γ. Lower Δ_A corresponds to $\gamma \to 100\%$ under one-sided non-adherence and $\gamma \to 50\%$ under two-sided non-adherence. a_i is generated via weighted sampling: $1000 \times \gamma$ non-adherers are selected with p_i^{nc} as sampling probability. a_i is set to $1 - t_i$ if sampled as non-adherer and t_i if otherwise.

Synthetic dataset B evaluates CATEA estimation under different Δ_Y. x_i and t_i generation follows synthetic dataset A. a_i and y_i are generated with:

$$
\begin{aligned}
a_i|t_i, x_i &= \text{Bern}\left(\sigma\left((1-t_i)\frac{w_{t0}x_i}{30} + t_i\frac{w_{t1}x_i}{30}\right)\right) \\
y_i|a_i, x_i &= \text{Bern}\left((1-a_i)0.1 + a_i\sigma\left(\frac{w_{a1}x_i}{30} + \eta\right)\right)
\end{aligned}
$$

Δ_Y is controlled via $\eta \in \mathbb{R}$, the average effect of treatment intake on the log-odds of y_i, with higher η giving higher average Δ_Y. w_{t0}, w_{t1}, and w_{a1} are sampled identically to the synthetic dataset A, except that under one-sided non-adherence, w_{t0} is zero.

Next, we consider the **IHDP** (Infant Health and Development Program) dataset ($n = 747$) (Hill, 2011), a semi-synthetic dataset created from an experiment that studied the effect of high-quality childcare and home visits on future cognitive test scores. It contains 25 features (X) and a binary treatment assignment (T). We simulate treatment intake (A) identically as the synthetic datasets and adopt the same outcome simulation function proposed by Hill (2011), except that we replace T with A as the conditioning variable.

To evaluate our approach with real potential outcomes, we also examine **AMR-UTI** ($n = 15,806$),

a dataset containing antimicrobial resistance for patients with urinary tract infections (UTIs) (Oberst et al., 2020). AMR-UTI covers four types of antimicrobial treatments (T) and uniquely provides ground-truth antimicrobial resistance labels for each treatment obtained via microbiology testing, such that the labels are representative of real-world counterfactual outcomes. We adapt AMR-UTI to one-sided and two-sided non-adherence. Under one-sided non-adherence, the treatment assignment indicates whether Nitrofurantoin was prescribed. Only patients prescribed Nitrofurantoin can be non-adherent. Under two-sided non-adherence, treatment assignment represents a prescription for one of two antimicrobial treatments (Nitrofurantoin and Ciprofloxacin). In this toy setup, patients prescribed either treatment may not adhere to their treatment by taking the other, unassigned treatment. In both settings, we simulate non-adherence behavior identically to the synthetic datasets. We define whether a patient recovers from the UTI as the outcome of interest (Y). Only patients infected with a pathogen susceptible (*i.e.*, non-resistant) to the antimicrobial treatment actually taken will recover. The goal is to estimate the individual-level difference in recovery rates between the two treatment assignments.

For IHDP and AMR-UTI, we do not vary Δ_Y because the potential outcome generation processes are pre-specified. Δ_A is controlled identically to the synthetic datasets. For all datasets, we evaluate the quality of CATEA estimation using the precision in the estimation of heterogeneous effect (PEHE) metric (Hill, 2011).

6.2. Implementation details

We use neural networks to implement CFD and SBD. For each dataset, we perform an 80:20 train-test split. We perform early stopping on a validation split sampled from 20% of the training split. To quantify uncertainty, we repeat data generation and model training 20 times and report the median and inter-quartile range (IQR) of the results. Full implementation details are in Appendix F.

6.3. Results on Synthetic Datasets

Overall, CFD achieves lower PEHE than SBD on both synthetic datasets A and B Figure 5 directly compares CFD and SBD on the synthetic datasets, with both approaches implemented via T-learners (Künzel et al., 2019).

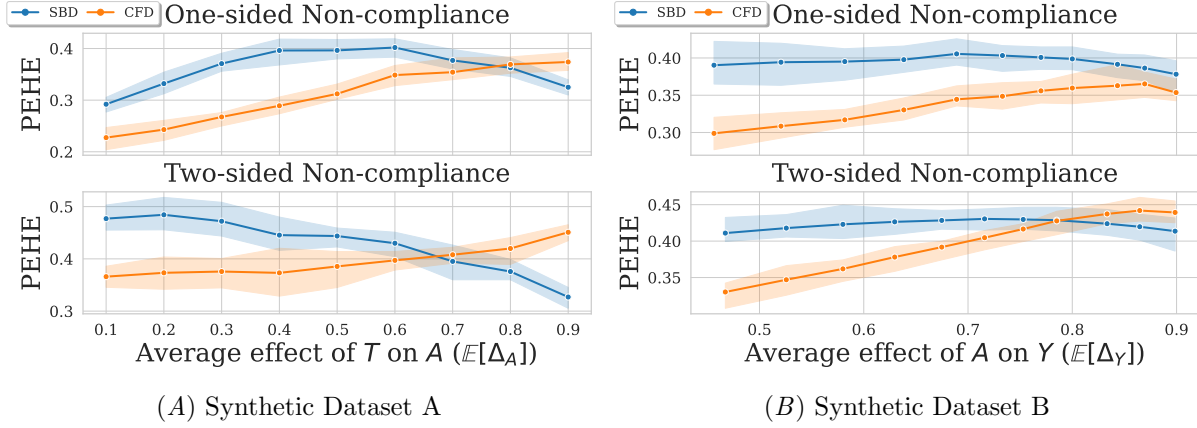(A) Synthetic Dataset A   (B) Synthetic Dataset B

Figure 5: CATEA estimation PEHE achieved by SBD and CFD on synthetic datasets. CFD outperforms SBD in most settings. The overall trend of CFD's performance improvement over SBD also supports the theoretical result.
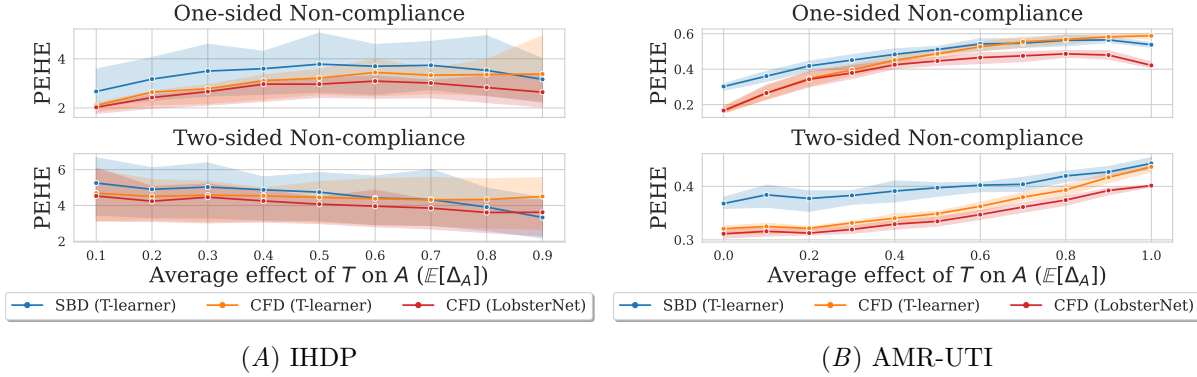

(A) IHDP   (B) AMR-UTI

Figure 6: CATEA estimation PEHE achieved by SBD, CFD, and LobsterNet on IHDP and AMR-UTI datasets. CFD-based methods overall outperform SBD, and LobsterNet yields additional improvements over baseline CFD.

With synthetic dataset A (Figure 5(A)), CFD achieves lower average PEHE than SBD under both one-sided (CFD: 0.309, 95%CI: 0.300-0.318 vs. SBD: 0.362, 95%CI: 0.355-0.368) and two-sided non-adherence (CFD: 0.394, 95%CI: 0.336-0.345 vs. SBD: 0.428, 95%CI: 0.422-0.435) across all levels of average Δ_A ($\mathbb{E}[\Delta_A]$). Similarly, with synthetic dataset B (Figure 5(B)), CFD shows an advantage over SBD in average PEHE under both one-sided (CFD: 0.341, 95%CI: 0.336-0.345 vs. SBD: 0.395, 95%CI: 0.390-0.400) and two-sided non-adherence (CFD: 0.398, 95%CI: 0.393-0.405 vs. SBD: 0.423, 95%CI: 0.419-0.426) across different average Δ_Y ($\mathbb{E}[\Delta_Y]$).

CFD's advantage increases as Δ_A decreases. As shown in Figure 5(A), CFD's performance improvement over SBD increases as the $\mathbb{E}[\Delta_A]$ decreases. This trend is consistent across both one-sided and

two-sided non-adherence, validating our theoretical finding that CFD has lower variance than SBD when Δ_A is small.

CFD's advantage inversely correlates with Δ_Y. Given our theoretical result that CFD has lower variance than SBD when Δ_Y is smaller, we expect the lift of CFD over SBD to increase as $\mathbb{E}[\Delta_Y]$ decreases. This is precisely what we observe in Figure 5(B). Under both one-sided and two-sided non-adherence, CFD's PEHE improvement over SBD increases as $\mathbb{E}[\Delta_Y]$ decreases, consistent with the theory.

6.4. Results on IHDP and AMR-UTI

CFD outperforms SBD, and LobsterNet yields additional gains across IHDP and AMR-UTI datasets. In Figure 6, we compare SBD and CFD

with LobsterNet in terms of CATEA estimation PEHE. On average, SBD (in blue) produces lower-quality CATEA estimates than either of the CFD-based approaches. Trends in relative performance between SBD and CFD in the IHDP and AMR-UTI datasets align with results in the two synthetic datasets: CFD provides greater advantages over SBD when $\mathbb{E}[\Delta_A]$ is small, validating our theoretical finding.

LobsterNet additionally improves CATEA estimation compared to CFD. In Figure 6, Lobster-Net yields a small improvement on CATEA estimation PEHE compared to CFD with a T-learner across all $\mathbb{E}[\Delta_A]$ settings. This result is consistent in both the IHDP and AMR-UTI datasets, illustrating shared representations' empirical benefit when modeling CFD nuisance parameters.

LobsterNet obtains greater improvements over CFD as $\mathbb{E}[\Delta_A]$ increases and is empirically advantageous even when $\mathbb{E}[\Delta_A]$ is large. As shown in Figure 6, across both IHDP and AMR-UTI datasets, LobsterNet's relative advantages over CFD consistently correlate with $\mathbb{E}[\Delta_A]$. This is because as $\mathbb{E}[\Delta_A]$ increases, individuals' treatment intake conditioned on treatment assignment becomes more deterministic (*i.e.*, adherence improves), leaving fewer observations for non-adherent outcomes. In such settings, Lobster-Net's shared modeling can efficiently estimate under-observed outcomes. More importantly, the additional boost from employing LobsterNet allows CFD to outperform or match SBD in PEHE across all settings, suggesting that CFD with LobsterNet is a desirable choice of CATEA estimation even when Δ_A is large.

7. Conclusion

We studied CATEA estimation in a setting without perfect treatment adherence. Specifically, we proposed CFD as a framework for incorporating individual treatment intake information. We showed that CFD-based approaches have lower asymptotic variance than conventional SBD-based approaches when the effect of treatment assignment is small. However, CFD requires estimating multiple nuisance parameters, which limits its estimation quality in settings where the distribution of treatment intake behavior is skewed. To overcome this limitation, we proposed LobsterNet, a neural network architecture that empirically surpasses existing methods due to its unified approach to modeling nuisance parameters.

Limitations. The CFD-based approach depends on treatment adherence information, which may be unavailable or unreliable (Evans and Spelman, 1983). Additionally, our work assumes binary treatments, but real-world treatments could be categorical (Lopez and Gutman, 2017), continuous (Bahadori et al., 2022), and graph-valued (Kaddour et al., 2021). Modeling non-adherence in these challenging settings is a valuable future direction.

Nonetheless, researchers aiming to estimate CATEAs in settings with non-adherence should consider utilizing CFD-based approaches. This approach is theoretically proven to be advantageous when the true CATEA is small, which is often the case in real-world treatment assignments (Ioannidis, 2005; Gu et al., 2016; Vivalt, 2020). Empirically, by leveraging our proposed LobsterNet architecture, CFD can reduce estimation error regardless of the true CATEA size. In real-world healthcare scenarios, choosing CFD with LobsterNetover the existing SBD can result in better CATEA estimate, potentially improving downstream treatment recommendations.

References

Taha Bahadori, Eric Tchetgen Tchetgen, and David Heckerman. End-to-end balancing for causal continuous treatment-effect estimation. In *International Conference on Machine Learning*, pages 1313–1326. PMLR, 2022.

Falco J Bargagli-Stoffi, Kristof De Witte, and Giorgio Gnecco. Heterogeneous causal effects with imperfect compliance: A bayesian machine learning approach. *The Annals of Applied Statistics*, 16(3):1986–2009, 2022.

Marc F Bellemare, Jeffrey R Bloem, and Noah Wexler. The paper of how: Estimating treatment effects using the front-door criterion. *Oxford Bulletin of Economics and Statistics*, 2020.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and causal parameters. Technical report, 2017.

Daniel R Cox. Regression analysis when there is prior information about supplementary variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 22(1):172–176, 1960.

Alicia Curth and Mihaela Van der Schaar. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34:15883–15894, 2021.

Pedro Domingos. A unified bias-variance decomposition. In *Proceedings of 17th international conference on machine learning*, pages 231–238. Morgan Kaufmann Stanford, 2000.

Larry Evans and Michael Spelman. The problem of non-compliance with drug therapy. *Drugs*, 25:63–76, 1983.

Ronald Aylmer Fisher. *Statistical methods for research workers*. Number 5. Oliver and Boyd, 1928.

Adam N Glynn and Konstantin Kashin. Front-door versus back-door adjustment with unmeasured confounding: Bias formulas for front-door and hybrid adjustments with application to a job training program. *Journal of the American Statistical Association*, 113(523):1040–1049, 2018.

Brett R Gordon, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. A comparison of approaches to advertising measurement: Evidence from big field experiments at facebook. *Marketing Science*, 38(2): 193–225, 2019.

Yuan-Yuan Gu, Koen Balcaen, Yicheng Ni, Jan Ampe, and Jan Goffin. Review on prevention of falls in hospital settings. *Chinese Nursing Research*, 3(1): 7–10, 2016.

Sandeep K Gupta. Intention-to-treat concept: a review. *Perspectives in clinical research*, 2(3):109–112, 2011.

Shantanu Gupta, Zachary C Lipton, and David Childers. Estimating treatment effects with observed confounders and mediators. In *Uncertainty in Artificial Intelligence*, pages 982–991. PMLR, 2021.

Daniel N Hill, Robert Moakler, Alan E Hubbard, Vadim Tsemekhman, Foster Provost, and Kiril Tsemekhman. Measuring causal impact of online actions via natural experiments: Application to display advertising. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1839–1847, 2015.

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

Keisuke Hirano, Guido W Imbens, Donald B Rubin, and Xiao-Hua Zhou. Assessing the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1(1):69–88, 2000.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal effect inference for structured treatments. *Advances in Neural Information Processing Systems*, 34:24841–24854, 2021.

Edward H Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

Jahangir Khan, Jwaad A Khan, Subhra Kumari, and Deepak Charan. Treatment non-adherence patterns among patients with mental illness: A study from the district mental health care center in india. *Cureus*, 16(2), 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2015.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10): 4156–4165, 2019.

Michael J Lopez and Roee Gutman. Estimation of causal effects with multiple treatments: a review and new ideas. *Statistical Science*, pages 432–454, 2017.

Sue M Marcus and Robert D Gibbons. Estimating the efficacy of receiving treatment in randomized clinical trials with noncompliance. *Health Services and Outcomes Research Methodology*, 2:247–258, 2001.

Branko Miladinovic, Ambuj Kumar, Iztok Hozo, and Benjamin Djulbegovic. Instrumental variable meta-analysis of individual patient data: application to adjust for treatment non-compliance. *BMC medical research methodology*, 11:1–7, 2011.

Victor M Montori and Gordon H Guyatt. Intention-to-treat principle. *Cmaj*, 165(10):1339–1341, 2001.

Nico Nagelkerke, Vaclav Fidler, Roos Bernsen, and Martien Borgdorff. Estimating treatment effects in

randomized clinical trials in the presence of non-compliance. *Statistics in medicine*, 19(14):1849–1864, 2000.

M Oberst, S Boominathan, H Zhou, S Kanjilal, and D Sontag. Amr-uti: antimicrobial resistance in urinary tract infections (version 1.0. 0), 2020.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Thibaud Rahier, Amélie Héliou, Matthieu Martin, Christophe Renaudin, and Eustache Diemert. Individual treatment prescription effect estimation in a low compliance setting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1399–1409, 2021.

Roland R Ramsahai. Supplementary variables for causal estimation. *Causality: Statistical perspectives and applications*, pages 218–233, 2012.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Stacy Tressler and Ruchi Bhandari. Interventions to increase completion of hepatitis b vaccination in people who inject drugs: a systematic review and meta-analysis. In *Open forum infectious diseases*, volume 6, page ofz521. Oxford University Press US, 2019.

Mark J Van Der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The international journal of biostatistics*, 2(1), 2006.

Eva Vivalt. How much can we generalize from impact evaluations? *Journal of the European Economic Association*, 18(6):3045–3089, 2020.

Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Kui Yu. Causal inference with conditional front-door adjustment and identifiable variational autoencoder. *International Conference on Learning Representations*, 2024.

Appendix A. Identifiability Assumptions

Assumption 1 (Consistency) *For every individual i, the potential treatment intake and potential outcome under the observed treatment assignment should equal the observed treatment intake and observed outcome, respectively: $A_i(t_i) = a_i, Y_i(A_i(t_i)) = y_i \ \forall i$.*

Assumption 2 (Positivity) *Every individual has a non-zero probability of receiving either treatment assignment: $0 < P(t_i = 1|x_i) < 1 \ \forall i$.*

Assumption 3 (Full Mediation) *The effect of treatment assignment on the outcome is fully mediated by treatment intake: $\mathbb{E}[Y|do(T = t), x_i] = \sum_{a' \in \{0,1\}} \mathbb{E}[Y|do(A = a'), x_i] P(a'|do(T = t), x_i)$.*

Assumption 4 (Assignment-Outcome Ignorability) *For every individual i, the potential outcomes given treatment assignment, $Y_i(A_i(0))$ and $Y_i(A_i(1))$, are conditionally independent of the observed treatment assignment given their pre-treatment covariates: $Y_i(A_i(0)), Y_i(A_i(1)) \perp\!\!\!\perp t_i|x_i \ \forall i$.*

Assumption 5 (Assignment-Intake Ignorability) *For every individual i, the potential treatment intakes given assignment, $A_i(0)$ and $A_i(1)$, are conditionally independent of the observed treatment assignment given their pre-treatment covariates: $A_i(0), A_i(1) \perp\!\!\!\perp t_i|x_i \ \forall i$.*

Assumption 6 (Intake-Outcome Ignorability) *For every individual i, the potential outcomes given treatment intake, $Y_i(0)$ and $Y_i(1)$, are conditionally independent of the observed treatment intake given pre-treatment covariates: $Y_i(0), Y_i(1) \perp\!\!\!\perp a_i|x_i \ \forall i$.*

Appendix B. Identifiability Proofs

B.1. Proof of SBD's identifiability

We show that $\mathbb{E}[Y|do(T = t), X]$ can be expanded as following:

$$
\begin{aligned}
&\mathbb{E}[Y|do(T = t), X = x_i] \\
&\overset{(1)}{=} \sum_{a' \in \{0,1\}} \mathbb{E}[Y|A = a', do(T = t), X = x_i] P(A = a'|do(T = t), X = x_i) \\
&\overset{(2)}{=} \sum_{a' \in \{0,1\}} \mathbb{E}[Y|A = a', do(T = t), X = x_i] P(A = a'|T = t, X = x_i) \\
&\overset{(3)}{=} \sum_{a' \in \{0,1\}} \mathbb{E}[Y|A = a', T = t, X = x_i] P(A = a'|T = t, X = x_i) \\
&\overset{(4)}{=} \mathbb{E}[Y|T = t, X = x_i]
\end{aligned}
$$

Equality (1) factorizes the conditional expectation. Equality (2) applies assignment-intake ignorability (Assumption 5). Equality (3) applies the assignment-outcome ignorability (Assumption 4) and the fact that any unobserved backdoors between treatment intake and outcome are blocked by conditioning on A. Equality (4) marginalizes the expectations to complete the proof.

B.2. Proof of identifiability of CFD under two-sided non-adherence

We show that $\mathbb{E}[Y|do(T = t), X = x_i]$ can be expanded as following:

$$\mathbb{E}[Y|do(T = t), X = x_i]$$

$$\overset{(1)}{=} \sum_{a' \in \{0,1\}} \mathbb{E}[Y|A = a', do(T = t), X = x_i] P(A = a'|do(T = t), X = x_i)$$

$$\overset{(2)}{=} \sum_{a' \in \{0,1\}} \mathbb{E}[Y|do(A = a'), X = x_i] P(A = a'|do(T = t), X = x_i)$$

$$\overset{(3)}{=} \sum_{a' \in \{0,1\}} (\mathbb{E}[Y|do(A = a'), X] P(A = a'|T = t, X = x_i))$$

$$\overset{(4)}{=} \sum_{a' \in \{0,1\}} \left(\sum_{t' \in \{0,1\}} (\mathbb{E}[Y|T = t', do(A = a'), X = x_i] P(T = t'|do(A = a'), X = x_i)) \right. \tag{1}$$

$$\left. P(A = a'|T = t, X = x_i) \right)$$

$$\overset{(5)}{=} \sum_{a' \in \{0,1\}} \left(\sum_{t' \in \{0,1\}} (\mathbb{E}[Y|T = t', do(A = a'), X = x_i] P(T = t'|X = x_i)) \, P(A = a'|T = t, X = x_i) \right)$$

$$\overset{(6)}{=} \sum_{a' \in \{0,1\}} \left(\sum_{t' \in \{0,1\}} (\mathbb{E}[Y|T = t', A = a', X = x_i] P(T = t'|X = x_i)) \, P(A = a'|T = t, X = x_i) \right)$$

$$\overset{(7)}{=} \sum_{t', a' \in \{0,1\}} \mathbb{E}[Y|T = t', A = a', X = x_i] P(T = t'|X = x_i) P(A = a'|T = t, X = x_i)$$

Equality (1) factorizes the conditional expectation. Equality (2) applies full mediation (Assumption 3). Equality (3) is allowed because of assignment-intake ignorability (Assumption 5). Equality (4) performs another factorization of conditional expectation. Equality (5) applies the definition of $do(\cdot)$ intervention on causal DAG assumed by our problem setup (*i.e.* A does not influence T), intervening on A does not affect the probability of T. Equality (6) applies intake-outcome ignorability (Assumption 6), and any unobserved backdoor between T and Y is blocked by conditioning on T. Finally, equality (7) combines the two summations to complete the proof.

B.3. Proof of CFD's identifiability under one-sided non-adherence

For identifying $\mathbb{E}[Y|do(T = 0), X = x_i]$, we start with equality (3) in Equation 1:

$$\mathbb{E}[Y|do(T = 0), X = x_i] = \sum_{a' \in \{0,1\}} \mathbb{E}[Y|do(A = a'), X = x_i] P(A = a'|T = 0, X = x_i)$$

Applying assumptions for one-sided non-adherence in Section 3.2 (*i.e.* $P(A|T = 0, X) = 0$), we simplify:

$$\mathbb{E}[Y|do(T = 0), X = x_i] = \mathbb{E}[Y|do(A = 0), X = x_i]$$

Then, applying the same techniques as equality (4), (5), and (6) in Equation 1 and assumptions for one-sided non-adherence in Section 3.2 (*i.e.* $1 - P(A|T = 0, X = x_i) = 1$), we can get:

$$\mathbb{E}[Y|do(T = 0), X = x_i] = \sum_{t' \in \{0,1\}} \mathbb{E}[Y|A = 0, T = t', X = x_i] P(T = t'|X = x_i)$$

For identifying $\mathbb{E}[Y|do(T=1), X=x_i]$, we start with equality (4) in Equation 1.

$$
\begin{aligned}
&\mathbb{E}[Y|do(T=1), X=x_i] \\
&= \sum_{a' \in \{0,1\}} \left(\sum_{t' \in \{0,1\}} (\mathbb{E}[Y|T=t', do(A=a'), X=x_i]P(T=t'|do(A=a'), X=x_iP) \right.\\
&\qquad\qquad\qquad \left. P(A=a'|T=1, X=x_i) \right) \\
&= \sum_{t' \in \{0,1\}} \left(\mathbb{E}[Y|T=t', do(A=0), X]P(T=t'|do(A=0), X=x_i)\right) P(A=0|T=1, X=x_i) \\
&\quad + \sum_{t' \in \{0,1\}} \left(\mathbb{E}[Y|T=t', do(A=1), X=x_i]P(T=t'|do(A=1), X=x_i)\right) P(A=1|T=1, X=x_i)
\end{aligned}
$$

Applying definitions for one-sided non-adherence in Section 3.2 (*i.e.* $P(T=0|do(A=1), X)=0$), we get:

$$
\begin{aligned}
&\mathbb{E}[Y|do(T=1), X=x_i] \\
&= P(A=0|T=1, X=x_i) \sum_{t' \in \{0,1\}} \mathbb{E}[Y|T=t', do(A=0), X=x_i]p[T=t'|do(A=0), X=x_i] \\
&\quad + \mathbb{E}[Y|T=1, do(A=1), X=x_i]P(A=1|T=1, X=x_i)
\end{aligned}
$$

Finally, following the same technique as equality (5) and (6) in Equation 1, we get the final form:

$$
\begin{aligned}
\mathbb{E}[Y|do(T=1), X=x_i] = P(A=0|T=1, X=x_i) \sum_{t' \in \{0,1\}} \mathbb{E}[Y|A=0, T=t', X]P(T=t'|X=x_i) \\
+ \mathbb{E}[Y|A=1, T=1, X=x_i]P(A=1|T=1, X=x_i)
\end{aligned}
$$

Appendix C. CFD-based CATEA Estimation Derivation

Given CFD's identifiability proof under one-sided non-adherence (Appendix B.3), we have its potential outcome estimation defined as follows:

$$
\begin{aligned}
\mathbb{E}[Y|do(T=0), X=x_i] =& \mathbb{E}[Y|A=0, T=0, X=x_i]p(T=0|X=x_i)+ \\
& \mathbb{E}[Y|A=0, T=1, X=x_i]P(T=1|X=x_i) \\
\mathbb{E}[Y|do(T=1), X=x_i] =& P(A=0|T=1, X=x_i)(\mathbb{E}[Y|A=0, T=0, X=x_i]P(T=0|X=x_i)+ \\
& \mathbb{E}[Y|A=0, T=1, X=x_i]P(T=1|X=x_i))+ \\
& \mathbb{E}[Y|A=1, T=1, X=x_i]P(A=1|T=1, X=x_i)
\end{aligned}
$$

Substitute in the following simplified notations:

$$
\hat{\pi}(x_i) = \hat{\mathbb{E}}[T|x_i] \quad \hat{A}(t_i, x_i) = \hat{\mathbb{E}}[A|T=t_i, x_i] \quad \hat{Y}(a_i, t_i, x_i) = \hat{\mathbb{E}}[Y|A=a_i, T=t_i, x_i] \tag{2}
$$

We get CFD's CATEA estimation under one-sided non-adherence as:

$$
\begin{aligned}
\Phi_{\mathrm{OS}}^{\hat{\mathrm{CFD}}} =& (1 - \hat{A}(T=1, x_i)) \underbrace{\left(\hat{Y}(A=0, T=0, x_i)(1 - \hat{\pi}(x_i)) + \hat{Y}(A=0, T=1, x_i)\hat{\pi}(x_i) \right)}_{\text{same}} + \\
& \hat{Y}(A=1, T=1, x_i)\hat{A}(T=1, x_i) - \underbrace{\left(\hat{Y}(A=0, T=0, x_i)(1 - \hat{\pi}(x_i)) + \hat{Y}(A=0, T=1, x_i)\hat{\pi}(x_i) \right)}_{\text{same}} \\
=& -\underbrace{\hat{A}(T=1, x_i)}_{\text{same}} \left(\hat{Y}(A=0, T=0, x_i)(1 - \hat{\pi}(x_i)) + \hat{Y}(A=0, T=1, x_i)\hat{\pi}(x_i) \right) + \hat{Y}(A=1, T=1, x_i)\underbrace{\hat{A}(T=1, x_i)}_{\text{same}} \\
=& \left(\underbrace{\hat{Y}(A=1, T=1, x_i)}_{\text{Split into two parts}} - \hat{Y}(A=0, T=0, x_i)(1 - \hat{\pi}(x_i)) - \hat{Y}(A=0, T=1, x_i)\hat{\pi}(x_i) \right) \hat{A}(T=1, x_i) \\
=& \left((\hat{Y}(A=1, T=1, x_i) - \hat{Y}(A=0, T=0, x_i))(1 - \hat{\pi}(x_i)) + \right. \\
& \left. (\hat{Y}(A=1, T=1, x_i) - \hat{Y}(A=0, T=1, x_i))\hat{\pi}(x_i) \right) \hat{A}(T=1, x_i)
\end{aligned}
$$

Given CFD's identifiability proof under two-sided non-adherence (Appendix B.2), we have its potential outcome estimation defined as follows:

$$
\begin{aligned}
\mathbb{E}[Y|do(T=t), X=x_i] =& \mathbb{E}[Y|A=0, T=0, X=x_i]P(T=0|X=x_i)P(A=0|T=t, X=x_i) + \\
& \mathbb{E}[Y|A=0, T=1, X=x_i]P(T=1|X=x_i)P(A=0|T=t, X=x_i) + \\
& \mathbb{E}[Y|A=1, T=0, X=x_i]P(T=0|X=x_i)P(A=1|T=t, X=x_i) + \\
& \mathbb{E}[Y|A=1, T=1, X=x_i]P(T=1|X=x_i)P(A=1|T=t, X=x_i)
\end{aligned}
$$

Substituting in notations from Equation 2, we get CFD's CATEA estimation under two-sided non-adherence as:

$$
\begin{aligned}
\Phi_{\mathrm{TS}}^{\hat{\mathrm{CFD}}} =& \hat{Y}(A=0, T=0, x_i)(1 - \hat{\pi}(x_i))\left(1 - \hat{A}(T=1, x_i)\right) + \hat{Y}(A=0, T=1, x_i)\hat{\pi}(x_i)\left(1 - \hat{A}(T=1, x_i)\right) + \\
& \hat{Y}(A=1, T=0, x_i)(1 - \hat{\pi}(x_i))\hat{A}(T=1, x_i) + \hat{Y}(A=1, T=1, x_i)\hat{\pi}(x_i)\hat{A}(T=1, x_i) - \\
& \hat{Y}(A=0, T=0, x_i)(1 - \hat{\pi}(x_i))\left(1 - \hat{A}(T=0, x_i)\right) - \hat{Y}(A=0, T=1, x_i)\hat{\pi}(x_i)\left(1 - \hat{A}(T=0, x_i)\right) - \\
& \hat{Y}(A=1, T=0, x_i)(1 - \hat{\pi}(x_i))\hat{A}(T=0, x_i) - \hat{Y}(A=1, T=0, x_i)\hat{\pi}(x_i)\hat{A}(T=0, x_i) \\
=& -\hat{Y}(A=0, T=0, x_i)(1 - \hat{\pi}(x_i))\left(\hat{A}(T=1, x_i) - \hat{A}(T=0, x_i)\right) \\
& -\hat{Y}(A=0, T=1, x_i)\hat{\pi}(x_i)\left(\hat{A}(T=1, x_i) - \hat{A}(T=0, x_i)\right) \\
& +\hat{Y}(A=1, T=0, x_i)(1 - \hat{\pi}(x_i))\left(\hat{A}(T=1, x_i) - \hat{A}(T=0, x_i)\right) \\
& +\hat{Y}(A=1, T=1, x_i)\hat{\pi}(x_i)\left(\hat{A}(T=1, x_i) - \hat{A}(T=0, x_i)\right) \\
=& \left(\hat{Y}(A=1, T=0, x_i)(1 - \hat{\pi}(x_i)) + \hat{Y}(A=1, T=1, x_i)\hat{\pi}(x_i) \right. \\
& \left. -\hat{Y}(A=0, T=0, x_i)(1 - \hat{\pi}(x_i)) - \hat{Y}(A=0, T=1, x_i)\hat{\pi}(x_i) \right) \left(\hat{A}(T=1, x_i) - \hat{A}(T=0, x_i) \right) \\
=& \left(\left(\hat{Y}(A=1, T=0, x_i) - \hat{Y}(A=0, T=0, x_i) \right)(1 - \hat{\pi}(x_i)) + \right. \\
& \left. \left(\hat{Y}(A=1, T=1, x_i) - \hat{Y}(A=0, T=1, x_i) \right)\hat{\pi}(x_i) \right) \left(\hat{A}(T=1, x_i) - \hat{A}(T=0, x_i) \right)
\end{aligned}
$$

Appendix D. Asymptotic Variance Derivation

We derive and compare the asymptotic variances of two estimators: 1) the standard backdoor adjustment (SBD) estimator that only considers treatment (T) and 2) the conditional front-door adjustment (CFD) estimator that considers both treatment (T) and action (A). The variance bounds for SBD under one-sided non-adherence with binary outcomes (Y) are first derived in Rahier et al. (2021); We follow their framework to develop variance bounds for SBD and CFD under both one-sided and two-sided non-adherence without assuming binary outcomes.

Our variance results are derived under:

Single-stratum setting. We focus on CATEA estimation for a single value $x_0 \in X$, for which we assume to observe n *i.i.d.* samples $\{(x_0, t_i, a_i, y_i)\}_{1 \le i \le n}$.

Full ignorability setting. Because SBD and CFD require slightly different ignorability assumptions, for a fair comparison, we assume that all ignorability assumptions required by SBD and CFD are met. In order words, we assume Assumption 4, 5, and 6 hold.

Notations. We first define compact notation for the propensity score

$$\pi = \mathbb{E}[t|x_0] \tag{3}$$

Then, we define the probability of four different treatment assignments and intake combinations (or non-adherence behaviors):

$$
\begin{aligned}
\omega_{a=0,t=0} = P(a=0, t=0|x_0) \qquad & \omega_{a=1,t=0} = P(a=1, t=0|x_0) \\
\omega_{a=0,t=1} = P(a=0, t=1|x_0) \qquad & \omega_{a=1,t=1} = P(a=1, t=1|x_0)
\end{aligned}
\tag{4}
$$

Next, we define compact notations for treatment intakes generated under different treatment assignments:

$$A_{t=0} \sim P(a|t=0, x_0) \qquad\qquad A_{t=1} \sim P(a|t=1, x_0)$$

We also define compact notations for outcomes generated under different treatment assignment or intakes.

$$
\begin{aligned}
Y_{t=0} \sim P(y|t=0, x_0) \qquad\qquad & Y_{t=1} \sim P(y|t=1, x_0) \\
Y_{a=0} \sim P(y|a=0, x_0) \qquad\qquad & Y_{a=1} \sim P(y|1=1, x_0)
\end{aligned}
$$

Finally, we use $\mathbb{E}(\cdot)$ and $\mathbb{V}(\cdot)$ denote expectation and variance of the corresponding variable. For instance:

$$\mathbb{E}(Y_{t=0}) = \mathbb{E}[y|t=0, x_0] \qquad\qquad \mathbb{V}(Y_{t=1}) = \text{Var}(y|t=0, x_0)$$

For notational simplicity, we also define Δ_A and Δ_Y as the effect of assignment on intake and the effect of intake on the outcome, which are decompositions of the CATEA (*i.e.* $\Phi(x_o) = \Delta_A \Delta_Y$):

$$\Delta_A = \mathbb{E}(A_{t=1}) - \mathbb{E}(A_{t=0}) \qquad\qquad \Delta_Y = \mathbb{E}(Y_{a=1}) - \mathbb{E}(Y_{a=0})$$

The derivation is organized in four steps:

1. Define maximum-likelihood CATEA estimators (Appendix D.1)

2. Derive variance of SBD and CFD estimators (Appendix D.2)

3. Derive asymptotic variance upper and lower bounds (Appendix D.3)

4. Compare SBD and CFD's variance bounds directly. (Appendix D.4)

D.1. Define maximum-likelihood and CATEA estimators.

Given n *i.i.d.* samples $\{(t_i, a_i, y_i)\}_{1 \leq i \leq n}$, we define maximum-likelihood (MLE) estimators of notations described above.

The MLE estimator of the propensity score (Equation 3) is:

$$\hat{\pi} = \frac{\sum\limits_{i=1}^{n} t_i}{n}$$

The MLE estimators of the four different treatment assignments and intake combinations (Equation 4) are:

$$\hat{\omega}_{a=0,t=0} = \frac{\sum\limits_{i=1}^{n} (1-a_i)(1-t_i)}{n} \qquad \hat{\omega}_{a=1,t=0} = \frac{\sum\limits_{i=1}^{n} a_i(1-t_i)}{n}$$

$$\hat{\omega}_{a=0,t=1} = \frac{\sum\limits_{i=1}^{n} (1-a_i)t_i}{n} \qquad \hat{\omega}_{a=1,t=1} = \frac{\sum\limits_{i=1}^{n} a_i t_i}{n}$$

The MLE estimators of the potential treatment intakes are defined as follows:

$$\hat{\mathbb{E}}(A_{t=0}) = \frac{\sum\limits_{i=1}^{n} (1-t_i)a_i}{\sum\limits_{i=1}^{n} (1-t_i)} \qquad \hat{\mathbb{E}}(A_{t=1}) = \frac{\sum\limits_{i=1}^{n} t_i a_i}{\sum\limits_{i=1}^{n} t_i}$$

The MLE estimators of the potential outcomes are defined as follows:

$$\hat{\mathbb{E}}(Y_{t=0}) = \frac{\sum\limits_{i=1}^{n} (1-t_i)y_i}{\sum\limits_{i=1}^{n} (1-t_i)} \qquad \hat{\mathbb{E}}(Y_{t=1}) = \frac{\sum\limits_{i=1}^{n} t_i y_i}{\sum\limits_{i=1}^{n} t_i}$$

$$\hat{\mathbb{E}}(Y_{a=0,t=0}) = \frac{\sum\limits_{i=1}^{n} (1-a_i)(1-t_i)y_i}{\sum\limits_{i=1}^{n} (1-a_i)(1-t_i)} \qquad \hat{\mathbb{E}}(Y_{a=1,t=0}) = \frac{\sum\limits_{i=1}^{n} a_i(1-t_i)y_i}{\sum\limits_{i=1}^{n} a_i(1-t_i)}$$

$$\hat{\mathbb{E}}(Y_{a=0,t=1}) = \frac{\sum\limits_{i=1}^{n} (1-a_i)t_i y_i}{\sum\limits_{i=1}^{n} (1-a_i)t_i} \qquad \hat{\mathbb{E}}(Y_{a=1,t=1}) = \frac{\sum\limits_{i=1}^{n} a_i t_i y_i}{\sum\limits_{i=1}^{n} a_i t_i}$$

Using the MLE estimator notations, we define the SBD estimator under both one-sided and two-sided non-adherence settings:

$$\hat{\Phi}^{\text{SBD}} = \hat{\mathbb{E}}(Y_{t=1}) - \hat{\mathbb{E}}(Y_{t=0}) \tag{5}$$

Under one-sided non-adherence, the CFD estimator is defined as:

$$\hat{\Phi}^{\text{CFD}}_{\text{os}} = (\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=0})(1-\hat{\pi}) - \hat{\mathbb{E}}(Y_{a=0,t=1})\hat{\pi})\hat{\mathbb{E}}(A_{t=1}) \tag{6}$$

Under two-sided non-adherence, the CFD estimator is defined as:

$$\hat{\Phi}^{\text{CFD}}_{\text{ts}} = ((\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1-\hat{\pi}) + (\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))\hat{\pi})(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0})) \tag{7}$$

D.2. Derive variance of SBD and CFD estimators

D.2.1. SBD ESTIMATOR'S VARIANCE DERIVATION DERIVATION

We start by applying the law of total variance to get:

$$\text{Var}\left(\hat{\Phi}^{\text{SBD}}\right) = \underbrace{\mathbb{E}\left[\text{Var}\left(\hat{\Phi}^{\text{SBD}} \mid \{t_k\}_k\right)\right]}_{\text{A}} + \underbrace{\text{Var}\left(\mathbb{E}\left[\hat{\Phi}^{\text{SBD}} \mid \{t_k\}_k\right]\right)}_{\text{B}}$$

where $\{t_k\}_k := \{t_1, \cdots, t_n\}$.

Computation of A

Given definition from Equation 5, term A decomposes as:

$$
\begin{aligned}
&\mathbb{E}\left[\text{Var}\left(\hat{\Phi}^{\text{SBD}} \mid \{t_k\}_k\right)\right] \\
&= \mathbb{E}\left[\underbrace{\text{Var}\left(\hat{\mathbb{E}}\left(Y_{t=1}\right) \mid \{t_k\}_k\right)}_{\text{A1}} + \underbrace{\text{Var}\left(\hat{\mathbb{E}}\left(Y_{t=0}\right) \mid \{t_k\}_k\right)}_{\text{A2}} \underbrace{-2\,\text{Cov}\left(\hat{\mathbb{E}}\left(Y_{t=1}\right), \hat{\mathbb{E}}\left(Y_{t=0}\right) \mid \{t_k\}_k\right)}_{\text{A3}} \right]
\end{aligned}
\tag{8}
$$

For A1, we write:

$$
\begin{aligned}
\text{Var}\left(\hat{\mathbb{E}}\left(Y_{t=1}\right) \mid \{t_k\}_k\right) &= \text{Var}\left(\underbrace{\frac{\sum_{i=1}^{n} t_i y_i}{\sum_{i=1}^{n} t_i}}_{\text{Conditional Independence}} \mid \{t_k\}_k \right) = \left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n} t_i}\right)^2 \text{Var}\left(\sum_{i=1}^{n} \underbrace{t_i y_i}_{i.i.d} \mid \{t_k\}_k\right) \\
&= \left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n} t_i}\right)^2 \sum_{i=1}^{n} \text{Var}\left(t_i y_i \mid \underbrace{\{t_k\}_k}_{t_i \text{ since samples are } i.i.d}\right) = \left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n} t_i}\right)^2 \sum_{i=1}^{n} \underbrace{\text{Var}\left(t_i y_i \mid t_i\right)}_{t_i \text{Var}(y_i \mid t_i) \ (t_i \text{ is binary})} \\
&= \left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n} t_i}\right)^2 \sum_{i=1}^{n} t_i \underbrace{\text{Var}\left(y_i \mid t_i\right)}_{\text{Var}(y \mid t=1) \ (\text{since } i.i.d. \text{ and multiplied by } t_i)} = \left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n} t_i}\right)^2 \text{Var}\left(y \mid t=1\right) \sum_{i=1}^{n} t_i = \frac{\mathbb{V}\left(Y_{t=1}\right)}{n\hat{\pi}}
\end{aligned}
\tag{9}
$$

For A2, we analogously get:

$$\text{Var}(\hat{\mathbb{E}}\left(Y_{t=0}\right) \mid \{t_k\}_k) = \frac{\mathbb{V}\left(Y_{t=0}\right)}{n\left(1-\hat{\pi}\right)}$$

For $\boxed{A3}$, we start by applying bi-linearity of covariance:

$$\text{Cov}\left(\hat{\mathbb{E}}\left(Y_{t=1}\right), \hat{\mathbb{E}}\left(Y_{t=0}\right) \mid \{t_k\}_k\right) = \text{Cov}\left(\frac{\sum_{i=1}^{n} t_i y_i}{\sum_{i=1}^{n} t_i}, \frac{\sum_{i=j}^{n}(1-t_j)y_j}{\sum_{j=1}^{n}(1-t_j)} \mid \{t_k\}_k\right) = \frac{\text{Cov}\left(\sum_{i=1}^{n} t_i y_i, \sum_{j=1}^{n}(1-t_j)y_j \mid \{t_k\}_k\right)}{\sum_{i=1}^{n} t_i \sum_{j=1}^{n}(1-t_j)}$$

$$= \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} t_i(1-t_j)\underbrace{\text{Cov}\left(y_i, y_j \mid \{t_k\}_k\right)}_{\neq 0 \text{ only if } i=j \text{ (since } i.i.d.)}}{\sum_{i=1}^{n} t_i \sum_{j=1}^{n}(1-t_j)} = \frac{\sum_{i=1}^{n} \underbrace{t_i(1-t_i)}_{=0}\text{Var}\left(y_i \mid \{t_k\}_k\right)}{\sum_{i=1}^{n} t_i \sum_{j=1}^{n}(1-t_j)} = 0$$

(10)

Injecting values of $\boxed{A1}$, $\boxed{A2}$, and $\boxed{A3}$ in Equation 8, we get expression for $\boxed{A}$:

$$\mathbb{E}[\text{Var}(\hat{\Phi}^{\text{SBD}} \mid \{t_k\}_k)] = \frac{\mathbb{V}\left(Y_{t=1}\right)}{\mathbb{E}\left[n\hat{\pi}\right]} + \frac{\mathbb{V}\left(Y_{t=0}\right)}{\mathbb{E}\left[n(1-\hat{\pi})\right]} = \frac{\mathbb{V}\left(Y_{t=1}\right)}{n\pi} + \frac{\mathbb{V}\left(Y_{t=0}\right)}{n(1-\pi)}$$

(11)

Computation of $\boxed{B}$

Given definition from Equation 5, term $\boxed{B}$ decomposes as:

$$\text{Var}\left(\mathbb{E}\left[\hat{\Phi}^{\text{SBD}} \mid \{t_k\}_k\right]\right) = \text{Var}\left(\underbrace{\mathbb{E}\left[\hat{\mathbb{E}}\left(Y_{t=1}\right) \mid \{t_k\}_k\right]}_{\boxed{B1}} - \underbrace{\mathbb{E}\left[\hat{\mathbb{E}}\left(Y_{t=0}\right) \mid \{t_k\}_k\right]}_{\boxed{B2}}\right)$$

(12)

For $\boxed{B1}$, we apply similar techniques from computing $\boxed{A}$ to get:

$$\mathbb{E}\left[\hat{\mathbb{E}}\left(Y_{t=1}\right) \mid \{t_k\}_k\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} t_i y_i}{\sum_{i=1}^{n} t_i} \mid \{t_k\}_k\right] = \frac{\mathbb{E}\left[\sum_{i=1}^{n} t_i y_i \mid \{t_k\}_k\right]}{\sum_{i=1}^{n} t_i} = \frac{\sum_{i=1}^{n} \mathbb{E}\left[t_i y_i \mid \{t_k\}_k\right]}{\sum_{i=1}^{n} t_i}$$

$$= \frac{\sum_{i=1}^{n} t_i \mathbb{E}\left[y_i \mid \{t_k\}_k\right]}{\sum_{i=1}^{n} t_i} = \frac{\sum_{i=1}^{n} t_i \underbrace{\mathbb{E}\left[y_i \mid t_i\right]}_{\mathbb{E}(Y_{t=1})}}{\sum_{i=1}^{n} t_i} = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} t_i}\mathbb{E}\left(Y_{t=1}\right) = \mathbb{E}\left(Y_{t=1}\right)$$

(13)

Analogously, $\boxed{B2}$ can be written as:

$$\mathbb{E}\left[\hat{\mathbb{E}}\left(Y_{t=0}\right) \mid \{t_k\}_k\right] = \mathbb{E}\left(Y_{t=0}\right)$$

Injecting $\boxed{B1}$ and $\boxed{B2}$ into Equation 12, we get the expression for $\boxed{B}$ as:

$$\text{Var}\left(\mathbb{E}\left[\hat{\Phi}^{\text{SBD}} \mid \{t_k\}_k\right]\right) = \text{Var}\left(\underbrace{\mathbb{E}\left(Y_{t=1}\right) - \mathbb{E}\left(Y_{t=0}\right)}_{\text{constant}}\right) = 0$$

(14)

Combining expression for Ⓐ (Equation 11) and Ⓑ (Equation 14), we get:

$$\text{Var}\left(\hat{\Phi}^{\text{SBD}}\right) = \frac{\mathbb{V}\left(Y_{t=1}\right)}{n\pi} + \frac{\mathbb{V}\left(Y_{t=0}\right)}{n\left(1-\pi\right)}$$

Given the full mediation assumption, we have the following decompositions:

$$P\left(y \mid t=0\right) = \begin{cases} P\left(y \mid a=0\right) & \text{(one-sided non-adherence)} \\ P\left(y \mid a=0\right)\left(1-P\left(a \mid t=0\right)\right) + P\left(y \mid a=1\right)P\left(a \mid t=0\right) & \text{(two-sided non-adherence)} \end{cases}$$

$$P\left(y \mid t=1\right) = P\left(y \mid a=0\right)\left(1-P\left(a \mid t=1\right)\right) + P\left(y \mid a=1\right)P\left(a \mid t=1\right)$$

Applying the above decompositions allow us to further expand $\mathbb{V}\left(Y_{t=0}\right)$ and $\mathbb{V}\left(Y_{t=1}\right)$. Under one-sided non-adherence, $\mathbb{V}\left(Y_{t=0}\right) = \mathbb{V}\left(Y_{a=0}\right)$, and under two-sided non-adherence, $\mathbb{V}\left(Y_{t=0}\right)$ decomposes as following:

$$\begin{aligned}
\mathbb{V}\left(Y_{t=0}\right) &= \text{Var}\left(Y_{a=0}(1-A_{t=0}) + Y_{a=1}A_{t=0}\right) \\
&= \text{Var}\left(Y_{a=0} - A_{t=0}Y_{a=0} + Y_{a=1}A_{t=0}\right) \\
&= \text{Var}\left(Y_{a=0}\right) + \text{Var}\left(A_{t=0}Y_{a=0}\right) + \text{Var}\left(A_{t=0}Y_{a=1}\right) - \\
&\quad 2\text{Cov}\left(Y_{a=0}, A_{t=0}Y_{a=0}\right) - 2\text{Cov}\left(A_{t=0}Y_{a=0}, A_{t=0}Y_{a=1}\right) + 2\text{Cov}\left(Y_{a=0}, A_{t=0}Y_{a=1}\right)
\end{aligned} \tag{15}$$

Leveraging the full ignorability assumption (*i.e.*, all potential outcomes and intakes are independent), we decompose the variances of produces:

$$\begin{aligned}
\text{Var}\left(A_{t=0}Y_{a=0}\right) &= \mathbb{E}\left(\left(A_{t=0}Y_{a=0}\right)^2\right) - \mathbb{E}\left(A_{t=0}Y_{a=0}\right)^2 \\
&= \mathbb{E}\left(A_{t=0}^2\right)\mathbb{E}\left(Y_{a=0}^2\right) - \mathbb{E}\left(A_{t=0}\right)^2\mathbb{E}\left(Y_{a=0}\right)^2 \quad \text{(Applying independence)} \\
&= \mathbb{E}\left(A_{t=0}^2\right)\left(\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(Y_{a=0}\right)^2\right) - \mathbb{E}\left(A_{t=0}\right)^2\mathbb{E}\left(Y_{a=0}\right)^2 \quad \text{(Definition of variance)} \\
&= \mathbb{E}\left(\underbrace{A_{t=0}^2}_{A_{t=0}\ \text{(binary)}}\right)\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(Y_{a=0}^2\right)\left(\underbrace{\mathbb{E}\left(A_{t=0}^2\right) - \mathbb{E}\left(A_{t=0}\right)^2}_{\text{Variance by definition}}\right) \\
&= \mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(Y_{a=0}\right)^2\mathbb{V}\left(A_{t=0}\right) \\
\mathbb{V}\left(A_{t=0}Y_{a=1}\right) &= \mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=1}\right) + \mathbb{E}\left(Y_{a=1}\right)^2\mathbb{V}\left(A_{t=0}\right) \quad \text{(Analogous as above)}
\end{aligned}$$

Similarly leveraging the full ignorability assumption, we decompose the covariance terms:

$$\begin{aligned}
\text{Cov}\left(Y_{a=0}, A_{t=0}Y_{a=0}\right) &= \mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=0}\right) \\
\text{Cov}\left(A_{t=0}Y_{a=0}, A_{t=0}Y_{a=1}\right) &= \mathbb{E}\left(Y_{a=0}\right)\mathbb{E}\left(Y_{a=1}\right)\mathbb{V}\left(A_{t=0}\right) \\
\text{Cov}\left(Y_{a=0}, A_{t=0}Y_{a=1}\right) &= 0 \quad \text{(Independence)}
\end{aligned}$$

Plug the above results back into the Equation 15, we get:

$$\begin{aligned}
\mathbb{V}\left(Y_{t=0}\right) &= \mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(Y_{a=0}\right)^2\mathbb{V}\left(A_{t=0}\right) + \mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=1}\right) + \mathbb{E}\left(Y_{a=1}\right)^2\mathbb{V}\left(A_{t=0}\right) \\
&\quad - 2\mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=0}\right) - 2\mathbb{E}\left(Y_{a=0}\right)\mathbb{E}\left(Y_{a=1}\right)\mathbb{V}\left(A_{t=0}\right) \\
&= \left(1 - 2\mathbb{E}\left(A_{t=0}\right) + \mathbb{E}\left(A_{t=0}\right)\right)\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=1}\right) + \left(\mathbb{E}\left(Y_{a=1}\right) - \mathbb{E}\left(Y_{a=1}\right)\right)^2\mathbb{V}\left(A_{t=0}\right) \\
&= \left(1 - \mathbb{E}\left(A_{t=0}\right)\right)\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=1}\right) + \left(\mathbb{E}\left(Y_{a=1}\right) - \mathbb{E}\left(Y_{a=0}\right)\right)^2\mathbb{V}\left(A_{t=0}\right) \\
&= \left(1 - \mathbb{E}\left(A_{t=0}\right)\right)\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=1}\right) + \Delta_Y^2\mathbb{V}\left(A_{t=0}\right)
\end{aligned}$$

Analogously, we have $\mathbb{V}\left(Y_{t=1}\right)$ as:

$$\mathbb{V}\left(Y_{t=1}\right) = \left(1 - \mathbb{E}\left(A_{t=1}\right)\right)\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=1}\right)\mathbb{V}\left(Y_{a=1}\right) + \Delta_Y^2\mathbb{V}\left(A_{t=1}\right)$$

Putting everything together, we have the variance of the SBD estimator:

$$\text{Var}\left(\hat{\Phi}_{\text{OS}}^{\text{SBD}}\right) = \frac{(1 - \mathbb{E}\left(A_{t=1}\right))\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=1}\right)\mathbb{V}\left(Y_{a=1}\right) + \Delta_A^2\mathbb{V}\left(A_{t=1}\right)}{n\pi} + \frac{\mathbb{V}\left(Y_{a=0}\right)}{n(1-\pi)}$$

$$\text{Var}\left(\hat{\Phi}_{\text{TS}}^{\text{SBD}}\right) = \frac{(1 - \mathbb{E}\left(A_{t=1}\right))\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=1}\right)\mathbb{V}\left(Y_{a=1}\right) + \Delta_A^2\mathbb{V}\left(A_{t=1}\right)}{n\pi} +$$

$$\frac{(1 - \mathbb{E}\left(A_{t=0}\right))\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=1}\right) + \Delta_A^2\mathbb{V}\left(A_{t=0}\right)}{n(1-\pi)}$$

D.2.2. CFD estimator' variance derivation under one-sided non-adherence

We start by applying the law of total variance to get:

$$\text{Var}(\hat{\Phi}_{\text{os}}^{\text{CFD}}) = \underbrace{\mathbb{E}\left[\text{Var}\left(\hat{\Phi}_{\text{os}}^{\text{CFD}}|\{a_k,t_k\}_k\right)\right]}_{\text{C}} + \underbrace{\text{Var}\left(\mathbb{E}\left[\hat{\Phi}_{\text{os}}^{\text{CFD}}|\{a_k,t_k\}_k\right]\right)}_{\text{D}} \tag{16}$$

Computation of C

Given definition from Equation 6, we can write term C as:

$$\mathbb{E}\left[\text{Var}\left(\hat{\Phi}_{\text{os}}^{\text{CFD}} \mid \{a_k,t_k\}_k\right)\right] = \mathbb{E}\left[\text{Var}((\hat{Y}_{a=1,t=1} - \hat{Y}_{a=0,t=0}(1-\hat{\pi}) - \hat{Y}_{a=0,t=1}\hat{\pi})\hat{\mathbb{E}}\left(A_{t=1}\right) \mid \{a_k,t_k\}_k)\right]$$

$$= \mathbb{E}[\hat{\mathbb{E}}\left(A_{t=1}\right)^2 \text{Var}(\hat{Y}_{a=1,t=1} - \hat{Y}_{a=0,t=0}(1-\hat{\pi}) - \hat{Y}_{a=0,t=1}\hat{\pi}|\{a_k,t_k\}_k)]$$

$$= \mathbb{E}\left[\hat{\mathbb{E}}\left(A_{t=1}\right)^2 \left(\underbrace{\text{Var}(\hat{Y}_{a=1,t=1}|\{a_k,t_k\}_k)}_{\text{C1}} + \underbrace{(1-\hat{\pi})^2\text{Var}(\hat{Y}_{a=0,t=0}|\{a_k,t_k\}_k)}_{\text{C2}} + \underbrace{\hat{\pi}^2\text{Var}(\hat{Y}_{a=0,t=1}|\{a_k,t_k\}_k)}_{\text{C3}}\right.\right.$$

$$\left.\left. - \underbrace{(1-\hat{\pi})\text{Cov}(\hat{Y}_{a=1,t=1}, \hat{Y}_{a=0,t=0}|\{a_k,t_k\}_k)}_{\text{C4}} - \underbrace{\hat{\pi}\text{Cov}(\hat{Y}_{a=1,t=1}, \hat{Y}_{a=0,t=1}|\{a_k,t_k\}_k)}_{\text{C5}}\right.\right.$$

$$\left.\left. + \underbrace{\hat{\pi}(1-\hat{\pi})\text{Cov}(\hat{Y}_{a=0,t=0}, \hat{Y}_{a=0,t=1}|\{a_k,t_k\}_k)}_{\text{C6}}\right)\right]$$

$$\tag{17}$$

For $\boxed{\text{C1}}$, we write:

$$\text{Var}(\hat{Y}_{a=1,t=1}|\{a_k,t_k\}_k)$$

$$= \text{Var}\left(\frac{\sum_{i=1}^{n}a_it_iy_i}{\sum_{i=1}^{n}a_it_i}\Big|\{a_k,t_k\}_k\right) = \left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n}a_it_i}\right)^2 \text{Var}\left(\sum_{i=1}^{n}a_it_iy_i|\underbrace{\{a_k,t_k\}_k}_{a_i,t_i, \text{ since samples are } i.i.d.}\right)$$

$$= \left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n}a_it_i}\right)^2 \sum_{i=1}^{n}\text{Var}\left(a_it_iy_i|a_i,t_i\right) \quad (\text{samples are } i.i.d.)$$

$$= \left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n}a_it_i}\right)^2 \sum_{i=1}^{n}\underbrace{(a_it_i)^2}_{a_it_i(\text{binary})}\underbrace{\text{Var}\left(y_i|a_i,t_i\right)}_{\text{Var}(y|a=1,t=1), \text{ since } i.i.d \text{ and multiplying with } a_i,t_i}$$

$$= \left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n}a_it_i}\right)^2 \text{Var}(y|a=1,t=1)\sum_{i=1}^{n}a_it_i = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}a_it_i}\text{Var}(y|a=1,t=1) = \frac{\mathbb{V}\left(Y_{a=1}\right)}{n\hat{\omega}_{a=1,t=1}}$$

(18)

Analogously, we get $\boxed{\text{C2}}$ and $\boxed{\text{C3}}$ as following:

$$(1-\hat{\pi})^2\text{Var}(\hat{Y}_{a=0,t=0}|\{a_k,t_k\}_k) = (1-\hat{\pi})^2\frac{\mathbb{V}\left(Y_{a=0}\right)}{n\hat{\omega}_{a=0,t=0}}$$

$$\hat{\pi}^2\text{Var}(\hat{Y}_{a=0,t=1}|\{a_k,t_k\}_k) = \hat{\pi}\frac{\mathbb{V}\left(Y_{a=0}\right)}{n\hat{\omega}_{a=0,t=1}}$$

(19)

For $\boxed{\text{C4}}$, we start by applying the bi-linearity of covariance:

$$(1-\hat{\pi})\text{Cov}(\hat{Y}_{a=1,t=1},\hat{Y}_{a=0,t=0}|\{a_k,t_k\}_k) = (1-\hat{\pi})\text{Cov}\left(\frac{\sum_{i=1}^{n}a_it_iy_i}{\sum_{i=1}^{n}a_it_i}, \frac{\sum_{j=1}^{n}(1-a_j)(1-t_j)y_j}{\sum_{j=1}^{n}(1-a_j)(1-t_j)}\Big|\{a_k,t_k\}_k\right)$$

$$= \frac{(1-\hat{\pi})}{\sum_{i=1}^{n}a_it_i\sum_{j=1}^{n}(1-a_j)(1-t_j)}\text{Cov}\left(\sum_{i=1}^{n}a_it_iy_i,\sum_{j=1}^{n}(1-a_j)(1-t_j)y_j|\{a_k,t_k\}_k\right)$$

$$= \frac{(1-\hat{\pi})}{\sum_{i=1}^{n}a_it_i\sum_{j=1}^{n}(1-a_j)(1-t_j)}\sum_{i=1}^{n}\sum_{j=1}^{n}a_it_i(1-a_j)(1-t_j)\underbrace{\text{Cov}\left(y_i,y_j|\{a_k,t_k\}_k\right)}_{\neq 0 \text{ only if } i=j \text{ because samples are } i.i.d.}$$

$$= \frac{(1-\hat{\pi})}{\sum_{i=1}^{n}a_it_i\sum_{j=1}^{n}(1-a_j)(1-t_j)}\sum_{i=1}^{n}\underbrace{a_it_i(1-a_i)(1-t_i)}_{0}\text{Cov}\left(y_i,y_j|\{a_k,t_k\}_k\right) = 0$$

(20)

Analogously, $\boxed{\text{C5}}$ and $\boxed{\text{C6}}$ are also 0.

Injecting $\boxed{\text{C1}}$, $\boxed{\text{C2}}$, $\boxed{\text{C3}}$ back into $\boxed{\text{C}}$, we get:

$$\mathbb{E}[\text{Var}(\hat{\Phi}_{\text{os}}^{\text{CFD}}|\{a_k, t_k\}_k)] = \mathbb{E}\left[\hat{\mathbb{E}}\left(A_{t=1}\right)^2 \left(\frac{\mathbb{V}\left(Y_{a=1}\right)}{n\hat{\omega}_{a=1,t=1}} + (1-\hat{\pi})^2 \frac{\mathbb{V}\left(Y_{a=0}\right)}{n\hat{\omega}_{a=0,t=0}} + \hat{\pi}^2 \frac{\mathbb{V}\left(Y_{a=0}\right)}{n\hat{\omega}_{a=0,t=1}}\right)\right]$$

$$= \frac{1}{n}\left(\mathbb{E}\left[\frac{\hat{\mathbb{E}}\left(A_{t=1}\right)^2}{\hat{\omega}_{a=1,t=1}}\right] \mathbb{V}\left(Y_{a=1}\right) + \left(\mathbb{E}\left[\frac{\hat{\mathbb{E}}\left(A_{t=1}\right)^2 (1-\hat{\pi})^2}{\hat{\omega}_{a=0,t=0}}\right] + \mathbb{E}\left[\frac{\hat{\mathbb{E}}\left(A_{t=1}\right)^2 \hat{\pi}^2}{\hat{\omega}_{a=0,t=1}}\right]\right)\mathbb{V}\left(Y_{a=0}\right)\right) \tag{21}$$

Computation of $\boxed{\text{D}}$

Given definition from Equation 6, we can write term $\boxed{\text{D}}$ as:

$$\text{Var}(\mathbb{E}[\hat{\Phi}^{\text{CFD}_{\text{os}}}|\{a_k, t_k\}_k]) = \text{Var}(\mathbb{E}[(\hat{Y}_{a=1,t=1} - \hat{Y}_{a=0,t=0}(1-\hat{\pi}) - \hat{Y}_{a=0,t=1}\hat{\pi})\hat{\mathbb{E}}\left(A_{t=1}\right) \mid \{a_k, t_k\}_k])$$

$$= \text{Var}(\hat{\mathbb{E}}\left(A_{t=1}\right)\mathbb{E}[\hat{Y}_{a=1,t=1} - \hat{Y}_{a=0,t=0}(1-\hat{\pi}) - \hat{Y}_{a=0,t=1}\hat{\pi} \mid \{a_k, t_k\}_k])$$

$$= \text{Var}\left(\hat{\mathbb{E}}\left(A_{t=1}\right)\left(\underbrace{\mathbb{E}[\hat{Y}_{a=1,t=1} \mid \{a_k, t_k\}_k]}_{\boxed{\text{D1}}} - \underbrace{(1-\hat{\pi})\mathbb{E}[\hat{Y}_{a=0,t=0} \mid \{a_k, t_k\}_k]}_{\boxed{\text{D2}}} - \underbrace{\hat{\pi}\mathbb{E}[\hat{Y}_{a=0,t=1} \mid \{a_k, t_k\}_k]}_{\boxed{\text{D3}}}\right)\right) \tag{22}$$

For $\boxed{\text{D1}}$, we write:

$$\mathbb{E}[\hat{Y}_{a=1,t=1} \mid \{a_k, t_k\}_k] = \mathbb{E}\left[\frac{\sum_{i=1}^n a_i t_i y_i}{\sum_{i=1}^n a_i t_i} \mid \{a_k, t_k\}_k\right] = \frac{1}{\sum_{i=1}^n a_i t_i}\mathbb{E}\left[\sum_{i=1}^n a_i t_i y_i \mid \{a_k, t_k\}_k\right]$$

$$= \frac{1}{\sum_{i=1}^n a_i t_i}\sum_{i=1}^n \mathbb{E}\left[a_i t_i y_i \mid \underbrace{\{a_k, t_k\}_k}_{a_i, t_i, \text{ because } i.i.d. \text{ samples}}\right] \tag{23}$$

$$= \frac{1}{\sum_{i=1}^n a_i t_i}\sum_{i=1}^n \mathbb{E}[a_i t_i y_i | a_i, t_i] = \frac{1}{\sum_{i=1}^n a_i t_i}\sum_{i=1}^n a_i t_i \underbrace{\mathbb{E}[y_i | a_i, t_i]}_{\mathbb{E}[y|a=1,t=1] \ i.i.d. \ and \ multiplying \ by \ a_i t_i}$$

$$= \frac{1}{\sum_{i=1}^n a_i t_i}\mathbb{E}[y|a=1,t=1]\sum_{i=1}^n a_i t_i = \mathbb{E}[y|a=1,t=1] = \mathbb{E}[Y_{a=1,t=1}] = \mathbb{E}(Y_{a=1})$$

Analogously, we get $\boxed{\text{D2}}$ and $\boxed{\text{D3}}$ as following:

$$(1-\hat{\pi})\mathbb{E}[\hat{Y}_{a=0,t=0}|\{a_k, t_k\}_k] = (1-\hat{\pi})\mathbb{E}(Y_{a=0})$$
$$\hat{\pi}\mathbb{E}[\hat{Y}_{a=0,t=1}|\{a_k, t_k\}_k] = \hat{\pi}\mathbb{E}(Y_{a=0}) \tag{24}$$

Injecting $\boxed{\text{D1}}$, $\boxed{\text{D2}}$, $\boxed{\text{D3}}$ back into $\boxed{\text{D}}$ (equation 22), we get:

$$\text{Var}(\mathbb{E}[\hat{\Phi}_{\text{os}}^{\text{CFD}}|\{a_k, t_k\}_k]) = \text{Var}\left(\hat{\mathbb{E}}\left(A_{t=1}\right)\left(\mathbb{E}\left(Y_{a=1}\right) - (1-\hat{\pi})\mathbb{E}\left(Y_{a=0}\right) - \hat{\pi}\mathbb{E}\left(Y_{a=0}\right)\right)\right)$$

$$= \text{Var}\left(\hat{\mathbb{E}}\left(A_{t=1}\right)\left(\mathbb{E}\left(Y_{a=1}\right) - \mathbb{E}\left(Y_{a=0}\right)\right)\right) = \Delta_Y^2 \underbrace{\text{Var}\left(\hat{\mathbb{E}}\left(A_{t=1}\right)\right)}_{\boxed{\text{D7}}} \tag{25}$$

To simplify $\boxed{D7}$, we again apply the law of total variance to get:

$$\text{Var}\left(\hat{\mathbb{E}}\left(A_{t=1}\right)\right) = \underbrace{\mathbb{E}[\text{Var}(\hat{\mathbb{E}}\left(A_{t=1}\right) \mid \{t_j\}_k)]}_{\boxed{D7.1}} + \underbrace{\text{Var}(\mathbb{E}[\hat{\mathbb{E}}\left(A_{t=1}\right) \mid \{t_j\}_k])}_{\boxed{D7.2}} \tag{26}$$

Analogous to the derivation of $\boxed{A1}$ (Equation 9), $\boxed{D7.1}$ can be written as:

$$\mathbb{E}[\text{Var}(\hat{\mathbb{E}}\left(A_{t=1}\right) \mid \{t_j\}_k)] = \frac{\mathbb{V}\left(A_{t=1}\right)}{n\pi} \tag{27}$$

Analogous to the derivation of $\boxed{B1}$ (Equation 13), $\boxed{D7.2}$ can be written as:

$$\text{Var}\left(\mathbb{E}[\hat{\mathbb{E}}\left(A_{t=1}\right) \mid \{t_j\}_k]\right) = \text{Var}\left(\mathbb{E}\left(A_{t=1}\right)\right) = 0$$

Injecting $\boxed{D7.1}$ and $\boxed{D7.2}$ back to $\boxed{D}$, we get:

$$\text{Var}\left(\mathbb{E}[\hat{\Phi}_{\text{os}}^{\text{CFD}} | \{a_k, t_k\}_k]\right) = \Delta_Y^2 \frac{\mathbb{V}\left(A_{t=1}\right)}{n\pi} \tag{28}$$

Now injecting $\boxed{C}$ and $\boxed{D}$ back into Equation 16, we get variance for $\hat{\Phi}^{\text{CFD}_{\text{os}}}$ as:

$$\text{Var}(\hat{\Phi}_{\text{os}}^{\text{CFD}}) = \frac{1}{n}\left(\mathbb{V}\left(Y_{a=0}\right)\left(\mathbb{E}\left[\frac{\hat{\mathbb{E}}\left(A_{t=1}\right)^2\left(1-\hat{\pi}\right)^2}{\hat{\omega}_{a=0,t=0}}\right] + \mathbb{E}\left[\frac{\hat{\mathbb{E}}\left(A_{t=1}\right)^2\hat{\pi}^2}{\hat{\omega}_{a=0,t=1}}\right]\right) + \right.$$
$$\left. + \mathbb{V}\left(Y_{a=1}\right)\mathbb{E}\left[\frac{\hat{\mathbb{E}}\left(A_{t=1}\right)^2}{\hat{\omega}_{a=1,t=1}}\right] + \mathbb{V}\left(A_{t=1}\right)\frac{\Delta_Y^2}{\pi}\right)$$

D.2.3. CFD ESTIMATOR' VARIANCE DERIVATION UNDER TWO-SIDED NON-ADHERENCE

We start by applying the law of total variance to get:

$$\text{Var}(\hat{\Phi}_{\text{ts}}^{\text{CFD}}) = \underbrace{\mathbb{E}\left[\text{Var}\left(\hat{\Phi}_{\text{ts}}^{\text{CFD}} \mid \{a_k, t_k\}_k\right)\right]}_{\boxed{E}} + \underbrace{\text{Var}\left(\mathbb{E}\left[\hat{\Phi}_{\text{ts}}^{\text{CFD}} \mid \{a_k, t_k\}_k\right]\right)}_{\boxed{F}} \tag{29}$$

Computation of $\left(\textbf{E}\right)$

Given definition from Equation 7, we can write term $\left(\text{E}\right)$ as:

$$\mathbb{E}[\text{Var}(\hat{\Phi}_{\text{ts}}^{\text{CFD}}|\{a_k, t_k\}_k)]$$
$$= \mathbb{E}[\text{Var}(((\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1 - \hat{\pi}) + (\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))\hat{\pi})(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))|\{a_k, t_k\}_k)]$$
$$= \mathbb{E}[(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 \text{Var}((\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1 - \hat{\pi}) + (\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))\hat{\pi}|\{a_k, t_k\}_k)]$$
$$= \mathbb{E}[(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 (2\underbrace{\text{Cov}((\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1 - \hat{\pi}), (\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))\hat{\pi}|\{a_k, t_k\}_k)}_{\text{E1}} +$$

$$\underbrace{\text{Var}((\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1 - \hat{\pi})|\{a_k, t_k\}_k)}_{\text{E2}} + \underbrace{\text{Var}((\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))\hat{\pi}|\{a_k, t_k\}_k))]}_{\text{E3}}$$

$$(30)$$

For $\left(\text{E1}\right)$, start by applying the bi-linearity of covariance:

$$\text{Cov}((\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1 - \hat{\pi}), (\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))\hat{\pi}|\{a_k, t_k\}_k)$$
$$= (1 - \hat{\pi})\hat{\pi}\text{Cov}((\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0})), (\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))|\{a_k, t_k\}_k)$$
$$= (1 - \hat{\pi})\hat{\pi}(\underbrace{\text{Cov}(\hat{\mathbb{E}}(Y_{a=1,t=0}), \hat{\mathbb{E}}(Y_{a=1,t=1})|\{a_k, t_k\}_k)}_{\text{E1.1}} + \underbrace{\text{Cov}(\hat{\mathbb{E}}(Y_{a=0,t=0}), \hat{\mathbb{E}}(Y_{a=0,t=1})|\{a_k, t_k\}_k)}_{\text{E1.2}}$$

$$- \underbrace{\text{Cov}(\hat{\mathbb{E}}(Y_{a=1,t=0}), \hat{\mathbb{E}}(Y_{a=0,t=1})|\{a_k, t_k\}_k)}_{\text{E1.3}} - \underbrace{\text{Cov}(\hat{\mathbb{E}}(Y_{a=0,t=0}), \hat{\mathbb{E}}(Y_{a=1,t=1})|\{a_k, t_k\}_k))}_{\text{E1.4}}$$

For $\left(\text{E1.1}\right)$, also applying bi-linearity of covariance, we write:

$$\text{Cov}(\hat{\mathbb{E}}(Y_{a=1,t=0}), \hat{\mathbb{E}}(Y_{a=1,t=1})|\{a_k, t_k\}_k) = \text{Cov}\left(\frac{\sum_{i=1}^{n} a_i(1 - t_i)y_i}{\sum_{i=1}^{n} a_i(1 - t_i)}, \frac{\sum_{j=1}^{n} a_j t_j y_j}{\sum_{j=1}^{n} a_j t_j}|\{a_k, t_k\}_k\right)$$
$$= \frac{1}{\sum_{i=1}^{n} a_i(1 - t_i)}\frac{1}{\sum_{j=1}^{n} a_j t_j}\text{Cov}\left(\sum_{i=1}^{n} a_i(1 - t_i)y_i, \sum_{j=1}^{n} a_j t_j y_j|\{a_k, t_k\}_k\right)$$
$$= \frac{1}{\sum_{i=1}^{n} a_i(1 - t_i)}\frac{1}{\sum_{j=1}^{n} a_j t_j}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j(1 - t_i)t_j \underbrace{\text{Cov}(y_i, y_j|\{a_k, t_k\}_k)}_{\neq 0 \text{ only if } i=j \text{ (since } i.i.d,)}$$
$$= \frac{1}{\sum_{i=1}^{n} a_i(1 - t_i)}\frac{1}{\sum_{j=1}^{n} a_j t_j}\sum_{i=1}^{n} a_i a_i \underbrace{(1 - t_i)t_i}_{0} \text{Cov}(y_i, y_i|\{a_k, t_k\}_k) = 0$$

$$(31)$$

Analogously, we have (E1.2), (E1.3), (E1.4) also equal to 0. Therefore, we have (E1):

$$\text{Cov}((\hat{\mathbb{E}}\,(Y_{a=1,t=0}) - \hat{\mathbb{E}}\,(Y_{a=0,t=0}))(1 - \hat{\pi}), (\hat{\mathbb{E}}\,(Y_{a=1,t=1}) - \hat{\mathbb{E}}\,(Y_{a=0,t=1}))\hat{\pi}|\{a_k, t_k\}_k) = 0$$

For (E2), we write:

$$\text{Var}((\hat{\mathbb{E}}\,(Y_{a=1,t=0}) - \hat{\mathbb{E}}\,(Y_{a=0,t=0}))(1 - \hat{\pi})|\{a_k, t_k\}_k) = (1 - \hat{\pi})^2 \text{Var}((\hat{\mathbb{E}}\,(Y_{a=1,t=0}) - \hat{\mathbb{E}}\,(Y_{a=0,t=0}))|\{a_k, t_k\}_k)$$

$$= (1 - \hat{\pi})^2 (\underbrace{\text{Var}(\hat{\mathbb{E}}\,(Y_{a=1,t=0})\,|\{a_k, t_k\}_k)}_{\text{E2.1}} + \underbrace{\text{Var}(\hat{\mathbb{E}}\,(Y_{a=0,t=0})\,|\{a_k, t_k\}_k)}_{\text{E2.2}}$$

$$- 2\underbrace{\text{Cov}(\hat{\mathbb{E}}\,(Y_{a=1,t=0})\,, \hat{\mathbb{E}}\,(Y_{a=0,t=0})\,|\{a_k, t_k\}_k))}_{\text{E2.3}}$$

$$(32)$$

For (E2.1), we write:

$$\text{Var}(\hat{\mathbb{E}}\,(Y_{a=1,t=0})\,|\{a_k, t_k\}_k) = \text{Var}\left(\frac{\sum_{i=1}^{n} a_i(1-t_i)y_i}{\sum_{i=1}^{n} a_i(1-t_i)}|\{a_k, t_k\}_k\right)$$

$$= \left(\frac{1}{\sum_{i=1}^{n} a_i(1-t_i)}\right)^2 \text{Var}\left(\sum_{i=1}^{n} \underbrace{a_i(1-t_i)y_i}_{i.i.d.}|\underbrace{\{a_k, t_k\}_k}_{a_i, t_i \text{ since } i.i.d.}\right) = \left(\frac{1}{\sum_{i=1}^{n} a_i(1-t_i)}\right)^2 \sum_{i=1}^{n} \text{Var}\,(a_i(1-t_i)y_i|a_i, t_i)$$

$$= \left(\frac{1}{\sum_{i=1}^{n} a_i(1-t_i)}\right)^2 \sum_{i=1}^{n} \underbrace{(a_i(1-t_i))^2}_{a_i(1-t_i) \text{ since binary.}} \underbrace{\text{Var}\,(y_i|a_i, t_i)}_{\text{Var}(y|a=1,t=0) \text{ (because } i.i.d. \text{ and multiply by } a_i(1-t_i))}$$

$$= \left(\frac{1}{\sum_{i=1}^{n} a_i(1-t_i)}\right)^2 \text{Var}(y|a=1, t=0) \sum_{i=1}^{n} a_i(1-t_i) = \frac{\mathbb{V}\,(Y_{a=1,t=0})}{\sum_{i=1}^{n} a_i(1-t_i)} = \frac{\mathbb{V}\,(Y_{a=1})}{n\hat{\omega}_{a=1,t=0}}$$

Analogously, we get expression for (E2.2) as:

$$\text{Var}(\hat{\mathbb{E}}\,(Y_{a=0,t=0})\,|\{a_k, t_k\}_k) = \frac{\mathbb{V}\,(Y_{a=0})}{n\hat{\omega}_{a=0,t=0}}$$

For (E2.3), analogous to Equation 31, we can get:

$$\text{Cov}(\hat{\mathbb{E}}\,(Y_{a=1,t=0})\,, \hat{\mathbb{E}}\,(Y_{a=0,t=0})\,|\{a_k, t_k\}_k) = 0$$

Injecting $\boxed{\text{E2.1}}, \boxed{\text{E2.2}}, \boxed{\text{E2.3}}$ back to Equation 32, we get expression for $\boxed{\text{E2}}$:

$$\mathrm{Var}((\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1 - \hat{\pi}) mid\{a_k, t_k\}_k) = (1 - \hat{\pi})^2 \left(\frac{\mathbb{V}(Y_{a=1})}{n\hat{\omega}_{a=2,t=0}} + \frac{\mathbb{V}(Y_{a=0})}{n\hat{\omega}_{a=0,t=0}} \right)$$

Analogously, we can get expression for $\boxed{\text{E3}}$:

$$\mathrm{Var}((\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))\hat{\pi} \mid \{a_k, t_k\}_k) = \hat{\pi}^2 \left(\frac{\mathbb{V}(Y_{a=1})}{n\hat{\omega}_{a=1,t=1}} + \frac{\mathbb{V}(Y_{a=0})}{n\hat{\omega}_{a=0,t=1}} \right)$$

Injecting $\boxed{\text{E1}}, \boxed{\text{E2}}, \boxed{\text{E3}}$ into Equation 30, we get expression for $\boxed{\text{E}}$:

$$\mathbb{E}[\mathrm{Var}(\hat{\Phi}_{ts}^{\mathrm{CFD}}|\{a_k, t_k\}_k)] = \mathbb{E}\left[\left(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}) \right)^2 \left((1 - \hat{\pi})^2 \left(\frac{\mathbb{V}(Y_{a=1})}{n\hat{\omega}_{a=1,t=0}} + \frac{\mathbb{V}(Y_{a=0})}{n\hat{\omega}_{a=0,t=0}} \right) \right. \right.$$

$$\left. \left. + \hat{\pi}^2 \left(\frac{\mathbb{V}(Y_{a=1})}{n\hat{\omega}_{a=1,t=1}} + \frac{\mathbb{V}(Y_{a=0})}{n\hat{\omega}_{a=0,t=1}} \right) \right) \right]$$

$$= \mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 (1 - \hat{\pi})^2}{n\hat{\omega}_{a=1,t=0}} \right] \mathbb{V}(Y_{a=1}) + \mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 (1 - \hat{\pi})^2}{n\hat{\omega}_{a=0,t=0}} \right] \mathbb{V}(Y_{a=0})$$

$$+ \mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 \hat{\pi}^2}{n\hat{\omega}_{a=1,t=1}} \right] \mathbb{V}(Y_{a=1}) + \mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 \hat{\pi}^2}{n\hat{\omega}_{a=0,t=1}} \right] \mathbb{V}(Y_{a=0})$$

$$= \frac{1}{n} \left(\mathbb{V}(Y_{a=1}) \left(\mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 (1 - \hat{\pi})^2}{\hat{\omega}_{a=1,t=0}} \right] + \mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 \hat{\pi}^2}{\hat{\omega}_{a=1,t=1}} \right] \right) \right.$$

$$\left. + \mathbb{V}(Y_{a=0}) \left(\mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 (1 - \hat{\pi})^2}{\hat{\omega}_{a=0,t=0}} \right] + \mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 \hat{\pi}^2}{\hat{\omega}_{a=0,t=1}} \right] \right) \right)$$

Computation of $\boxed{\text{F}}$

Given definition from Equation 7, we can write term $\boxed{\text{F}}$ as:

$$\mathrm{Var}\left(\mathbb{E}\left[\hat{\Phi}_{ts}^{\mathrm{CFD}}|\{a_k, t_k\}_k \right] \right)$$

$$= \mathrm{Var}\left(\mathbb{E}\left[((\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1 - \hat{\pi}) + (\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))\hat{\pi})(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))|\{a_k, t_k\}_k) \right] \right)$$

$$= \mathrm{Var}\left((\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0})) \left(\underbrace{\mathbb{E}\left[(\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1 - \hat{\pi})|\{a_k, t_k\}_k \right]}_{\boxed{\text{F1}}} \right. \right.$$

$$\left. \left. + \underbrace{\mathbb{E}\left[(\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))\hat{\pi}|\{a_k, t_k\}_k \right]}_{\boxed{\text{F2}}} \right) \right)$$

$$(33)$$

For $\boxed{F1}$ we write:

$$\mathbb{E}\left[(\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1-\hat{\pi})|\{a_k,t_k\}_k\right] = (1-\hat{\pi})\mathbb{E}\left[\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0})|\{a_k,t_k\}_k\right]$$

$$= (1-\hat{\pi})\left(\underbrace{\mathbb{E}\left[\hat{\mathbb{E}}(Y_{a=1,t=0})|\{a_k,t_k\}_k\right]}_{F1.1} - \underbrace{\mathbb{E}\left[\hat{\mathbb{E}}(Y_{a=0,t=0})|\{a_k,t_k\}_k\right]}_{F1.2}\right)$$

(34)

For $\boxed{F1.1}$ we write:

$$\mathbb{E}\left[\hat{\mathbb{E}}(Y_{a=1,t=0})|\{a_k,t_k\}_k\right] = \mathbb{E}\left[\frac{\sum_{i=1}^{n} a_i(1-t_i)y_i}{\sum_{i=1}^{n} a_i(1-t_i)}|\{a_k,t_k\}_k\right] = \frac{1}{\sum_{i=1}^{n} a_i(1-t_i)}\mathbb{E}\left[\sum_{i=1}^{n} a_i(1-t_i)y_i|\{a_k,t_k\}_k\right]$$

$$= \frac{1}{\sum_{i=1}^{n} a_i(1-t_i)}\sum_{i=1}^{n}\mathbb{E}\left[a_i(1-t_i)y_i|\{a_k,t_k\}_k\right] = \frac{1}{\sum_{i=1}^{n} a_i(1-t_i)}\sum_{i=1}^{n} a_i(1-t_i)\mathbb{E}\left[y_i|\{a_k,t_k\}_k\right]$$

$$= \frac{1}{\sum_{i=1}^{n} a_i(1-t_i)}\sum_{i=1}^{n} a_i(1-t_i)\underbrace{\mathbb{E}\left[y_i|a_i,t_i\right]}_{Y_{a=1,t=0}\ (i.i.d.\ \text{and multiply with } a_i(1-t_i))}$$

$$= \frac{\mathbb{E}(Y_{a=1,t=0})}{\sum_{i=1}^{n} a_i(1-t_i)}\sum_{i=1}^{n} a_i(1-t_i) = \mathbb{E}(Y_{a=1,t=0}) = \mathbb{E}(Y_{a=1})$$

Analogously, we get expression for $\boxed{F1.2}$:

$$\mathbb{E}\left[\hat{\mathbb{E}}(Y_{a=0,t=0})|\{a_k,t_k\}_k\right] = \mathbb{E}(Y_{a=0})$$

Injecting $\boxed{F1.1}$ and $\boxed{F1.2}$ into Equation 34, we get expression for $\boxed{F1}$ as:

$$\mathbb{E}\left[(\hat{\mathbb{E}}(Y_{a=1,t=0}) - \hat{\mathbb{E}}(Y_{a=0,t=0}))(1-\hat{\pi})|\{a_k,t_k\}_k\right] = (1-\hat{\pi})\left(\mathbb{E}(Y_{a=1}) - \mathbb{E}(Y_{a=0})\right)$$

Analogously, we get expression for $\boxed{F2}$ as:

$$\mathbb{E}\left[(\hat{\mathbb{E}}(Y_{a=1,t=1}) - \hat{\mathbb{E}}(Y_{a=0,t=1}))\hat{\pi}|\{a_k,t_k\}_k\right] = \hat{\pi}\left(\mathbb{E}(Y_{a=1}) - \mathbb{E}(Y_{a=0})\right)$$

Injecting $\boxed{F1}$, $\boxed{F2}$ into Equation 33, we get expression for $\boxed{F}$ as:

$$
\begin{aligned}
&\text{Var}\left(\mathbb{E}\left[\hat{\Phi}_{\text{ts}}^{\text{CFD}}|\{a_k, t_k\}_k\right]\right) \\
&= \text{Var}\left((\hat{\mathbb{E}}\left(A_{t=1}\right) - \hat{\mathbb{E}}\left(A_{t=0}\right))\left((1-\hat{\pi})\left(\mathbb{E}\left(Y_{a=1}\right) - \mathbb{E}\left(Y_{a=0}\right)\right) + \hat{\pi}\left(Y_{a=1,t=1} - \mathbb{E}\left(Y_{a=0}\right)\right)\right)\right) \\
&= \text{Var}\left((\hat{\mathbb{E}}\left(A_{t=1}\right) - \hat{\mathbb{E}}\left(A_{t=0}\right))\left(\mathbb{E}\left(Y_{a=1}\right) - \mathbb{E}\left(Y_{a=0}\right)\right)\right) = \Delta_Y^2 \underbrace{\text{Var}\left(\hat{\mathbb{E}}\left(A_{t=1}\right) - \hat{\mathbb{E}}\left(A_{t=0}\right)\right)}_{\boxed{F3}}
\end{aligned} \tag{35}
$$

Analogous to the derivation for $\text{Var}\left(\Phi^{\hat{\text{SBD}}}\right)$ (Section D.2.1), we can write expression for $\boxed{F3}$ as:

$$
\text{Var}\left(\hat{\mathbb{E}}\left(A_{t=1}\right) - \hat{\mathbb{E}}\left(A_{t=0}\right)\right) = \frac{\mathbb{V}\left(A_{t=1}\right)}{n\pi} + \frac{\mathbb{V}\left(A_{t=0}\right)}{n(1-\pi)}
$$

Injecting $\boxed{F3}$ into Equation 35, we get expression for $\boxed{F}$ as:

$$
\text{Var}\left(\mathbb{E}\left[\hat{\Phi}_{\text{ts}}^{\text{CFD}}|\{a_k, t_k\}_k\right]\right) = \left(\mathbb{E}\left(Y_{a=1}\right) - \mathbb{E}\left(Y_{a=0}\right)\right)^2 \left(\frac{\mathbb{V}\left(A_{t=1}\right)}{n\pi} + \frac{\mathbb{V}\left(A_{t=0}\right)}{n(1-\pi)}\right)
$$

Injecting $\boxed{E}$, $\boxed{F}$ back into Equation 29, we get variance for $\hat{\Phi}_{\text{ts}}^{\text{CFD}}$ as:

$$
\begin{aligned}
\text{Var}(\hat{\Phi}_{\text{ts}}^{\text{CFD}}) = \frac{1}{n}&\left(\mathbb{V}\left(Y_{a=1}\right)\left(\mathbb{E}\left[\frac{(\hat{\mathbb{E}}\left(A_{t=1}\right) - \hat{\mathbb{E}}\left(A_{t=0}\right))^2(1-\hat{\pi})^2}{\hat{\omega}_{a=1,t=0}}\right] + \mathbb{E}\left[\frac{(\hat{\mathbb{E}}\left(A_{t=1}\right) - \hat{\mathbb{E}}\left(A_{t=0}\right))^2\hat{\pi}^2}{\hat{\omega}_{a=1,t=1}}\right]\right) \right. \\
&+ \mathbb{V}\left(Y_{a=0}\right)\left(\mathbb{E}\left[\frac{(\hat{\mathbb{E}}\left(A_{t=1}\right) - \hat{\mathbb{E}}\left(A_{t=0}\right))^2(1-\hat{\pi})^2}{\hat{\omega}_{a=0,t=0}}\right] + \mathbb{E}\left[\frac{(\hat{\mathbb{E}}\left(A_{t=1}\right) - \hat{\mathbb{E}}\left(A_{t=0}\right))^2\hat{\pi}^2}{\hat{\omega}_{a=0,t=1}}\right]\right) \\
&\left. + \Delta_Y^2\left(\frac{\mathbb{V}\left(A_{t=1}\right)}{\pi} + \frac{\mathbb{V}\left(A_{t=0}\right)}{1-\pi}\right)\right)
\end{aligned}
$$

D.3. Derive asymptotic variance upper and lower bounds

From previous sections, we get variance for $\hat{\Phi}_{\text{os}}^{\text{SBD}}$, $\hat{\Phi}_{\text{ts}}^{\text{SBD}}$, $\hat{\Phi}_{\text{os}}^{\text{CFD}}$, and $\hat{\Phi}_{\text{ts}}^{\text{CFD}}$ as;

$$
\text{Var}\left(\hat{\Phi}_{\text{OS}}^{\text{SBD}}\right) = \frac{\left(1 - \mathbb{E}\left(A_{t=1}\right)\right)\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=1}\right)\mathbb{V}\left(Y_{a=1}\right) + \Delta_Y^2 \mathbb{V}\left(A_{t=1}\right)}{n\pi} + \frac{\mathbb{V}\left(Y_{a=0}\right)}{n(1-\pi)}
$$

$$
\begin{aligned}
\text{Var}\left(\hat{\Phi}_{\text{TS}}^{\text{SBD}}\right) = &\frac{\left(1 - \mathbb{E}\left(A_{t=1}\right)\right)\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=1}\right)\mathbb{V}\left(Y_{a=1}\right) + \Delta_Y^2 \mathbb{V}\left(A_{t=1}\right)}{n\pi} + \\
&\frac{\left(1 - \mathbb{E}\left(A_{t=0}\right)\right)\mathbb{V}\left(Y_{a=0}\right) + \mathbb{E}\left(A_{t=0}\right)\mathbb{V}\left(Y_{a=1}\right) + \Delta_Y^2 \mathbb{V}\left(A_{t=0}\right)}{n(1-\pi)}
\end{aligned}
$$

$$\mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{CFD}}) = \frac{1}{n} \left(\mathbb{V}(Y_{a=0}) \left(\mathbb{E}\left[\frac{\hat{\mathbb{E}}(A_{t=1})^2 (1-\hat{\pi})^2}{\hat{\omega}_{a=0,t=0}} \right] + \mathbb{E}\left[\frac{\hat{\mathbb{E}}(A_{t=1})^2 \hat{\pi}^2}{\hat{\omega}_{a=0,t=1}} \right] \right) + \mathbb{V}(Y_{a=1}) \mathbb{E}\left[\frac{\hat{\mathbb{E}}(A_{t=1})^2}{\hat{\omega}_{a=1,t=1}} \right] + \right.$$

$$\left. \mathbb{V}(A_{t=1}) \frac{\Delta_Y^2}{\pi} \right)$$

$$\mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{CFD}}) = \frac{1}{n} \left(\mathbb{V}(Y_{a=1}) \left(\mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 (1-\hat{\pi})^2}{\hat{\omega}_{a=1,t=0}} \right] + \mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 \hat{\pi}^2}{\hat{\omega}_{a=1,t=1}} \right] \right) \right.$$

$$+ \mathbb{V}(Y_{a=0}) \left(\mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 (1-\hat{\pi})^2}{\hat{\omega}_{a=0,t=0}} \right] + \mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 \hat{\pi}^2}{\hat{\omega}_{a=0,t=1}} \right] \right)$$

$$\left. + \Delta_Y^2 \left(\frac{\mathbb{V}(A_{t=1})}{\pi} + \frac{\mathbb{V}(A_{t=0})}{1-\pi} \right) \right)$$

D.3.1. ASYMPTOTIC LOWER BOUNDS OF $\mathrm{VAR}(\hat{\Phi}^{\mathrm{SBD}})$

Based on the derivation for $\hat{\Phi}_{\mathrm{os}}^{\mathrm{SBD}}$ and $\hat{\Phi}_{\mathrm{ts}}^{\mathrm{SBD}}$, we get following with a slight rewriting:

$$n\mathrm{Var}\left(\hat{\Phi}_{\mathrm{OS}}^{\mathrm{SBD}}\right) = \frac{(1-\mathbb{E}(A_{t=1}))\mathbb{V}(Y_{a=0}) + \mathbb{E}(A_{t=1})\mathbb{V}(Y_{a=1}) + \Delta_Y^2 \mathbb{V}(A_{t=1})}{\pi} + \frac{\mathbb{V}(Y_{a=0})}{(1-\pi)}$$

$$n\mathrm{Var}\left(\hat{\Phi}_{\mathrm{TS}}^{\mathrm{SBD}}\right) = \frac{(1-\mathbb{E}(A_{t=1}))\mathbb{V}(Y_{a=0}) + \mathbb{E}(A_{t=1})\mathbb{V}(Y_{a=1}) + \Delta_Y^2 \mathbb{V}(A_{t=1})}{\pi} +$$

$$\frac{(1-\mathbb{E}(A_{t=0}))\mathbb{V}(Y_{a=0}) + \mathbb{E}(A_{t=0})\mathbb{V}(Y_{a=1}) + \Delta_Y^2 \mathbb{V}(A_{t=0})}{(1-\pi)}$$

Then, by the law of large numbers, we get:

$$\lim_{n\to\infty} n\mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{SBD}}) = \frac{1}{\pi}\left((1-\mathbb{E}(A_{t=1}))\mathbb{V}(Y_{a=0}) + \mathbb{E}(A_{t=1})\mathbb{V}(Y_{a=1}) + \Delta_Y^2 \mathbb{V}(A_{t=1}) \right) + \frac{1}{1-\pi}\mathbb{V}(Y_{a=0})$$

$$\lim_{n\to\infty} n\mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{SBD}}) = \frac{1}{\pi}\left((1-\mathbb{E}(A_{t=1}))\mathbb{V}(Y_{a=0}) + \mathbb{E}(A_{t=1})\mathbb{V}(Y_{a=1}) + \Delta_Y^2 \mathbb{V}(A_{t=1}) \right) +$$

$$\frac{1}{1-\pi}\left((1-\mathbb{E}(A_{t=0}))\mathbb{V}(Y_{a=0}) + \mathbb{E}(A_{t=0})\mathbb{V}(Y_{a=1}) + \Delta_Y^2 \mathbb{V}(A_{t=0}) \right)$$

Because of the positivity assumption (Assumption 2), we have:

$$\pi < 1, \quad 1 - \pi < 1$$

Substituting above inequalities back to $\lim\limits_{n\to\infty} n\mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{SBD}})$ and $\lim\limits_{n\to\infty} n\mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{SBD}})$, we get following asymptotic variance lower bounds:

$$\lim_{n\to\infty} n\mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{SBD}}) > (2 - \mathbb{E}(A_{t=1}))\mathbb{V}(Y_{a=0}) + \mathbb{E}(A_{t=1})\mathbb{V}(Y_{a=1}) + \Delta_Y^2 \mathbb{V}(A_{t=1})$$

$$\lim_{n\to\infty} n\mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{SBD}}) > \left((1-\mathbb{E}(A_{t=1}))\mathbb{V}(Y_{a=0}) + \mathbb{E}(A_{t=1})\mathbb{V}(Y_{a=1}) + \Delta_Y^2 \mathbb{V}(A_{t=1}) \right) +$$

$$\left((1-\mathbb{E}(A_{t=0}))\mathbb{V}(Y_{a=0}) + \mathbb{E}(A_{t=0})\mathbb{V}(Y_{a=1}) + \Delta_Y^2 \mathbb{V}(A_{t=0}) \right)$$

$$= (2 - \mathbb{E}(A_{t=1}) - \mathbb{E}(A_{t=0}))\mathbb{V}(Y_{a=0}) + (\mathbb{E}(A_{t=0}) + \mathbb{E}(A_{t=1}))\mathbb{V}(Y_{a=1}) +$$

$$\Delta_Y^2 \left(\mathbb{V}(A_{t=0}) + \mathbb{V}(A_{t=1}) \right)$$

31

D.3.2. Asymptotic variance upper bounds of $\text{Var}(\hat{\Phi}^{\text{CFD}})$

Based on the derivation for $\hat{\Phi}_{\text{os}}^{\text{CFD}}$ and $\hat{\Phi}_{\text{ts}}^{\text{CFD}}$, we get following with a slight rewriting:

$$n\text{Var}(\hat{\Phi}_{\text{os}}^{\text{CFD}}) = \mathbb{V}(Y_{a=0}) \left(\mathbb{E}\left[\frac{\hat{\mathbb{E}}(A_{t=1})^2 (1-\hat{\pi})^2}{\hat{\omega}_{a=0,t=0}} \right] + \mathbb{E}\left[\frac{\hat{\mathbb{E}}(A_{t=1})^2 \hat{\pi}^2}{\hat{\omega}_{a=0,t=1}} \right] \right) + \mathbb{V}(Y_{a=1}) \mathbb{E}\left[\frac{\hat{\mathbb{E}}(A_{t=1})^2}{\hat{\omega}_{a=1,t=1}} \right] +$$

$$\mathbb{V}(A_{t=1}) \frac{\Delta_Y^2}{\pi}$$

$$n\text{Var}(\hat{\Phi}_{\text{ts}}^{\text{CFD}}) = \mathbb{V}(Y_{a=1}) \left(\mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 (1-\hat{\pi})^2}{\hat{\omega}_{a=1,t=0}} \right] + \mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 \hat{\pi}^2}{\hat{\omega}_{a=1,t=1}} \right] \right)$$

$$+ \mathbb{V}(Y_{a=0}) \left(\mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 (1-\hat{\pi})^2}{\hat{\omega}_{a=0,t=0}} \right] + \mathbb{E}\left[\frac{(\hat{\mathbb{E}}(A_{t=1}) - \hat{\mathbb{E}}(A_{t=0}))^2 \hat{\pi}^2}{\hat{\omega}_{a=0,t=1}} \right] \right)$$

$$+ \Delta_Y^2 \left(\frac{\mathbb{V}(A_{t=1})}{\pi} + \frac{\mathbb{V}(A_{t=0})}{1-\pi} \right)$$

Then, by the law of large numbers, we get:

$$\lim_{n\to\infty} n\text{Var}(\hat{\Phi}_{\text{os}}^{\text{CFD}}) = \mathbb{V}(Y_{a=0}) \left(\frac{\Delta_A^2 (1-\pi)^2}{\omega_{a=0,t=0}} + \frac{\Delta_A^2 \pi^2}{\omega_{a=0,t=1}} \right) + + \mathbb{V}(Y_{a=1}) \frac{\Delta_A^2}{\omega_{a=1,t=1}} + \mathbb{V}(A_{t=1}) \frac{\Delta_Y^2}{\pi}$$

$$\lim_{n\to\infty} n\text{Var}(\hat{\Phi}_{\text{ts}}^{\text{CFD}}) = \mathbb{V}(Y_{a=1}) \left(\frac{\Delta_A^2 (1-\pi)^2}{\omega_{a=1,t=0}} + \frac{\Delta_A^2 \pi^2}{\omega_{a=1,t=1}} \right) + \mathbb{V}(Y_{a=0}) \left(\frac{\Delta_A^2 (1-\pi)^2}{\omega_{a=0,t=0}} + \frac{\Delta_A^2 \pi^2}{\omega_{a=0,t=1}} \right) \qquad (36)$$

$$+ \Delta_Y^2 \left(\frac{\mathbb{V}(A_{t=1})}{\pi} + \frac{\mathbb{V}(A_{t=0})}{1-\pi} \right)$$

As shown in Equation 36, the asymptotic variance of CFD estimations depends on π, $1 - \pi$, $\omega_{a=0,t=0}$, $\omega_{a=1,t=0}$, $\omega_{a=0,t=1}$, and $\omega_{a=1,t=1}$ as their denominators. The larger these quantities are, the smaller the asymptotic variances are, which intuitively means more samples for each nuisance parameter will allow better estimation.

To ensure reasonable estimation quality, we define the minimum probability for the *possible* treatment assignment and intake combinations as ρ. Formally, we write:

$$\rho = \begin{cases} \min_{\substack{t',a'\in\{0,1\} \\ (a',t')\neq(1,0)}} \{P(a',t' \mid x_0)\} & \text{if one-sided non-adherence} \\ \min_{t',a'\in\{0,1\}} \{P(a',t' \mid x_0)\} & \text{if two-sided non-adherence} \end{cases} \qquad (37)$$

Substitute Equation 37 into Equation 36, we get the following asymptotic variance upper bounds:

$$\lim_{n\to\infty} n\text{Var}(\hat{\Phi}_{\text{os}}^{\text{CFD}}) \leq \frac{1}{\rho} \left(\mathbb{V}(Y_{a=0}) \left(\Delta_A^2 ((1-\pi)^2 + \pi^2) \right) + \mathbb{V}(Y_{a=1}) \Delta_A^2 + \Delta_Y^2 \mathbb{V}(A_{t=1}) \right)$$

$$\lim_{n\to\infty} n\text{Var}(\Phi_{\text{ts}}^{\text{CFD}}) \leq \frac{1}{\rho} \left((\mathbb{V}(Y_{a=1}) + \mathbb{V}(Y_{a=0})) \Delta_A^2 ((1-\pi)^2 + \pi^2) + (\mathbb{V}(A_{t=1}) + \mathbb{V}(A_{t=0})) \Delta_Y^2 \right)$$

Lastly, we simplify the supper bound by using the fact that $(1-\pi)^2 + \pi^2 < 1$ when $0 < \pi < 1$ to get the following upper bounds:

$$\lim_{n\to\infty} n\text{Var}(\hat{\Phi}_{\text{os}}^{\text{CFD}}) < \frac{1}{\rho} \left((\mathbb{V}(Y_{a=0}) + \mathbb{V}(Y_{a=1})) \Delta_A^2 + \Delta_Y^2 \mathbb{V}(A_{t=1}) \right)$$

$$\lim_{n\to\infty} n\text{Var}(\hat{\Phi}_{\text{ts}}^{\text{CFD}}) < \frac{1}{\rho} \left((\mathbb{V}(Y_{a=0}) + \mathbb{V}(Y_{a=1})) \Delta_A^2 + (\mathbb{V}(A_{t=1}) + \mathbb{V}(A_{t=0})) \Delta_Y^2 \right)$$

D.4. Compare SBD and CFD's variance bounds directly

From the previous sections, we derive SBD's variance lower bounds under one-sided and two-sided non-adherence as follows:

$$\lim_{n \to \infty} n \mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{SBD}}) > (2 - \mathbb{E}(A_{t=1})) \mathbb{V}(Y_{a=0}) + \mathbb{E}(A_{t=1}) \mathbb{V}(Y_{a=1}) + \Delta_Y^2 \mathbb{V}(A_{t=1})$$

$$\lim_{n \to \infty} n \mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{SBD}}) > (2 - \mathbb{E}(A_{t=1}) - \mathbb{E}(A_{t=0})) \mathbb{V}(Y_{a=0}) + (\mathbb{E}(A_{t=0}) + \mathbb{E}(A_{t=1})) \mathbb{V}(Y_{a=1}) +$$
$$\Delta_Y^2 (\mathbb{V}(A_{t=0}) + \mathbb{V}(A_{t=1}))$$

And CFD's variance upper bounds under one-sided and two-sided non-adherence are as follows:

$$\lim_{n \to \infty} n \mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{CFD}}) < \frac{1}{\rho} \left((\mathbb{V}(Y_{a=0}) + \mathbb{V}(Y_{a=1})) \Delta_A^2 + \Delta_Y^2 \mathbb{V}(A_{t=1}) \right)$$

$$\lim_{n \to \infty} n \mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{CFD}}) < \frac{1}{\rho} \left((\mathbb{V}(Y_{a=0}) + \mathbb{V}(Y_{a=1})) \Delta_A^2 + (\mathbb{V}(A_{t=1}) + \mathbb{V}(A_{t=0})) \Delta_Y^2 \right)$$

To simplify the comparison, we start by assuming treatment intake under different treatment assignments, and outcome under different treatment intakes have the same variance:

$$\begin{aligned} \mathbb{V}(A_{t=0}) &= \mathbb{V}(A_{t=1}) = V_A \\ \mathbb{V}(Y_{a=0}) &= \mathbb{V}(Y_{a=1}) = V_Y \end{aligned} \tag{38}$$

Which leads to the following simplified lower bounds for SBD:

$$\begin{aligned} \lim_{n \to \infty} n \mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{SBD}}) &> 2V_Y + V_A \Delta_Y^2 \\ \lim_{n \to \infty} n \mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{SBD}}) &> 2V_Y + 2V_A \Delta_Y^2 \end{aligned} \tag{39}$$

And the following simplified upper bounds for SBD:

$$\begin{aligned} \lim_{n \to \infty} n \mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{CFD}}) &< \frac{1}{\rho} \left(2V_Y \Delta_A^2 + V_A \Delta_Y^2 \right) \\ \lim_{n \to \infty} n \mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{CFD}}) &< \frac{1}{\rho} (2V_Y \Delta_A^2 + 2V_A \Delta_Y^2) \end{aligned} \tag{40}$$

By taking the difference between SBD's lower bound and CFD's upper bound, we get the following:

$$\lim_{n \to \infty} n \left(\mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{SBD}}) - \mathrm{Var}(\hat{\Phi}_{\mathrm{os}}^{\mathrm{CFD}}) \right) > 2V_Y - \frac{2}{\rho} V_Y \Delta_A^2 - \frac{1 - \rho}{\rho} V_A \Delta_Y^2$$

$$\lim_{n \to \infty} n \left(\mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{SBD}}) - \mathrm{Var}(\hat{\Phi}_{\mathrm{ts}}^{\mathrm{CFD}}) \right) > 2V_Y - \frac{2}{\rho} V_Y \Delta_A^2 - \frac{2(1 - \rho)}{\rho} V_A \Delta_Y^2$$

Finally, applying the big-O notation, we get the following asymptotic variance difference bound for both one-sided and two-sided non-adherence:

$$\lim_{n \to \infty} n \left(\mathrm{Var}(\hat{\Phi}^{\mathrm{SBD}}) - \mathrm{Var}(\hat{\Phi}^{\mathrm{CFD}}) \right) > \mathcal{O} \left(V_Y - \frac{V_Y \Delta_A^2}{\rho} - \frac{(1 - \rho) V_A \Delta_Y^2}{\rho} \right)$$

Appendix E. Variance Reduction Bound Numerical Visualization

To compare the variance bounds difference between SBD and CFD without assuming Equation 38, we consider a special case where outcome Y is binary. In such a scenario, we apply a simple trick to represent variance in terms of expectation. For instance:

$$\mathbb{V}(Y_{a=0}) = \mathbb{E}(Y_{a=0}^2) - \mathbb{E}(Y_{a=0})^2 = \mathbb{E}(Y_{a=0}) - \mathbb{E}(Y_{a=0})^2 = \mathbb{E}(Y_{a=0})(1 - \mathbb{E}(Y_{a=0}))$$

This trick allows us to replace all variance terms in the variance bounds with their corresponding expectations, effectively reducing the number of variables that our visualization needs to consider.
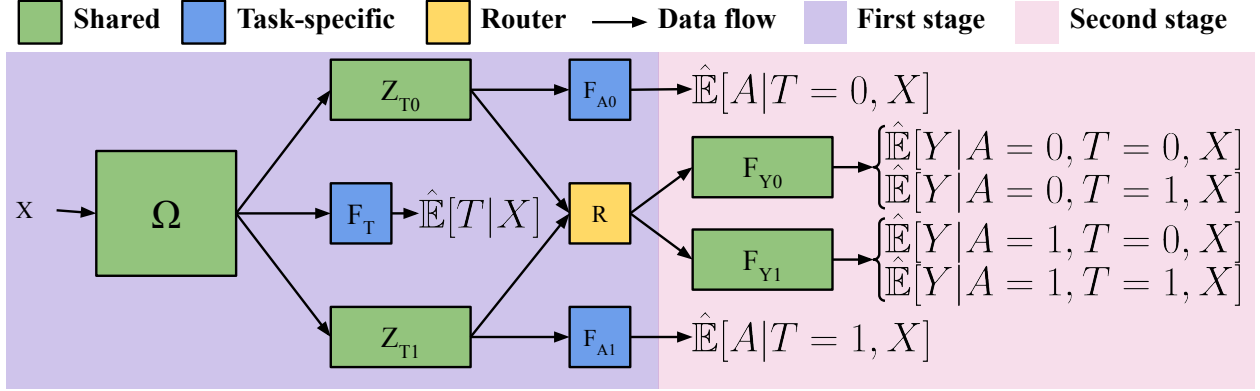
Figure 7: Illustration of our proposed LobsterNet architecture,which jointly models all required nuisance parameters

Appendix F. Implementation Details

F.1. T-learner Details

Each T-learner is implemented with two independent multi-layer perceptions (MLPs) with three hidden ELU activated layers of dimensions of 300. Only one T-learner (two MLPs) is used for the SBD-based method. Each MLP in the T-learner estimates one type of potential outcome. For the CFD-based method, seven MLPs are used (one for propensity score, two for treatment intake prediction, and four for outcome prediction).

F.2. LobsterNet Details

As shown in Figure 7, LobsterNet is composed of five shared neural network modules and three task-specific neural network modules. Each shared module is implemented using an MLP with three hidden ELU activated layers of dimensions of 300. Each task-specific module is implemented using an MLP with three hidden ELU activated layers of dimensions of 100. The router module takes in two different hidden representations from the first stage (*i.e.*, representations for predicting $\mathbb{E}[A|T=0,X]$ or $\mathbb{E}[A|T=1,X]$), and dynamically determines which representation to output for the second stage based on which outcome is being predicted. For example, when predicting $\mathbb{E}[Y|A=0,T=0,X]$ and $\mathbb{E}[Y|A=1,T=0,X]$, the router module outputs hidden representation for predicting $\mathbb{E}[A|T=0,X]$. When predicting $\mathbb{E}[Y|A=0,T=1,X]$ and $\mathbb{E}[Y|A=1,T=1,X]$, the router module outputs hidden representations for predicting $\mathbb{E}[A|T=1,X]$. This design allows for sharing the same hidden MLP representations when predicting outcomes conditioned on the same type of treatment assignment (T). Lastly, we conceptually break down LobsterNet into first stage and second stage in the illustration, however, the entire model is end-to-end trained with a multi-task objective function. This design enables efficient optimization in practice.

F.3. Training Details

80% of all data are randomly selected for training. Among all training data, another 20% is randomly selected as the validation set. All neural networks are trained using Adam optimizer (Kingma and Ba, 2015) with a fixed learning rate of 10^{-4} for up to 1000 epochs. L2 regularization is selected via grid search over $\{10^{-2}, 5 \times 10^{-3}, 10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}, 10^{-5}, 0.0\}$ by selecting the parameter resulting in the lowest validation loss. Early stop is used with a patience of 5 epochs based on the validation loss. We also applied a learning rate schedule, which reduces the learning rate by 0.5 after five epochs of non-improvement to the validation loss.

When the outcome prediction is a regression task, the α and β in LobsterNet's training objective are set to the root mean square of the outcomes in the training data to ensure a similar magnitude between regression and classification loss. The α and β are set to 1 when the outcome prediction is a classification

task. Experiments were conducted on a server with 256 CPU cores and 8 NVIDIA RTX A6000 GPUs for training and evaluation.

Appendix G. Experiments Details

G.1. Data Access

We accessed the IHDP dataset via the NPCI open-source package (https://github.com/vdorie/npci). Specifically, the IHDP data can be located under npci/examples/ihdp_sim /data/ihdp.RData in the NPCI GitHub repository. The data were accessed from the latest version of the repository, updated on Mar 16, 2022. No license is associated with this repository.

The AMR-UTI data was accessed via PhysioNet (https://physionet.org/content/antimicrobial-resistance-uti/1.0.0/). We used the 1.0.0 version of AMR-UTI data, which is shared under the PhysioNet Credentialed Health Data License 1.5.0 (https://physionet.org/content/antimicrobial-resistance-uti/view-license/1.0.0/). During our research, we followed the data use agreement required by AMR-UTI, PhysioNet Credentialed Health Data Use Agreement 1.5.0 (https://physionet.org/content/antimicrobial-resistance-uti/view-dua/1.0.0/).

Appendix H. Additional Experimental Results

H.1. Sensitivity Analysis with Synthetic Datasets

To assess the sensitivity of our synthetic experimental results (Section 6.3), we repeated our experiments with different simulation parameters. In particular, we first modified the number of features from 30 to 50. As shown in Figure 8, the overall trend in both Synthetic Dataset A and Synthetic Dataset B remains the same as Section 6.3: CFD is more advantageous than SBD when effect sizes are smaller. We also modified the random weights' sampling parameter from $\mathcal{U}(-10, 10)$ to $\mathcal{U}(-5, 5)$ and repeated the experiments. As shown in Figure 9, the overall trend remains the same, CFD is advantageous over SBD when effect sizes are smaller.

H.2. DragonNet vs. LobsterNet

We additionally compare DragonNet (Shi et al., 2019) against our proposed LobsterNet using IHDP and AMR-UTI. As shown in Figure 10, LobsterNet either matches or outperforms DragonNet in terms of CATAE estimation PEHE, despite both methods jointly modeling all of their nuisance parameters, illustrating the practical advantage of CFD over SBD.

H.3. Relative PEHE Improvement on IHDP and AMR-UTI

We also evaluated CFD methods in terms of their relative performance improvements compared to the SBD method. We measure the relative improvement with $\frac{\text{PEHE}_{\text{SBD}} - \text{PEHE}_{\text{CFD}}}{\text{PEHE}_{\text{SBD}}} \times 100$. As shown in Figure 11, across most settings, CFD methods deliver significant improvements in PEHE. In particular, under one-sided non-adherence, LobsterNetconsistently delivers close to 20% PEHE reduction across all settings.
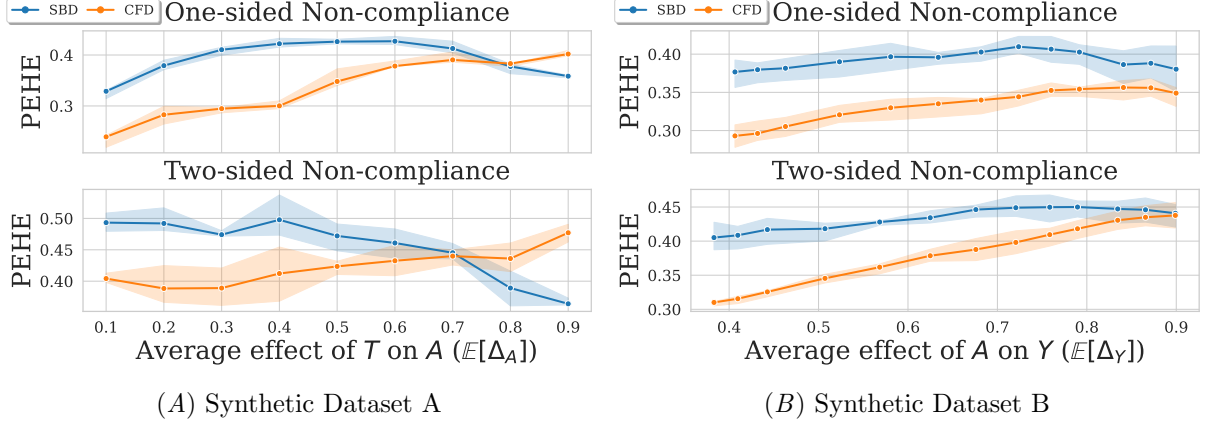
Figure 8: CATEA estimation PEHE achieved by SBD and CFD on synthetic datasets with number of features set to 50 instead of 30.
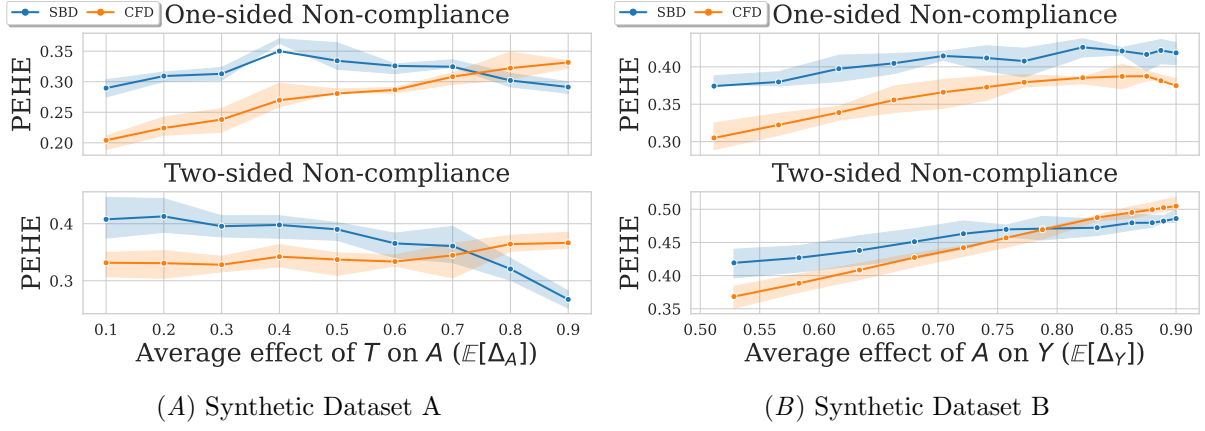

Figure 9: CATEA estimation PEHE achieved by SBD and CFD on synthetic datasets with random weights' sampling distribution changes to $\mathcal{U}(-5, 5)$ instead of $\mathcal{U}(-10, 10)$.
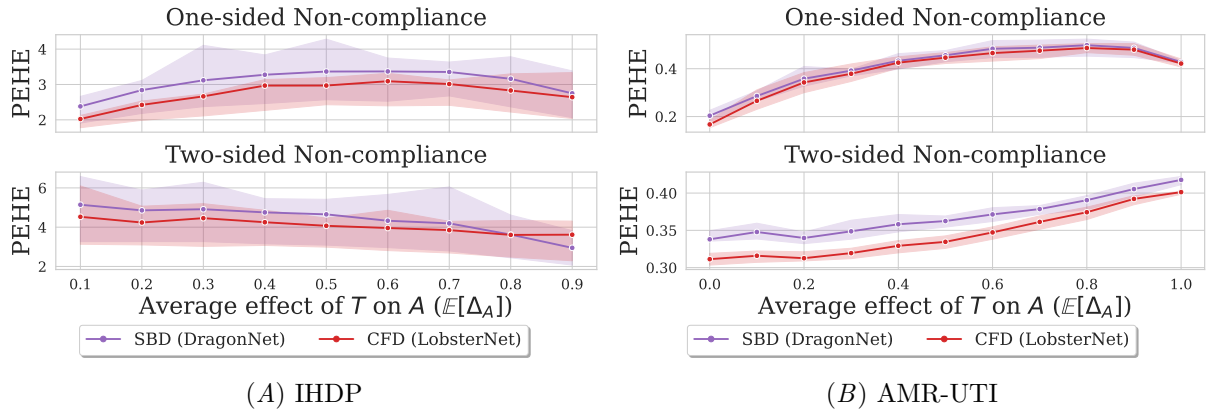

Figure 10: CATEA estimation PEHE achieved by DragonNet and LobsterNet on IHDP and AMR-UTI datasets. Across majority of the settings, LobsterNet either matches or out performs DragonNet.
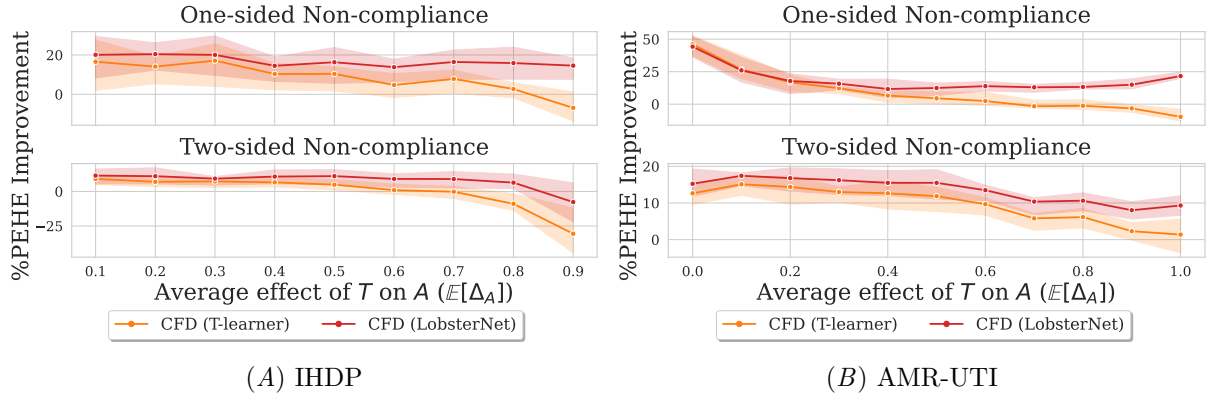
Figure 11: Relative CATEA estimation PEHE improvement achieved by CFD methods compared to SBD method on IHDP and AMR-UTI datasets. Across majority of the settings, CFD methods achieve significant amount of improvements.