TECHNICAL REPORT
R-112
May, 2024

Unified Covariate Adjustment for Causal Inference

Yonghan Jung[1], Jin Tian[2], and Elias Bareinboim[3]

[1] Purdue University
[2] Iowa State University
[3] Columbia University
jung222@purdue.edu, jtian@iastate.edu, eb@cs.columbia.edu

Abstract

Causal effect identification and estimation are fundamental tasks found throughout the data sciences. Although causal effect identification has been solved in theory, many existing estimators only address a subset of scenarios, known as the sequential back-door adjustment (SBD) (Pearl and Robins, 1995a) or g-formula (Robins, 1986). Recent efforts for developing general-purpose estimators with broader coverage, incorporating the front-door adjustment (FD) (Pearl, 2000) and others, are not scalable due to the high computational cost of summing over a high-dimensional set of variables. In this paper, we introduce a novel approach that achieves broad coverage of causal estimands beyond the SBD, incorporating various sum-product functionals like the FD, while achieving scalability – estimated in polynomial time relative to the number of variables and samples in the problem. Specifically, we present the class of *unified covariate adjustment (UCA)* for which we develop a scalable and doubly robust estimator. In particular, we illustrate the expressiveness of UCA for a wide spectrum of causal estimands (e.g., SBD, FD, and others) in causal inference. We then develop an estimator that exhibits computational efficiency and double robustness. Experiments corroborate the scalability and robustness of the proposed framework.

1 Introduction

Causal inference is a crucial aspect of scientific research, with broad applications ranging from social sciences to economics, and from biology to medicine. Two significant tasks in causal inference are causal effect identification and estimation. *Causal effect identification* concerns determining the conditions under which the causal effect can be inferred from a combination of available data distributions and a causal graph depicting the data-generating process. *Causal effect estimation*, on the other hand, develops an estimator for the identified causal effect expression using finite samples.

Causal effect identification theories have been well-established across various scenarios. These include cases where the input distribution is purely observational (Tian and Pearl, 2003; Shpitser and Pearl, 2006; Huang and Valtorta, 2006) (known as *observational identification* or obsID) or a combination of observational and interventional (Bareinboim and Pearl, 2012a; Lee et al., 2019) (referred to as *generalized identification* or gID); scenarios where the target query and input distributions originate from different populations (Bareinboim and Pearl, 2012b; Bareinboim et al., 2014; Bareinboim and Pearl, 2016; Correa et al., 2018; Lee et al., 2020) (known as *recoverability* or *transportability*); or cases where the target query is *counterfactual* (Rung 3) (Correa et al., 2021) (referred to as Ctf-ID) beyond interventional (Rung 2) of the *Ladder of Causation* (Pearl and Mackenzie, 2018; Bareinboim et al., 2020). In these situations, algorithmic solutions have been devised that take input distributions along with specified target queries and formulate identification functionals as arithmetic operations (sums/integration, products, ratios) on conditional distributions induced from input distributions.

Despite all the progress, existing estimators cover only a subset of all identification scenarios. Specifically, well-established estimators for the back-door (BD) adjustment (Pearl, 1995), represented as $\sum_z \mathbb{E}[Y \mid x, z]P(z)$, and sequential back-door adjustment (SBD) (Robins, 1986; Pearl and Robins, 1995b), are known for their robustness to the bias (Bang and Robins, 2005; Robins et al., 2009; van der Laan and Gruber, 2012; Rotnitzky et al., 2017; Luedtke et al., 2017; Díaz et al., 2023). These estimators are also *scalable*; i.e., evaluable in polynomial time relative to the number of covariates ($|Z|$) and capable in the presence of mixed discrete and continuous covariates. However, SBDs only address a fraction of the broader spectrum of identification scenarios.

Beyond SBD, recent efforts have expanded to developing estimators for the front-door (FD) adjustment $\sum_{z,x'} \mathbb{E}[Y \mid x', z]P(z \mid x)P(x')$ (Pearl, 1995). At first glance, this adjustment appears similar to SBD, as both involve the sum-product of conditional probabilities. However, FD involves treatments variables in dual roles – one being summed (x' in $\sum_{x'} \mathbb{E}[Y \mid x', z]P(x')$) and the other being fixed (x in $P(z \mid x)$). While FD estimators achieving doubly robustness have been developed (Fulcher et al., 2019; Guo et al., 2023), they lack scalability due to the necessity of summing over the values of Z (i.e., $\sum_z$), thereby limiting its practicality when Z is high-dimensional or continuous.

Similar challenges arise in more general identification scenarios beyond SBD and FD. Recent efforts have focused on developing estimators for broad causal estimands, such as *Tian's adjustment* (Tian and Pearl, 2002a), which incorporates FD and other cases where causal effects are represented as sum-product functionals (Bhattacharya et al., 2022). These efforts also include work on covering any identification functional (Jung et al., 2021a; Xia et al., 2021, 2022; Bhattacharya et al., 2022; Jung et al., 2023a). While these estimators are designed to achieve a wide coverage of functionals, they lack scalability due to the necessity of summing over high-dimensional variables.

Thus far, we have assessed the pair (functional class, estimator) based on two criteria: (1) *coverage* of the functional class, and (2) *scalability* of the corresponding estimators. Scalable estimators achieving doubly robustness have been established predominantly for BD/SBD classes. While recent studies have developed estimators with a strong emphasis on coverage (e.g., any identification functional), less attention has been given to achieving scalability.

Function class	Coverage		Scalability	
	Prior	**UCA**	**Prior**	**UCA**
BD/SBD	✓	✓	✓	✓
FD	✓	✓	✗	✓
Tian's	✓	✓	✗	✓
obsID/gID	✓	▲	✗	✓
Ctf-ID	▲	▲	?	✓
Transportability	▲	▲	?	✓

Table 1: Scope. ✓ denotes the addressed area (by UCA or prior works). ✗ denotes the unaddressed area. ▲ denotes the partially addressed area. ? indicates areas where no known results are present. Compared aspects are across back-door ('BD/SBD'), front-door ('FD'), Tian's adjustment, obsID/gID, Ctf-ID, transportability.

In this paper, we establish a novel pair of a functional class and its corresponding estimation frameworks designed to ensure scalability while covering a broad spectrum of identification functionals. Our work strives maximizing coverage such that scalable estimators with doubly robustness property can be effectively developed. This functional class, termed *unified covariate adjustment* (UCA), integrates a sum-product of conditional distributions appearing in many causal inference scenarios such as BD/FD, Tian's adjustment, S-admissibility in transportability/recoverability (Bareinboim and Pearl, 2016), effect-of-treatment-on-the-treated (ETT) (Heckman, 1992), and nested counterfactuals (Correa et al., 2021). The coverage of the proposed class is further demonstrated through the application to a novel estimand for the counterfactual directed effect (Ctf-DE) derived from fairness analysis (Plečko et al., 2024). For the proposed UCA class, we develop a scalable and doubly robust estimator evaluable computationally efficiently relative to the number of samples. Table 1 visualizes the scope of our framework. The contributions of this paper are as follows:

1. We introduce *unified covariate adjustment (UCA)*, a comprehensive framework that encompasses a broad class of sum-product causal estimands. This framework's expressiveness is demonstrated across various scenarios beyond SBD, including Tian's adjustment that incorporates FD and others as well novel counterfactual scenarios in fairness analysis.
2. We develop a corresponding estimator that is computationally efficient and doubly robust and provide its finite sample guarantee. We demonstrate scalability and robustness to bias both theoretically and empirically through simulations.

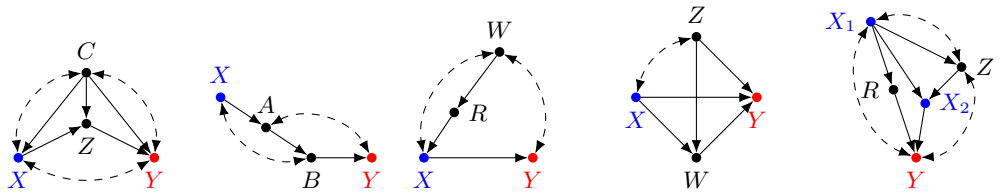


Figure 1: **(a)** Front-door in Example 1, **(b)** Verma in Example 2, **(c)** Napkin, **(d)** Standard fairness model in Example 3, and **(e)** Example graph from (Jung et al., 2021a, Fig. 1b)

Preliminaries. We use ($\mathbf{X}$, X, $\mathbf{x}$, x) to denote a random vector, variable, and their realized values, respectively. For a function $f(\mathbf{z}_i)$ for $i = 1, 2, \cdots$,, we use $\sum_i f(\mathbf{z}_i) = f(\mathbf{z}_1) + f(\mathbf{z}_2) \cdots$. Also, for a function $f(\mathbf{z})$, we use $\sum_{\mathbf{z}} f(\mathbf{z})$ to denote the summation/integration over a mixture of discrete/continuous random variables $\mathbf{Z}$. For example, we write the back-door adjustment as $\sum_{\mathbf{z}} \mathbb{E}_P[Y \mid x, \mathbf{z}]P(\mathbf{z})$ when $\mathbf{Z}$ is a mixture of discrete/continuous variables. Given an ordered set $\mathbf{X} = \{X_1, \cdots, X_n\}$, we denote $\mathbf{X}^{(i)} \coloneqq \{X_1, \cdots, X_i\}$ and $\mathbf{X}^{\geq i} \coloneqq \{X_{i+1}, \cdots, X_m\}$ for $m = |\mathbf{X}|$. For a discrete $\mathbf{X}$, we use $\mathbb{1}_{\mathbf{x}}(\mathbf{X})$ as a function such that $\mathbb{1}_{\mathbf{x}}(\mathbf{X}) = 1$ if $\mathbf{X} = \mathbf{x}$; $\mathbb{1}_{\mathbf{x}}(\mathbf{X}) = 0$ otherwise. $P(\mathbf{V})$ denotes a distribution over $\mathbf{V}$ and $P(\mathbf{v})$ as a probability at $\mathbf{V} = \mathbf{v}$, We use $\mathbb{E}_P[f(\mathbf{V})]$ and $\mathbb{V}_P[f(\mathbf{V})]$ to denote the mean and variance of $f(\mathbf{V})$ relative to $P(\mathbf{V})$. We use $\|f\|_P \coloneqq \sqrt{\mathbb{E}_P[\{f(\mathbf{V})\}^2]}$ as L2-norm of f with P. If a function $\widehat{f}$ is a consistent estimator of f having a rate r_n, we will use $\widehat{f} - f = o_P(r_n)$. We will say $\hat{f}$ is L_2-consistent if $\|\hat{f} - f\|_P = o_P(1)$. We will use $\widehat{f} - f = O_P(1)$ if $\widehat{f} - f$ is bounded in probability. Also, $\widehat{f} - f$ is said to be bounded in probability at rate r_n if $\widehat{f} - f = O_P(r_n)$. $[n] \coloneqq \{1, \cdots, n\}$ is a collection of index. $\mathcal{D} \coloneqq \{\mathbf{V}_{(i)} : i \in [n]\}$ denotes a sample set, where $\mathbf{V}_{(i)}$ denote the ith sample in $\mathcal{D}$. The empirical average of $f(\mathbf{V})$ with samples $\mathcal{D}$ is $\mathbb{E}_{\mathcal{D}}[f(\mathbf{V})] \coloneqq (1/|\mathcal{D}|) \sum_{i:\mathbf{V}_{(i)} \in \mathcal{D}} f(\mathbf{V}_{(i)})$.

Structural causal models. We use structural causal models (SCMs) (Pearl, 2000; Bareinboim et al., 2020) as our framework. An SCM $\mathcal{M}$ is a quadruple $\mathcal{M} = \langle \mathbf{U}, \mathbf{V}, P(\mathbf{U}), \mathcal{F} \rangle$, where $\mathbf{U}$ is a set of exogenous (latent) variables following a joint distribution $P(\mathbf{U})$, and $\mathbf{V}$ is a set of endogenous (observable) variables whose values are determined by functions $\mathcal{F} = \{f_{V_i}\}_{V_i \in \mathbf{V}}$ such that $V_i \leftarrow f_{V_i}(\mathbf{pa}_i, \mathbf{u}_i)$ where $\mathbf{PA}_i \subseteq V$ and $\mathbf{U}_i \subseteq \mathbf{U}$. Each SCM $\mathcal{M}$ induces a distribution $P(\mathbf{V})$ and a causal graph $\mathcal{G} = \mathcal{G}(\mathcal{M})$ over $\mathbf{V}$ in which directed edges exists from every variable in $\mathbf{PA}_i$ to V_i and dashed-bidirected arrows encode common latent variables. Performing an intervention fixing $\mathbf{X} = \mathbf{x}$ is represented through the do-operator, $\mathrm{do}(\mathbf{X} = \mathbf{x})$, which encodes the operation of replacing the original equations of X (i.e., $f_X(\mathbf{pa}_x, \mathbf{u}_x)$) by the constant x for all $X \in \mathbf{X}$ and induces an interventional distribution $P(\mathbf{V} \mid \mathrm{do}(\mathbf{x}))$. For any $\mathbf{Y} \subseteq \mathbf{V}$, the *potential response* $\mathbf{Y}_{\mathbf{x}}(\mathbf{u})$ is defined as the solution of $\mathbf{Y}$ in the submodel $\mathcal{M}_{\mathbf{x}}$ given $\mathbf{U} = \mathbf{u}$, which induces a *counterfactual variable* $\mathbf{Y}_{\mathbf{x}}$.

Related work. Our work is an extension of existing sequential back-door adjustment (SBD) estimators (Mises, 1947; Bickel et al., 1993; Bang and Robins, 2005; Robins et al., 2009; van der Laan and Gruber, 2012; Rotnitzky et al., 2017; Luedtke et al., 2017; Díaz et al., 2023) to a broader class of sum-product functionals, such as the front-door adjustment (FD) and Tian's adjustment (Tian and Pearl, 2002a) which generalizes FD and more, and nested counterfactuals, which will be detailed in later sections. Our work is aligned with recent works of Chernozhukov et al. (2022); Li and Luedtke (2023); Quintas-Martinez et al. (2024), which examined SBD derived from various joint distributions. Specifically, Li and Luedtke (2023) considered the SBD setting where conditional distributions are induced from different sources. In contrast, we study a broader class of sum-product functionals from multiple populations. Also, Quintas-Martinez et al. (2024) considered the Markovian model $\prod_{i=1}^{n} P^i(V_i \mid \mathbf{PA}_i)$ where each P^i can be distinct. In contrast, we study a broader class of estimands that are not confined to conditioning on $\mathbf{PA}_i$. On the other hands, (Chernozhukov et al., 2022) considered the case where covariate distributions are allowed to be changed, and demonstrated that FD can be captured through this technique. Our work expands on these findings by covering a broader class, such as the Tian's adjustment and a nested counterfactual in fairness literature, and by providing a more formal theory that includes finite sample guarantees and asymptotic analysis.

2 Unified Covariate Adjustment

In this section, we define a class of causal estimands termed *unified covariate adjustment (UCA)*, and exemplify its expressiveness with various causal scenarios.

Definition 1 (Unified Covariate Adjustment (UCA)). *Let $\Psi[\mathbf{P}; \boldsymbol{\sigma}]$ denote the following probability measure over an ordered set $\mathbf{V} \coloneqq (\mathbf{C}_1, \mathbf{R}_1, \cdots, \mathbf{C}_m, \mathbf{R}_m, Y \coloneqq \mathbf{C}_{m+1})$: $\Psi[\mathbf{P}; \boldsymbol{\sigma}] \coloneqq P^{m+1}(Y \mid \mathbf{S}_m) \prod_{i=1}^{m} P^i(\mathbf{C}_i \mid \mathbf{S}_{i-1}) \sigma^i_{\mathbf{R}_i}(\mathbf{R}_i \mid \mathbf{S}_i \setminus \mathbf{R}_i)$, where*

- *$\mathbf{P} \coloneqq \{P^i(\mathbf{V}) : i \in [m+1]\}$ is a set of distributions in the form of $P^i(\mathbf{V}) = Q^i(\mathbf{V} \mid \mathbf{S}^b_{i-1} = \mathbf{s}^b_{i-1})$, where Q^i is a distribution, $\mathbf{S}^b_{i-1}$ is a (potentially empty) set such that $\mathbf{S}^b_{i-1} \cap \mathbf{C}^{\geq 2} = \emptyset$. Each pairs $P^i(\mathbf{V})$ and $P^j(\mathbf{V})$ can be the same ($P^i(\mathbf{V}) = P^j(\mathbf{V})$) or distinct ($P^i(\mathbf{V}) \neq P^j(\mathbf{V})$).*
- *For $i \in [m+1]$, $\mathbf{S}_{i-1} \coloneqq (\mathbf{C}^{(i-1)} \cup \mathbf{R}^{(i-1)}) \setminus \mathbf{S}^b_{i-1}$.*
- *Each $\mathbf{R}_i$ is controlled by a pre-specified / known probability measure $\sigma^i_{\mathbf{R}_i} \coloneqq \sigma^i_{\mathbf{R}_i}(\mathbf{r}_i \mid \mathbf{s}_i \setminus \mathbf{r}_i)$ where $\sum_{\mathbf{r}_i} \sigma^i_{\mathbf{R}_i}(\mathbf{r}_i \mid \mathbf{s}_i \setminus \mathbf{r}_i) = 1$ and $0 \leq \sigma^i_{\mathbf{R}_i} \leq 1$ almost surely (e.g., $\sigma^i_{\mathbf{R}_i} \coloneqq \mathbb{1}_{\mathbf{r}_i}(\mathbf{R}_i)$).*

*Then, the expectation of Y over $\Psi[\mathbf{P}; \boldsymbol{\sigma}]$ is called a **Unified Covariate Adjustment** (UCA):*

$$\psi_0 \coloneqq \mathbb{E}_{\Psi[\mathbf{P};\boldsymbol{\sigma}]}[Y] = \sum_{\mathbf{c} \cup \mathbf{r}} \mathbb{E}_{P^{m+1}}[Y \mid \mathbf{s}_m] \prod_{i=1}^{m} P^i(\mathbf{c}_i \mid \mathbf{s}_{i-1}) \sigma^i_{\mathbf{R}_i}(\mathbf{r}_i \mid \mathbf{s}_i \setminus \mathbf{r}_i). \tag{1}$$

The *UCA-class* is a collection of causal estimands expressible in the form of a UCA. The UCA-class is a sub-class of sum-product functionals composed of $\{P^i(\mathbf{C}_i \mid \mathbf{S}_{i-1}) : P^i \in \mathbf{P}\}$, with the restriction $\mathbf{S}^b_i \cap \mathbf{C}^{\geq 2} = \emptyset$ to enable scalable computation (the reason is discussed in Appendix C.3.2).

At first glance, UCA closely resembles the sequential back-door adjustment (SBD) (Robins, 1986; Pearl and Robins, 1995b). Indeed, UCA is reduced to SBD in the special case where $P^i = P(\mathbf{V})$ for all $i = 1, \cdots, m+1$ and $\sigma^i_{\mathbf{R}_i} \coloneqq \mathbb{1}_{\mathbf{r}_i}(\mathbf{R}_i)$; i.e., $\psi_0 = \sum_{\mathbf{c}} \mathbb{E}_P[Y \mid \mathbf{c}^{(m)} \cup \mathbf{r}^{(m)})] \prod_{i=1}^{m} P(\mathbf{c}_i \mid \mathbf{c}^{(i-1)} \cup \mathbf{r}^{(i-1)})$. However, UCA provides flexibility to represent target estimands beyond SBD by allowing P^i to be any distribution that aligns with the target estimand, permitting arbitrary conditional distributions beyond the observational distribution P. To demonstrate, consider the front-door adjustment (FD) scenario (Pearl, 1995) depicted in Fig. 1a:

$$\mathbb{E}[Y \mid \mathrm{do}(x)] = \sum_{c,z,x'} \mathbb{E}[Y \mid c, x', z] P(z \mid c, x) P(c, x'). \tag{2}$$

Even though FD cannot be expressed using SBD because the treatment variable X is being fixed (in $P(z \mid c, x)$) and summed (with $\sum_{x'}$) simultaneously, it can be represented through UCA as follows:

Example 1 (FD as UCA). *FD can be written as the expectation of Y over $P(Y \mid Z, X, C)P(Z \mid x, C)P(X, C)$. We set $\mathbf{C}_1 \coloneqq \{X, C\}$, $\mathbf{C}_2 \coloneqq \{Z\}$, $\mathbf{R} = \emptyset$, $P^1(\mathbf{C}_1) = P(X, C)$, $P^2(\mathbf{C}_2 \mid \mathbf{S}_1) = P(Z \mid x, C)$ with $\mathbf{S}^b_1 = \{X\}$, $\mathbf{S}_1 = \{C\}$, and $P^3(Y \mid \mathbf{S}_3) = P(Y \mid Z, X, C)$ with $\mathbf{S}_2 = \{Z, X, C\}$.*

Next, consider *Verma's equation* (Verma and Pearl, 1990; Tian and Pearl, 2002b) with Fig. 1b:

$$\mathbb{E}[Y \mid \mathrm{do}(x)] = \sum_{b,a,x'} \mathbb{E}[Y \mid b, a, x] P(b \mid a, x') P(a \mid x) P(x'), \tag{3}$$

where X is fixed to x in $\mathbb{E}[Y \mid x, a, b]$ and $P(a \mid x)$ while summed in $P(b \mid a, x')$ and $P(x')$. Similar to FD, due to the dual role of X, the existing SBD framework is not suitable to express Verma's equation, which can be represented through UCA as follows:

Example 2 (Verma as UCA). *Verma's equation is expressible as the expectation of Y over $P(Y \mid B, A, x)P(B \mid A, X)P(A \mid x)P(X)$. We set $\mathbf{C}_1 = \{X\}$, $\mathbf{C}_2 = \{A\}$, $\mathbf{C}_3 = \{B\}$, and $\mathbf{R} = \emptyset$. We map $P^1(\mathbf{C}_1) \coloneqq P(X)$, $P^2(\mathbf{C}_2 \mid \mathbf{S}_1) = P(A \mid x)$ with $\mathbf{S}_1 = \emptyset$, $\mathbf{S}^b_1 = \{X\}$, $P^3(\mathbf{C}_3 \mid \mathbf{S}_2) = P(B \mid A, X)$ with $\mathbf{S}_2 = \{A, X\}$, and $P^4(Y \mid \mathbf{S}_3) = P(Y \mid B, A, x)$ with $\mathbf{S}_3 = \{B, A\}$, $\mathbf{S}^b_3 = \{X\}$.*

In both examples, a variable $\mathbf{S}^b_i = \{X\}$ is *bifurcated*, fixed in some conditional distributions (e.g., $P(z \mid x, c)$ in FD) and summed with $\sum_{x'}$ in others (e.g., $P(y \mid z, x', c)$ in FD). These FD and

Algorithm 1: `Tian-to-UCA`$(\mathcal{G}, \mathbf{V} \coloneqq (V_1, \cdots, V_K, Y))$

Set $\mathbf{C}_1 \coloneqq (V_1, \cdots, V_{k-1}, X)$, where $(V_1, \cdots, V_{k-1})$ are predecessors of X. If a node V_{k+1} located right next to X is in $\mathbf{S}_X$ (i.e., $V_{k+1} \in \mathbf{S}_X$), append $\{V_{k+1}, \cdots, V_{k+i_1}\}$ to this $\mathbf{C}_1$, where i_1 is an index such that $\{V_{k+1}, \cdots, V_{k+i_1}\} \subseteq \mathbf{S}_X$.
Set $P^1(\mathbf{C}_1) \coloneqq P(\mathbf{C}_1)$, $i \coloneqq 2$ and $\mathbf{R} \coloneqq \emptyset$.
while $\mathbf{V} \setminus (\{Y\} \cup \mathbf{C}^{(i-1)}) \neq \emptyset$ **do**
  if $\mathbf{C}_{i-1} \subseteq \mathbf{S}_X$, set $\mathbf{C}_i$ as the next series of vertices in $\mathbf{V} \setminus (\{Y\} \cup \mathbf{C}^{(i-1)})$ such that $\mathbf{C}_i \not\subseteq \mathbf{S}_X$; $\mathbf{S}_{i-1} \coloneqq \mathbf{C}^{(i-1)} \setminus \{X\}$; and $P^i(\mathbf{C}_i \mid \mathbf{S}_{i-1}) \coloneqq P(\mathbf{C}_i \mid \mathbf{S}_{i-1}, x)$ with $\mathbf{S}^b_{i-1} \coloneqq \{X\}$.
  else, set $\mathbf{C}_i$ as the next series of vertices in $\mathbf{V} \setminus (\{Y\} \cup \mathbf{C}^{(i-1)})$ such that $\mathbf{C}_i \subseteq \mathbf{S}_X$; $\mathbf{S}_{i-1} \coloneqq \mathbf{C}^{(i-1)}$; and $P^i(\mathbf{C}_i \mid \mathbf{S}_{i-1}) \coloneqq P(\mathbf{C}_i \mid \mathbf{S}_{i-1})$ with $\mathbf{S}^b_{i-1} \coloneqq \emptyset$.
  $i \leftarrow i + 1$.
end
Set $m \leftarrow i$. If $Y \in \mathbf{S}_X$, set $\mathbf{S}_m \coloneqq \mathbf{C}^{(m)}$, $\mathbf{S}^b_m = \emptyset$, and $P^{m+1}(Y \mid \mathbf{S}_m) = P(Y \mid \mathbf{S}_m)$. Otherwise, set $\mathbf{S}_m \coloneqq \mathbf{C}^{(m)} \setminus \{X\}$, $\mathbf{S}^b_m = \{X\}$, and $P^{m+1}(Y \mid \mathbf{S}_m) = P(Y \mid \mathbf{S}_m, x)$.
return $\mathbb{E}_{\Psi[\mathbf{P}]}[Y]$ *where* $\Psi[\mathbf{P}] \coloneqq P^{m+1}(Y \mid \mathbf{S}_m) \prod_{i=1}^{m} P^i(\mathbf{C}_i \mid \mathbf{S}_{i-1})$.

Verma's equations are special cases of *Tian's adjustment* (Tian and Pearl, 2002a), which states that $\mathbb{E}[Y \mid \mathrm{do}(x)]$ is identifiable if X and its children $\mathrm{ch}_{\mathcal{G}}(X)$ in the graph $\mathcal{G}$ are not connected by bidirected edges:

$$\mathbb{E}[Y \mid \mathrm{do}(x)] = \sum_{\mathbf{v} \setminus xy} \sum_{x'} \mathbb{E}_{P'}[Y \mid \mathbf{v}^{(K)}] \prod_{i=1}^{K} P'(v_i \mid \mathbf{v}^{(i-1)}), \tag{4}$$

where $\mathbf{V} \coloneqq (V_1, V_2, \cdots, V_K, Y)$ is a topologically ordered set with $V_k \coloneqq X$ for some k being the treatment variable X, $P'(v_i \mid \mathbf{v}^{(i-1)}) \coloneqq P(v_i \mid \mathbf{v}^{(k-1)}, x, v_{k+1}, \cdots, v_{i-1})$ (i.e., X is fixed to x) if $V_i \notin \mathbf{S}_X$ where $\mathbf{S}_X$ is the set of vertices connected with X through bidirected edges, and $P'(v_i \mid \mathbf{v}^{(i-1)}) \coloneqq P(v_i \mid \mathbf{v}^{(k-1)}, x', v_{k+1}, \cdots, v_{i-1})$ (i.e., X is summed with $\sum_{x'}$) if $V_i \in \mathbf{S}_X$. In Tian's adjustment, X is bifurcated into *summed* through $\sum_{x'}$ and *fixed* to $X = x$. We exhibit the expressiveness of UCA for Tian's adjustment:

Proposition 1. *Tian's adjustment in Eq. (4) is UCA-expressible through Algo. 1.*

Next, we exhibit the coverage of the UCA for a counterfactual quantity in the fairness literature. Specifically, we focus on the counterfactual directed effect (Ctf-DE) in the *Standard fairness model* (SFM) (Plečko et al., 2024), as illustrated in Fig. 1d. This model includes several key components: the protected (discrete) attribute (X), such as race; the baseline covariates (Z), like age; the mediator variables (W) affected by X, for example, educational level; and the outcome variable (Y), such as salary. Consider a scenario where we investigate the the query, "*What would be the expected salary for someone who is Black, but hypothetically of Asian race and had been educated as a White person typically would be?*". The query is represented as Ctf-DE: $\mathbb{E}[Y_{X=x_0, W_{X=x_1}} \mid X = x_2]$, where x_0, x_1, and x_2 correspond to the races Asian, White, and Black, respectively. This query can be identified through the algorithm in (Correa et al., 2021) under the SFM in Fig. 1d:

$$\mathbb{E}[Y_{X=x_0, W_{X=x_1}} \mid X = x_2] = \sum_{w,z} \mathbb{E}[Y \mid X = x_0, w, z] P(w \mid X = x_1, z) P(z \mid X = x_2). \tag{5}$$

This identification functional is UCA-expressible:

Example 3 (Ctf-DE as UCA). *The Ctf-DE is expressible through the expectation of Y over $P(Y \mid X = x_0, W, Z) P(W \mid X, Z) P(Z \mid X = x_2) \mathbb{1}_{x_1}(X)$. Set $\mathbf{R}_1 \coloneqq \{X\}$, $\sigma^1_{\mathbf{R}_1} \coloneqq \mathbb{1}_{x_1}(X)$, $P^1(\mathbf{C}_1) = P(Z \mid X = x_2)$ with $\mathbf{C}_1 = \{Z\}$ and $\mathbf{S}^b_0 = \{X\}$, $P^2(\mathbf{C}_2 \mid \mathbf{S}_1) = P(W \mid X, Z)$ with $\mathbf{C}_2 = \{W\}$ and $\mathbf{S}_1 = \{X, Z\}$, $P^3(Y \mid \mathbf{S}_2) = P(Y \mid X = x_0, W, Z)$ with $\mathbf{S}_2 = \{W, Z\}$ and $\mathbf{S}^b_2 = \{X\}$.*

Beyond Tian's adjustment and Ctf-DE, more examples estimands, including S-admissibility in transportability (Bareinboim and Pearl, 2016), effect-of-treatment-on-the-treated (ETT) (Heckman, 1991), off-policy evaluation (Precup, 2000), and a fusion of multiple experimental studies (Jung et al. (2023b,a)), can be expressed through UCA. Detailed examples are provided in Appendix B.

Clearly not all causal estimand functionals are UCA-expressible. To witness, consider the 'napkin' estimand described in (Pearl and Mackenzie, 2018; Jung et al., 2021a) with $\mathcal{G}$ in Fig. 1c, defined as

$P(y \mid \mathrm{do}(x)) = \frac{\sum_w P(y,x|r,w)P(w)}{\sum_w P(x|r,w)P(w)}$. Here, the functional for $\mathbb{E}[Y \mid \mathrm{do}(x)]$ is represented not as the expectation of a product of conditional distributions, but rather as a quotient of sums of conditional distributions. The napkin estimand is not UCA-expressible. We provide a detailed analysis on the cases where a target estimand is not UCA-expressible in Appendix C.3.

3 Scalable Estimator for Unified Covariate Adjustment

So far, we discussed the *coverage* of UCA. In this section, we construct a *scalable* estimator for UCA that achieves doubly robustness property and provide its finite sample guarantee. We define the estimator with two sets of nuisance parameters $\boldsymbol{\mu}$ and $\boldsymbol{\pi}$. $\boldsymbol{\mu}$ is a collection of regression parameters, and $\boldsymbol{\pi}$ is a collection of ratio parameters.

To define the regression nuisances, we first define some sets. Set $\mathbf{B}_i \coloneqq \mathbf{S}_{i-1}^b \cap \mathbf{C}_1$ for $i = 2, \cdots, m+1$ and $\mathbf{B}_1 \coloneqq \emptyset$. Define $\check{\mathbf{S}}_i \coloneqq \mathbf{B}_i' \cup (\mathbf{S}_i \setminus (\mathbf{R}_i \cup \mathbf{B}_i))$, where $\mathbf{B}_i'$ is a copy of $\mathbf{B}_i$ (variables following the same distribution as $\mathbf{B}_i$ but independent of $\mathbf{B}_i$). Set $\check{\mu}_0^{m+1} \coloneqq Y$. For $i = m, \cdots, 1$, regression nuisances are defined as follows:

$$\mu_0^i(\mathbf{S}_i) \coloneqq \mathbb{E}_{P^{i+1}}[\check{\mu}_0^{i+1}(\check{\mathbf{S}}_{i+1}) \mid \mathbf{S}_i] \tag{6}$$

$$\check{\mu}_0^i(\check{\mathbf{S}}_i) \coloneqq \sum_{\mathbf{r}_i} \mu_0^i(\check{\mathbf{S}}_i, \mathbf{r}_i)\sigma_{\mathbf{R}_i}^i(\mathbf{r}_i \mid \check{\mathbf{S}}_i). \tag{7}$$

Equipped with the regression nuisances, UCA can be computed as follows:

Proposition 2. *UCA in Eq. (1) can be parameterized as* $\psi_0 = \mathbb{E}_{P^1}[\check{\mu}_0^1(\check{\mathbf{S}}_1)]$.

Whenever no variables are being summed and fixed simultaneously (i.e., $\mathbf{B}_i = \emptyset$) in the UCA functional, as in Eq. (5) in Ctf-DE, we can estimate $\boldsymbol{\mu}$ through nested regression methods with off-the-shelf regression models and compute UCA in Eq. (1) as $\psi_0 = \mathbb{E}_{P^1}[\check{\mu}_0^1(\check{\mathbf{S}}_1)]$. This approach aligns with existing SBD estimators (Bang and Robins, 2005; Robins et al., 2009; van der Laan and Gruber, 2012; Rotnitzky et al., 2017; Luedtke et al., 2017; Díaz et al., 2023). For instance, in Ctf-DE in Example 3, $\mu_0^2(W,Z) \coloneqq \mathbb{E}_P[Y \mid W, Z, x_0]$, $\check{\mu}_0^2(W,Z) = \mu_0^2(W,Z)$, $\mu_0^1(X,Z) \coloneqq \mathbb{E}_P[\check{\mu}_0^2(W,Z) \mid X, Z]$, $\check{\mu}_0^1(Z) = \mu_0^1(x_1, Z)$, and $\psi_0 = \mathbb{E}_P[\check{\mu}_0^1(Z) \mid x_2]$. These nuisances can be estimated efficiently with regression models run in polynomial time relative to the number of variables and samples (e.g., neural networks (LeCun et al., 2015) or XGBoost (Chen and Guestrin, 2016)).

Going beyond the SBD framework, regression nuisances in Eq. (6) are capable of representing functionals in the presence of variables being summed and fixed simultaneously (e.g., FD in Eq. (2) or Verma in Eq. (3)). For example, consider FD in Eq. (2) with its UCA representation in Example 1. First, we have $\mu_0^2(Z,X,C) = \mathbb{E}_P[Y \mid Z, X, C]$ with $\mathbf{S}_2 = \{Z, X, C\}$. Next, we have $\mathbf{B}_2 = \mathbf{S}_1^b \cap \mathbf{C}_1 = \{X\}$ and, therefore, $\check{\mathbf{S}}_2 = \{Z, X', C\}$, where X' is an independent copy of X. Then, we have $\check{\mu}_0^2(Z,X',C) = \mu_0^2(Z,X',C)$. Next, $\mu_0^1(C) = \mathbb{E}_P[\check{\mu}_0^2(Z,X',C) \mid x, C]$. Finally, we have $\mu_0^1(C) = \check{\mu}_0^1(C)$ since $\check{\mathbf{S}}_1 = \mathbf{S}_1 = \{C\}$. We witness that $\mathbb{E}_P[\check{\mu}_0^1(C)]$ correctly specified FD in Eq. (2) since $\mathbb{E}_P[\check{\mu}_0^1(C)] = \mathbb{E}_P[\mathbb{E}_P[\mathbb{E}_P[Y \mid Z, X', C] \mid x, C]]$ where X' is an independent copy of X.

Evaluating regression nuisances may need $\mathbf{B}_i'$ as an independent copy of $\mathbf{B}_i$. Empirically, generating $\mathbf{B}_i'$ involves permuting copied samples of $\mathbf{B}_i$, an approach used in recent works in (Chernozhukov et al., 2022; Xu and Gretton, 2022). We formally name this approach *empirical bifurcation*:

Definition 2 (Empirical bifurcation). *An **empirical bifurcation** for* $\mathbf{B}$ *following a distribution* P *is the procedure of copying samples of* $\mathbf{B} \sim P$ *and randomly permuting to obtain new samples* $\mathbf{B}'$.

In general, the regression nuisances can be estimated from data by employing empirical bifurcation and off-the-shelf regression models.

Next, we define the ratio nuisance parameters $\boldsymbol{\pi}$. For $i \in [m]$, $\pi_0^i(\mathbf{C}^{(i)} \cup \mathbf{R}^{(i)})$ is defined as the solution to the following equation:

$$\pi_0^i(\mathbf{C}^{(i)} \cup \mathbf{R}^{(i)}) : \text{ such that } \mathbb{E}_{P^{i+1}}[\pi_0^i(\mathbf{C}^{(i)} \cup \mathbf{R}^{(i)})\mu_0^i(\mathbf{S}_i)] = \psi_0. \tag{8}$$

For the example of FD in Example 1, we have $\mathbb{E}_P[\pi_0^2(Z,X,C)\mu_0^2(Z,X,C)] = \sum_{z,x',c} \mathbb{E}_P[Y \mid z, x', c]P(z,x',c)\pi_0^2(z,x',c)$. To match this functional to FD adjustment in Eq. (2), we have

Algorithm 2: DML-UCA($\{\mathcal{D}^i\}, L$)

1 **Sample split**: $\forall i \in [m+1]$, randomly split $\mathcal{D}^i \overset{iid}{\sim} P^i$ into L-fold. $\mathcal{D}^i_\ell$ is the ℓ-th partition, and $\mathcal{D}^i_{-\ell} := \mathcal{D}^i \setminus \mathcal{D}^i_\ell$.

2 **Learning $\hat{\boldsymbol{\mu}}$**: For $i = m, \cdots, 1$, learn a function $\hat{\mu}^i_\ell(\mathbf{S}_i)$ by regressing $\hat{\check{\mu}}^{i+1}_\ell(\check{\mathbf{S}}_{i+1})$ evaluated from $\{\mathcal{D}^j_{-\ell} : j \in [m+1]\}$ with Def. 2 onto samples of $\mathbf{S}_i$ in $\mathcal{D}^{i+1}_{-\ell}$.

3 **Learning $\hat{\boldsymbol{\pi}}$**: For $i \in [m]$, learn $\hat{\pi}^i_\ell(\mathbf{C}^{(i)} \cup \mathbf{R}^{(i)})$ using $\{\mathcal{D}^j_{-\ell} : j \in [i+1]\}$ through Eqs. (8, 9)

4 Return DML-UCA estimator $\hat{\psi}$:

$$\frac{1}{L}\sum_{\ell=1}^{L}\sum_{i=1}^{m}\mathbb{E}_{\mathcal{D}^{i+1}_\ell}[\hat{\pi}^i_\ell\{\hat{\mu}^{i+1}_\ell - \hat{\mu}^i_\ell\}] + \mathbb{E}_{\mathcal{D}^1_\ell}[\hat{\mu}^1_\ell]. \tag{10}$$

$\pi^2_0(Z,X,C) = \frac{P(Z|x,C)}{P(Z|X,C)}$. The closed-form solution for Eq. (8), with $\mathbf{S}^{\pi_i}_j := (\mathbf{S}_j \setminus \mathbf{B}_{i+1}) \cup \mathbf{B}'_{i+1}$, is

$$\pi^i_0(\mathbf{C}^{(i)} \cup \mathbf{R}^{(i)}) = \frac{P^1(\mathbf{C}_1 \setminus \mathbf{B}^{\geq i+1})\prod_{j=2}^{i} P^j(\mathbf{C}_j \mid \mathbf{S}^{\pi_i}_{j-1})\sigma^j_{\mathbf{R}_j}(\mathbf{R}_j \mid \mathbf{S}^{\pi_i}_j \setminus \mathbf{R}_j)}{P^{i+1}(\mathbf{S}_i)}. \tag{9}$$

For the example of FD, we have $\pi^1_0 = \frac{P^1(\mathbf{C}_1 \setminus \mathbf{B}^{\geq 2})}{P^2(\mathbf{S}_1)} = \frac{P(C)}{P(C|x)} = \frac{P(x)}{P(x|C)}$ with $\mathbf{B}_2 = \{X\}$. Equipped with the ratio nuisances, UCA can be computed as follows:

Proposition 3. *UCA in Eq. (1) can be parameterized as* $\psi_0 = \mathbb{E}_{P^{m+1}}[\pi^m_0(\mathbf{C}^{(m)} \cup \mathbf{R}^{(m)})Y]$.

Estimating the ratio nuisances may be challenging due to the distribution ratio of continuous/high-dimensional variables. To address the challenge, we use Bayes' rule to transform the distribution ratio into a more tractable form. For example, in FD, if the treatment X is a singleton binary, instead of estimating $\pi^2_0 = \frac{P(Z|x,C)}{P(Z|X,C)}$, an equivalent estimand $\pi^2_0 = \frac{P(x|Z,C)P(X|C)}{P(X|Z,C)P(x|C)}$ can be estimated. This approach allows to use off-the-shelf probabilistic classification methods for estimating distribution ratios, allowing scalable computation. A detailed procedure for ratio estimation is in Appendix C.2.

Combining regression and ratio-nuisances, we present a double/debiased machine learning (DML) (Chernozhukov et al., 2018)-based estimator $\hat{\psi}$ for the UCA, titled 'DML-UCA', in Algo. 2. We provide detailed nuisance specification for various examples in Appendix A and B.

DML-UCA provides a scalable estimator for functionals within UCA class. When the target query is BD/SBD, DML-UCA aligns with existing scalable SBD estimators (Bang and Robins, 2005; Robins et al., 2009; van der Laan and Gruber, 2012; Rotnitzky et al., 2017; Luedtke et al., 2017; Díaz et al., 2023). Going beyond SBD, DML-UCA can be estimated through nested regressions, estimating distribution ratios, and empirical bifurcation, which can be conducted in polynomial time relative to the number of variables and samples, ensuring the scalability of DML-UCA:

Theorem 1 (Scalability). *Algo. 2 runs in* $O(m \times \{n_{\max} + L \times (T_{\boldsymbol{\mu}} + T_{\boldsymbol{\pi}})\})$ *time, where* $n_{\max} := \max\{|\mathcal{D}^i| : i \in [m+1]\}$, $T_{\boldsymbol{\mu}} := \max\{T_{\hat{\mu}^i_\ell} : i \in [m], \ell \in [L]\}$, $T_{\boldsymbol{\pi}} := \max\{T_{\hat{\pi}^i_\ell} : i \in [m], \ell \in [L]\}$, *and* $T_{\hat{\mu}^i_\ell}$ *and* $T_{\hat{\pi}^i_\ell}$ *denote the time complexity for learning and evaluating* $\hat{\pi}^i_\ell$ *and* $\hat{\mu}^i_\ell$, *respectively.*

An an example, for XGBoost (Chen and Guestrin, 2016), $T_{\boldsymbol{\pi}} = T_{\boldsymbol{\mu}} = O(\text{num}_{\text{tree}} \times \text{depth}_{\text{tree}} \times n_{\max} \log n_{\max})$, where num_{tree} and $\text{depth}_{\text{tree}}$ are the number and depth of trees in XGBoost.

DML-UCA is a novel estimator achieving coverage beyond SBD and also scalability, as illustrated in Table 1. Existing estimators beyond SBD often lack scalability. For instance, existing FD estimators (Fulcher et al., 2019; Guo et al., 2023) face exponential time complexity in the dimension of mediators. This issue also plagues the estimators in (Bhattacharya et al., 2022; Jung et al., 2021b) which cover Tian's adjustment and any identification functionals. In contrast, DML-UCA's polynomial time complexity positions it as a uniquely scalable solution within the UCA functional class, which includes FD and Tian's adjustment as special cases.

3.1 Error analysis

In this section, we show that DML-UCA exhibits doubly robustness, in addition to scalability. Since UCA is composed of multiple (possibly distinct) distributions, we provide a tool to distinguish them.

Definition 3 (**Index set**). *The index sets $\mathcal{I}_1, \cdots, \mathcal{I}_K$ partition $\{1, \cdots, m+1\}$ such that indices i and j are in the same set $\mathcal{I}_k$ if and only if $P^i(\mathbf{V}) = P^j(\mathbf{V})$.*

We will use $\mathtt{P}^k$ for $k = 1, \cdots, K$ to denote the distribution P^i for $i \in \mathcal{I}_k$. Then,

$$\Psi[\mathbf{P}; \boldsymbol{\sigma}] = \Psi[\{\mathtt{P}^k : k = 1, \cdots, K\}; \boldsymbol{\sigma}]. \tag{11}$$

Since multiple distributions are involved in UCA, deriving an influence function for each distribution $\mathtt{P}^k$ becomes necessary. A standard influence functions is typically defined for a single distribution P, and thus, does not suffice for studying multi-distribution setting. To address the issue, we employ a *partial influence function* (PIF) (Pires and Branco, 2002), an influence function defined relative to each $\mathtt{P}^k$. A formal definition is in Appendix C. For UCA, PIFs are given as follows:

Theorem 2 (**PIF for UCA**). *Assume that $\mu_0^i < \infty$ and $0 < \pi_0^i < \infty$ almost surely for $i = 1, \cdots, m$. Define $\eta_0^1 := \{\mu_0^1\}$ and $\eta_0^i := \{\pi_0^{i-1}, \mu_0^i, \mu_0^{i-1}\}$ for $i = 1, \cdots, m+1$, and*

$$\varphi^i(\mathbf{S}^i; \eta_0^i, \psi_0) := \begin{cases} \pi_0^{i-1}\{\breve{\mu}_0^i - \mu_0^{i-1}\} & \text{if } i > 1 \\ \breve{\mu}_0^1 - \psi_0 & \text{if } i = 1, \end{cases} \tag{12}$$

where $\mathbf{S}^i := (\mathbf{C}^{(i-1)} \cup \mathbf{R}^{(i-1)})$. Let $\mathbf{V}^k := \cup_{i \in \mathcal{I}_k} \mathbf{S}^i$ and $\boldsymbol{\eta}_0^k := \cup_{i \in \mathcal{I}_k} \eta_0^i$. Then, the k-th PIF for UCA is $\phi_0^k := \phi^k(\mathbf{V}^k; \boldsymbol{\eta}_0^k, \psi_0) := \sum_{i \in \mathcal{I}_k} \varphi^i(\mathbf{S}^i; \eta_0^i, \psi_0)$.

Equipped with PIFs, we provide a finite-sample guarantee for DML-UCA, extending Chernozhukov et al. (2023) which analyzed DML estimators for BDs.

Theorem 3 (**Finite sample guarantee**). *Assume that $\mu_0^i, \hat{\mu}_\ell^i < \infty$ and $0 < \pi_0^i, \hat{\pi}_\ell^i < \infty$ almost surely for $i = 1, \cdots, m$. Suppose the third moment of ϕ_0^k for $k = 1, \cdots, K$ exist. Let $\phi_0^k := \phi^k(\mathbf{V}^k; \boldsymbol{\eta}_0^k, \psi_0)$ and $\hat{\phi}_\ell^k := \phi^k(\mathbf{V}^k; \hat{\boldsymbol{\eta}}_\ell^k, \psi_0)$. Then, the error is $\hat{\psi} - \psi_0 = \sum_{k=1}^K R_1^k + \frac{1}{L}\sum_{\ell=1}^L R_2^\ell$, where $R_1^k := (1/L)\sum_{\ell=1}^L (\mathbb{E}_{\mathcal{D}^k}[\hat{\phi}_\ell^k] - \mathbb{E}_{\mathtt{P}^k}[\hat{\phi}_\ell^k])$, and*

$$R_2^\ell := \sum_{i=1}^m \mathbb{E}_{P^{i+1}}[(\hat{\pi}_\ell^i - \pi_0^i)(\mu_0^i - \hat{\mu}_\ell^i)] + \sum_{i=2}^m \mathbb{E}_{P^{i+1}}[(\pi_0^i/\pi_0^{i-1})(\hat{\pi}_\ell^{i-1} - \pi_0^{i-1})(\mu_0^i - \hat{\mu}_\ell^i)],$$

and, with probability greater than $1 - \epsilon$, the difference between the cumulative distribution function (CDF) of R_1^k and the standard normal CDF $\texttt{NORMAL}(x)$ is upper bounded as follows:

$$\left| \mathtt{P}^k \left(\frac{\sqrt{|\mathcal{D}^k|}}{\rho_{k,0}} R_1^k < x \right) - \texttt{NORMAL}(x) \right| \leq \frac{1}{\sqrt{2\pi}} \sqrt{\frac{L^2}{\epsilon} \sum_{\ell=1}^L \frac{\|\hat{\phi}_\ell^k - \phi_0^k\|_{\mathtt{P}^k}^2}{|\mathcal{D}_\ell^k|}} + \frac{0.4748\kappa_0^3}{\rho_{k,0}^3 \sqrt{|\mathcal{D}^k|}}, \tag{13}$$

where $\rho_{k,0}^2 := \mathbb{V}_{\mathtt{P}^k}[\phi_0^k]$ and $\kappa_0^3 := \mathbb{E}_{\mathtt{P}^k}[|\phi_0^k|^3]$.

This is a novel finite sample guarantee of DML-based estimators for functionals beyond SBD. For example, only asymptotic analyses were provided for FD (Fulcher et al., 2019; Guo et al., 2023), Tian's adjustment (Bhattacharya et al., 2022), and obsID (Jung et al., 2021b). Thm. 3 elucidates that the error can be decomposed into two terms R_1^k and R_2^ℓ. The term R_1^k closely approximates a standard normal distribution variable, and R_2^ℓ, comprises the error of $(\hat{\pi}_\ell^i, \hat{\pi}_\ell^{i-1})$ and $\hat{\mu}^i$, exhibiting doubly-robustness behavior. Specifically, if the nuisance parameters $\hat{\mu}_\ell^i$, $\hat{\pi}_\ell^i$, and $\hat{\pi}_\ell^{i-1}$ converge at a rate of $n^{-1/4}$ (where n represents the size of the smallest sample set), then DML-UCA converges at a faster rate of $n^{-1/2}$. This point becomes evident in the corresponding asymptotic analysis:

Corollary 3 (**Asymptotic error**). *Assume $\mu_0^i, \hat{\mu}_0^i < \infty$ and $0 < \pi_0^i, \hat{\pi}_0^i < \infty$ almost surely. Suppose the third moment of ϕ_0^k exist. Suppose $\hat{\breve{\mu}}_\ell^1$ and $\hat{\pi}_\ell^i\{\hat{\breve{\mu}}_\ell^{i+1} - \hat{\mu}_\ell^i\}$ are L_2-consistent. Then, $\hat{\psi} - \psi_0 = \sum_{k=1}^K R_1^k + \frac{1}{L}\sum_{\ell=1}^L \sum_{i=1}^m O_{P^{i+1}}\left(\|\hat{\mu}_\ell^i - \mu_0^i\|(\|\hat{\pi}_\ell^i - \pi_0^i\| + \|\hat{\pi}_\ell^{i-1} - \pi_0^{i-1}\|)\right)$ where R_1^k is a random variable such that $\sqrt{|\mathcal{D}^k|}R_1^k$ converges in distribution to normal$(0, \rho_{k,0}^2)$.*

4 Experiments

In this section, we demonstrate the *scalability* and *doubly robustness* of the DML-UCA estimator, where nuisances are learned through gradient boosting models called XGBoost (Chen and Guestrin,

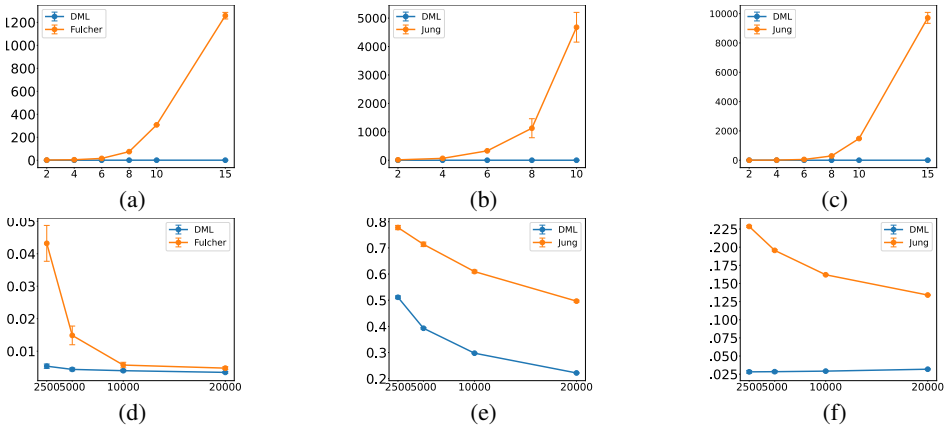


Figure 2: Comparison of DML-UCA ('DML') with existing estimators using **(Top)** running-time-plots (*x-axis*: the dimension of summed variables, *y-axis*: running time); and **(Bottom)** AAE-plots (*x-axis*: the sample size, *y-axis*: errors). DML-UCA is compared with **(a,e)** Fulcher et al. (2019) for FD; **(b.f)** (Jung et al., 2021a) for Verma's equation; and **(c,f)** Jung et al. (2021a) for Jung's equation.

2016). We specify an SCM $\mathcal{M}$ for FD (Fig. 1a), Verma (Fig. 1b), and the example graph in (Jung et al., 2021a) (Fig. 1e), and generate datasets $\mathtt{D}^k \sim \mathtt{P}^k$ from the SCM. For each example, we estimate the target estimand ψ_0 (e.g., $\mathbb{E}[Y \mid \mathrm{do}(x)]$ in FD and Verma). Details of simulations are in Appendix F. Further simulations are provided in Appendix E.

Scalability. To demonstrate scalability of DML-UCA, we compare the running time with existing estimators of (Fulcher et al., 2019) (FD) and (Jung et al., 2021a) (Verma's equation and the causal estimand with Fig. 1e — $\mathbb{E}[Y \mid \mathrm{do}(x_1, x_2)] = \sum_{x_1', r, z} \mathbb{E}_P[Y \mid r, x_1', x_2, z] P(r \mid x_1, z) P(z, x_1')$ — which we call 'Jung's equation'). For each example, we increment the dimension of the summed variables, run 100 simulations, take the average of running times, and compare this average. We label this plot as 'run-time-plot', presented in the top side of Fig. 2. In the comparison with (Fulcher et al., 2019) for FD in Fig. 2a, we fix $|C| = 2$ and increment $|Z| = \{2, 4, 6, 8, 10, 15\}$. When comparing with (Jung et al., 2021a), for Verma's equations in Figs. (2b), we fix $|A| = 2$ and increment $|B| = \{2, 4, 6, 8, 10\}$. For Jung's equation in Fig. 2c, we fix $|Z| = 2$ and $|R| = \{2, 4, 6, 8, 10, 15\}$. For all scenarios, the run-time of existing estimators increases rapidly over dimensions due to the summation operation, while DML-UCA scales well.

Doubly robustness. To demonstrate doubly robustness, we compare the error of DML-UCA with existing estimators for FD of Fulcher et al. (2019) and for Verma's and Jung's equations of Jung et al. (2021a) We use $\hat{\psi}^{\text{est}}$ for est $\in \{\text{DML}, \text{Fulcher}, \text{Jung}\}$ to denote each estimator. We use the average absolute error (AAE), which is an average of the error of the estimated versus true causal effect of $\mathbf{X} = \mathbf{x}$: $\frac{1}{|\text{domain}(\mathbf{X})|} \sum_{\mathbf{x} \in \text{domain}(\mathbf{X})} |\hat{\psi}^{\text{est}}(\mathbf{x}) - \psi_0(\mathbf{x})|$. To witness the fast convergence of DML-UCA, we enforce the convergence rate of nuisance estimates to be no faster than the decaying rate $n^{-1/4}$ by adding the noise term $\epsilon \sim \mathtt{normal}(n^{-1/4}, n^{-1/4})$ to nuisances, inspired by the experimental design in (Kennedy, 2023). We ran 100 simulations for each number of samples $n = \{2500, 5000, 10000, 20000\}$. We label the plot as 'AAE-plot', presented in the bottom side of Fig. 2. For each example, DML-UCA outperforms other estimators, exhibiting fast convergence.

5 Conclusions

We introduce a framework that encompasses a broad class of sum-product causal estimands, called UCA class, for which scalable estimators were previously unavailable. We demonstrate the expressiveness of the UCA class, which includes not only BD/SBD but also broader classes such as Tian's adjustment incorporating FD and Verma, and Ctf-DE, for which the existing SBD-based framework is not applicable. We develop an estimator for UCA called DML-UCA that can estimate the target estimand in polynomial time relative to the number of samples and variables, ensuring scalability. We provide finite-sample guarantees and corresponding asymptotic error analysis for DML-UCA, demonstrating its fast convergence. These scalability and fast convergence properties are empirically verified through simulations. Our results pave the way toward developing an estimation framework maximizing both coverage and scalability in Table 1.

Acknowledgements

This research is supported in part by the NSF, ONR, AFOSR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

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

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. (2020). On pearl's hierarchy and the foundations of causal inference. Technical Report R-60. Causal Artificial Intelligence Laboratory, Columbia University.

Bareinboim, E. and Pearl, J. (2012a). Causal inference by surrogate experiments: z-identifiability. In *In Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence*, pages 113–120. AUAI Press.

Bareinboim, E. and Pearl, J. (2012b). Transportability of causal effects: Completeness results. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, pages 698–704.

Bareinboim, E. and Pearl, J. (2016). Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352.

Bareinboim, E., Tian, J., and Pearl, J. (2014). Recovering from selection bias in causal and statistical inference. In *Proceedings of the 28th AAAI Conference on Artificial Intelligence*, pages 2410–2416.

Berry, A. C. (1941). The accuracy of the gaussian approximation to the sum of independent variates. *Transactions of the american mathematical society*, 49(1):122–136.

Bhattacharya, R., Nabi, R., and Shpitser, I. (2022). Semiparametric inference for causal effects in graphical models with hidden variables. *Journal of Machine Learning Research*, 23:1–76.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A., and Ritov, Y. (1993). *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

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

Chernozhukov, V., Newey, W., Singh, R., and Syrgkanis, V. (2022). Automatic debiased machine learning for dynamic treatment effects and general nested functionals. *arXiv preprint arXiv:2203.13887*.

Chernozhukov, V., Newey, W. K., and Singh, R. (2023). A simple and general debiased machine learning theorem with finite-sample guarantees. *Biometrika*, 110(1):257–264.

Correa, J., Lee, S., and Bareinboim, E. (2021). Nested counterfactual identification from arbitrary surrogate experiments. *Advances in Neural Information Processing Systems*, 34.

Correa, J. D., Tian, J., and Bareinboim, E. (2018). Generalized adjustment under confounding and selection biases. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Díaz, I., Williams, N., Hoffman, K. L., and Schenck, E. J. (2023). Nonparametric causal effects based on longitudinal modified treatment policies. *Journal of the American Statistical Association*, 118(542):846–857.

Esseen, C.-G. (1942). On the liapunov limit error in the theory of probability. *Ark. Mat. Astr. Fys.*, 28:1–19.

Fulcher, I. R., Shpitser, I., Marealle, S., and Tchetgen Tchetgen, E. J. (2019). Robust inference on population indirect causal effects: the generalized front door criterion. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Guo, A., Benkeser, D., and Nabi, R. (2023). Targeted machine learning for average causal effect estimation using the front-door functional. *arXiv preprint arXiv:2312.10234*.

Heckman, J. J. (1991). *Randomization and social policy evaluation*. National Bureau of Economic Research Cambridge, MA.

Heckman, J. J. (1992). Randomization and Social Policy Evaluation. In Manski, C. and Garfinkle, I., editors, *Evaluations: Welfare and Training Programs*, pages 201–230. Harvard University Press, Cambridge, MA.

Huang, Y. and Valtorta, M. (2006). Pearl's calculus of intervention is complete. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 217–224. AUAI Press.

Jung, Y., Díaz, I., Tian, J., and Bareinboim, E. (2023a). Estimating causal effects identifiable from combination of observations and experiments. In *Proceedings of the 37th Neural Information Processing Systems*.

Jung, Y., Tian, J., and Bareinboim, E. (2021a). Estimating identifiable causal effects on markov equivalence class through double machine learning. In *Proceedings of the 38th International Conference on Machine Learning*.

Jung, Y., Tian, J., and Bareinboim, E. (2021b). Estimating identifiable causal effects through double machine learning. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*.

Jung, Y., Tian, J., and Bareinboim, E. (2023b). Estimating joint treatment effects by combining multiple experiments. In *Proceedings of the 40th International Conference on Machine Learning*.

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

Kennedy, E. H., Balakrishnan, S., G'Sell, M., et al. (2020). Sharp instruments for classifying compliers and generalizing causal effects. *Annals of Statistics*, 48(4):2008–2030.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, S., Correa, J., and Bareinboim, E. (2020). General transportability–synthesizing observations and experiments from heterogeneous domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10210–10217.

Lee, S., Correa, J. D., and Bareinboim, E. (2019). General identifiability with arbitrary surrogate experiments. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

Li, S. and Luedtke, A. (2023). Efficient estimation under data fusion. *Biometrika*, 110(4):1041–1054.

Luedtke, A. R., Sofrygin, O., van der Laan, M. J., and Carone, M. (2017). Sequential double robustness in right-censored longitudinal models. *arXiv preprint arXiv:1705.02459*.

Mises, R. v. (1947). On the asymptotic distribution of differentiable statistical functions. *The annals of mathematical statistics*, 18(3):309–348.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–710.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

Pearl, J. and Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.

Pearl, J. and Robins, J. (1995a). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, pages 444–453. Morgan Kaufmann Publishers Inc.

Pearl, J. and Robins, J. (1995b). Probabilistic evaluation of sequential plans from causal models with hidden variables. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 444–453.

Pires, A. M. and Branco, J. A. (2002). Partial influence functions. *Journal of Multivariate Analysis*, 83(2):451–468.

Plečko, D., Bareinboim, E., et al. (2024). Causal fairness analysis: A causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, 17(3):304–589.

Precup, D. (2000). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80.

Quintas-Martinez, V., Bahadori, M. T., Santiago, E., Mu, J., Janzing, D., and Heckerman, D. (2024). Multiply-robust causal change attribution. *arXiv preprint arXiv:2404.08839*.

Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512.

Robins, J., Li, L., Tchetgen, E., and van der Vaart, A. W. (2009). Quadratic semiparametric von mises calculus. *Metrika*, 69:227–247.

Rotnitzky, A., Robins, J., and Babino, L. (2017). On the multiply robust estimation of the mean of the g-functional. *arXiv preprint arXiv:1705.08582*.

Shevtsova, I. (2014). On the absolute constants in the berry-esseen-type inequalities. In *Doklady Mathematics*, volume 89, pages 378–381. Springer.

Shpitser, I. and Pearl, J. (2006). Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Tian, J. and Pearl, J. (2002a). A general identification condition for causal effects. In *Proceedings of the 18th National Conference on Artificial Intelligence*, pages 567–573.

Tian, J. and Pearl, J. (2002b). On the testable implications of causal models with hidden variables. In *Proceedings of the 18th conference on Uncertainty in artificial intelligence*, pages 519–527. Morgan Kaufmann Publishers Inc.

Tian, J. and Pearl, J. (2003). On the identification of causal effects. Technical Report R-290-L.

van der Laan, M. J. and Gruber, S. (2012). Targeted minimum loss based estimation of causal effects of multiple time point interventions. *The international journal of biostatistics*, 8(1).

Verma, T. and Pearl, J. (1990). Causal networks: Semantics and expressiveness. In *Machine intelligence and pattern recognition*, volume 9, pages 69–76. Elsevier.

Xia, K., Lee, K.-Z., Bengio, Y., and Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34.

Xia, K. M., Pan, Y., and Bareinboim, E. (2022). Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations*.

Xu, L. and Gretton, A. (2022). A neural mean embedding approach for back-door and front-door adjustment. In *The Eleventh International Conference on Learning Representations*.

Supplement to "Unified Covariate Adjustment for Causal Inference"

Contents

A Nuisance Specification

A.1 Front-door adjustment in Example 1

First, $\mathbf{C}_1 = \{X, C\}$, $\mathbf{C}_2 = \{Z\}$, $\mathbf{S}_1 = \{C\}$, $\mathbf{S}_1^b = \{X\}$, $\mathbf{S}_2 = \{Z, X\}$. Also, $\mathbf{B}_2 \coloneqq \mathbf{S}_1^b \cap \mathbf{C}_1 = \{X\}$. $\mathbf{R} = \emptyset$. Then, $\check{\mathbf{S}}_1 = \mathbf{S}_1 = \{C\}$ and $\check{\mathbf{S}}_2 = \{Z, X', C\}$.

The regression nuisances are the followings:

$$\begin{aligned}
&\mu_0^2(\mathbf{S}_2) \coloneqq \mu_0^2(Z, X, C) \coloneqq \mathbb{E}_P[Y \mid Z, X, C] \\
&\check{\mu}_0^2(\check{\mathbf{S}}_2) \coloneqq \mu_0^2(Z, X', C) \\
&\mu_0^1(\mathbf{S}_1) \coloneqq \mu_0^1(C) \coloneqq \mathbb{E}_P[\mu_0^2(Z, X', C) \mid x, C] \\
&\check{\mu}_0^1(\mathbf{S}_1) = \mu_0^1(C).
\end{aligned}$$

The ratio nuisances are the followings:

$$\pi_0^2(Z, X, C) = \frac{P(Z \mid x, C)}{P(Z \mid X, C)}, \tag{A.1}$$

$$\pi_0^1(C) = \frac{P(x)}{P(x \mid C)}. \tag{A.2}$$

The representation for DML-UCA is

$$\begin{aligned}
&\mathbb{E}_P[\pi_0^2(Z, X, C)\{Y - \mu_0^2(Z, X, C)\}] \\
&+ \mathbb{E}_P[\pi_0^1(C)\{\mu_0^2(Z, X', C) - \mu_0^1(C)\} \mid x] \\
&+ \mathbb{E}_P[\mu_0^1(C)].
\end{aligned}$$

A.2 Verma's equation in Example 2

From the fact that Verma's equation in Eq. (3) is represented as the expectation of Y over $P(Y \mid B, A, x)P(B \mid A, X)P(A \mid x)P(X)$. Set

- $\mathbf{R} = \emptyset$.
- $\mathbf{C}_1 = \{X\}$, $\mathbf{C}_2 = \{A\}$, $\mathbf{C}_3 = \{B\}$.
- $\mathbf{S}_1 = \emptyset$, $\mathbf{S}_2 = \{A, X\}$, $\mathbf{S}_3 = \{B, A\}$.
- $\mathbf{S}_1^b = \{X\}$, $\mathbf{S}_3^b = \{X\}$.
- $\mathbf{B}_2 = \{X\}$, $\mathbf{B}_4 = \{X\}$.
- $\check{\mathbf{S}}_2 = \{A, X'\}$. For $i \neq 2$, $\mathbf{S}_i = \check{\mathbf{S}}_i$.

The regression nuisances are the followings:

$$\begin{aligned}
&\mu_0^3(\mathbf{S}_3) \coloneqq \mu_0^3(B, A, x) \coloneqq \mathbb{E}_P[Y \mid B, A, x] \\
&\check{\mu}_0^3(\check{\mathbf{S}}_3) \coloneqq \mu_0^3(B, A, x) = \mathbb{E}_P[Y \mid B, A, x] \\
&\mu_0^2(\mathbf{S}_2) \coloneqq \mu_0^2(A, X) \coloneqq \mathbb{E}_P[\mu_0^3(B, A, x) \mid A, X] \\
&\check{\mu}_0^2(\check{\mathbf{S}}_2) = \mu_0^2(A, X') \\
&\mu_0^1(\mathbf{S}_1) \coloneqq \mathbb{E}_P[\mu_0^2(A, X') \mid x] \\
&\quad \check{\mu}_0^1 \coloneqq \mu_0^1 \coloneqq \mathbb{E}_P[\mu_0^2(A, X') \mid x].
\end{aligned}$$

The ratio nuisances are the followings:

$$\pi_0^3(B, A, X') = \frac{P(B \mid A, X')P(A \mid x)}{P(B, A \mid x)} = \frac{P(B \mid A, X')}{P(B \mid A, x)} \frac{P(A \mid X')}{P(A \mid x)}, \tag{A.3}$$

$$\pi_0^2(A, X) = \frac{P(A \mid x)P(X)}{P(A, X)} = \frac{P(A \mid x)}{P(A \mid X)}. \tag{A.4}$$

The representation for DML-UCA is

$$\begin{aligned}
&\mathbb{E}_P[\pi_0^3(B, A)\{Y - \mu_0^3(B, A, x)\} \mid x] \\
&+ \mathbb{E}_P[\pi_0^2(A, X)\{\mu_0^3(B, A, x) - \mu_0^2(A, X)\}] \\
&+ \mathbb{E}_P[\mu_0^1(x)].
\end{aligned}$$

A.3 Counterfactual directed effect in Example 3

From the fact that Ctf-DE in Eq. (5) is represented as the expectation of Y over $P(Y \mid X = x_0, W, Z)P(W \mid X, Z)P(Z \mid X = x_2)\mathbb{1}_{x_1}(X)$. Set

- $\mathbf{R}_1 = \{X\}$.
- $\mathbf{C}_1 = \{Z\}$, $\mathbf{C}_2 = \{W\}$
- $\mathbf{S}_1 = \{X, Z\}$, $\mathbf{S}_2 = \{W, Z\}$
- $\mathbf{S}_1^b = \emptyset$, $\mathbf{S}_2^b = \{X\}$.
- $\check{\mathbf{S}}_i = \mathbf{S}_i \setminus \mathbf{R}_i$ for all i

The regression nuisances are the followings:

$$\begin{aligned}
&\mu_0^2(\mathbf{S}_2) \coloneqq \mu_0^2(W, Z) \coloneqq \mathbb{E}_P[Y \mid W, x_0, Z] \\
&\check{\mu}_0^2(\check{\mathbf{S}}_2) \coloneqq \mu_0^2(W, Z) = \mathbb{E}_P[Y \mid W, x_0, Z] \\
&\mu_0^1(\mathbf{S}_1) \coloneqq \mu_0^1(X, Z) \coloneqq \mathbb{E}_P[\mu_0^2(W, Z) \mid X, Z] \\
&\check{\mu}_0^1(\check{\mathbf{S}}_1) = \mu_0^1(x_1, Z).
\end{aligned}$$

The ratio nuisances are the followings:

$$\pi_0^2(W, Z) = \frac{P(W \mid x_1, Z)P(Z \mid x_2)}{P(W, Z \mid x_0)}, \tag{A.5}$$

$$\pi_0^1(X, Z) = \frac{\mathbb{1}_{x_1}(X)P(Z \mid x_2)}{P(X \mid Z)P(Z)}. \tag{A.6}$$

The representation for DML-UCA is

$$\begin{aligned}
&\mathbb{E}_P[\pi_0^2(W, Z)\{Y - \mu_0^2(W, X, Z)\} \mid X = x_0] \\
&+ \mathbb{E}_P[\pi_0^1(X, Z)\{\mu_0^2(W, x_0, Z) - \mu_0^1(X, Z)\}] \\
&+ \mathbb{E}_P[\mu_0^1(x_1, Z) \mid x_2].
\end{aligned}$$

A.4 Example Estimand for Fig. 1e

Given Fig. 1e, the causal effect is given as

$$\mathbb{E}[Y \mid \mathrm{do}(x_1, x_2)] = \sum_{r, z, x_1'} \mathbb{E}_P[Y \mid r, x_2, z, x_1']P(r \mid x_1, z)P(x_1', z),$$

which is the expectation of Y over the probability measure

$$P(Y \mid R, X_2, Z, X_1)P(R \mid x_1, Z)P(X_1, Z)\mathbb{1}_{x_2}(X_2).$$

Set

- $\mathbf{C}_1 = \{X_1, Z\}$ and $\mathbf{C}_2 = \{R\}$.
- $\mathbf{R}_1 = \emptyset$ and $\mathbf{R}_2 = \{X_2\}$ with $\sigma^2_{\mathbf{R}_2} = \mathbb{1}_{x_2}(X_2)$.
- $\mathbf{S}_1 = \{Z\}$, $\mathbf{S}_2 = \{R, X_2, Z, X_1\}$.
- $\mathbf{S}_1^b = \{X_1\}$.
- $\mathbf{B}_2 = \{X_1\}$.
- $\check{\mathbf{S}}_2 = \{R, X_1', Z\}$. $\check{\mathbf{S}}_1 = \mathbf{S}_1$.

The regression nuisances are the followings:

$$\begin{aligned}
&\mu_0^2(\mathbf{S}_2) \coloneqq \mu_0^2(R, X_2, Z, X_1) \coloneqq \mathbb{E}_P[Y \mid R, X_2, Z, X_1] \\
&\check{\mu}_0^2(\check{\mathbf{S}}_2) \coloneqq \check{\mu}_0^2(R, Z, X_1') = \mathbb{E}_P[Y \mid R, x_2, Z, X_1'] \\
&\mu_0^1(\mathbf{S}_1) \coloneqq \mu_0^1(Z) \coloneqq \mathbb{E}_P[\check{\mu}_0^2(R, Z, X_1') \mid Z] \\
&\check{\mu}_0^1(\check{\mathbf{S}}_1) = \mu_0^1(Z).
\end{aligned}$$

The ratio nuisances are the followings:

$$\pi_0^2(X_2, R, X_1, Z) = \frac{\mathbb{1}_{x_2}(X_2)}{P(X_2 \mid R, X_1, Z)} \frac{P(R \mid x_1, Z)}{P(R \mid X_1, Z)}, \tag{A.7}$$

$$\pi_0^1(Z) = \frac{P(Z)}{P(Z \mid x)} = \frac{P(x)P(Z)}{P(x \mid Z)P(Z)} = \frac{1}{P(x \mid Z)}. \tag{A.8}$$

The representation for DML-UCA is

$$\begin{aligned}
&\mathbb{E}_P[\pi_0^2(X_2, R, X_1, Z)\{Y - \mu_0^2(R, X_2, Z, X_1)\}] \\
&+ \mathbb{E}_P[\pi_0^1(Z)\{\check{\mu}_0^2(R, Z, X_1) - \mu_0^1(Z)\} \mid X_1 = x_1] \\
&+ \mathbb{E}_P[\mu_0^1(Z)].
\end{aligned}$$

B More UCA Examples

B.1 Effect of the treatment on the treated (ETT)

Let $\mathbf{V} = \{\mathbf{Z}, X, Y\}$ be a set of variables where $\mathbf{Z}$ is a covariate, X is a treatment and Y is an outcome. The target estimand is

$$\mathbb{E}[Y(x) \mid x'] = \sum_{\mathbf{z}} \mathbb{E}_P[Y \mid x, \mathbf{z}] P(\mathbf{z} \mid x'). \tag{B.1}$$

The ETT estimand can be written as an expectation of Y over the probability measure

$$\Psi = P(Y \mid X, \mathbf{Z}) P(\mathbf{Z} \mid x') \mathbb{1}_x(X).$$

This factorization implies that $\mathbf{C}_1 \coloneqq \{\mathbf{Z}\}$, $\mathbf{R} \coloneqq \mathbf{V} \setminus \mathbf{C}_1 \cup \{Y\} = \{X\}$, where $\mathbf{R}_1 = \{X\}$, and $\sigma^1_{\mathbf{R}_1} \coloneqq \mathbb{1}_x(X)$. Also, $\mathbf{S}_1 = \{X\} \cup \mathbf{Z}$. Finally,

$$\begin{aligned}
P^1(\mathbf{C}_1) &= P(\mathbf{Z} \mid x') \\
P^2(Y \mid \mathbf{S}_1) &= P(Y \mid X, \mathbf{Z}).
\end{aligned}$$

The regression nuisances are the followings:

$$\begin{aligned}
\mu_0^1(\mathbf{S}_1) &\coloneqq \mu_0^1(X, \mathbf{Z}) \coloneqq \mathbb{E}_P[Y \mid X, \mathbf{Z}] \\
\check{\mu}_0^1(\mathbf{S}_1 \setminus \mathbf{R}_1) &\coloneqq \check{\mu}_0^1(\mathbf{Z}) \mu_0^1(x, \mathbf{Z}).
\end{aligned}$$

The ratio nuisances are the followings:

$$\pi_0^1(X, \mathbf{Z}) = \frac{P(Z \mid x')\mathbb{1}_x(X)}{P(X, \mathbf{Z})} = \frac{P(x' \mid \mathbf{Z})P(\mathbf{Z})}{P(x')} \frac{\mathbb{1}_x(X)}{P(X \mid \mathbf{Z})P(\mathbf{Z})} = \frac{P(x' \mid \mathbf{Z})}{P(X \mid \mathbf{Z})} \frac{\mathbb{1}_x(X)}{P(x')}.$$

The representation for DML-UCA is

$$\mathbb{E}_P[\pi_0^1(X, \mathbf{Z})\{Y - \mu_0^1(X, \mathbf{Z})\}] + \mathbb{E}_P[\breve{\mu}_0^1(\mathbf{Z}) \mid x'].$$

B.2 Transportability (S-admissibility)

Let $\mathbf{V} = \{\mathbf{Z}, X, Y\}$ be a set of variables where $\mathbf{Z}$ is a covariate, X is a treatment and Y is an outcome. Let S denote the domain indicator such that $S = 0$ means the target domain, and $S = 1$ denotes the source. The S-admissibility estimand appeared in transportability scenario is

$$\mathbb{E}[Y \mid \mathrm{do}(x)] = \sum_{\mathbf{z}} \mathbb{E}_P[Y \mid x, \mathbf{z}, S = 1]P(\mathbf{z} \mid S = 0). \tag{B.2}$$

The estimand can be written as an expectation of Y over the probability measure

$$\Psi = P(Y \mid X, \mathbf{Z}, S = 1)P(\mathbf{Z} \mid S = 0)\mathbb{1}_x(X).$$

From this factorization, we have $\mathbf{C}_1 \coloneqq \mathbf{Z}$ and $\mathbf{R}_1 \coloneqq X$. Also, set $P^1(\mathbf{C}_1) \coloneqq P(\mathbf{Z} \mid S = 0)$ with $\mathbf{S}_0^b = S$. Set $P^2(Y \mid \mathbf{S}_1) \coloneqq P(Y \mid X, \mathbf{Z} \mid S = 1)$ with $\mathbf{S}_1^b = S$ and $\mathbf{S}_1 \coloneqq \{X\} \cup \mathbf{Z}$.

The regression nuisances are the followings:

$$\mu_0^1(\mathbf{S}_1) \coloneqq \mu_0^1(X, \mathbf{Z}) \coloneqq \mathbb{E}_P[Y \mid X, \mathbf{Z}, S = 1]$$
$$\breve{\mu}_0^1(\breve{\mathbf{S}}_1) \coloneqq \breve{\mu}_0^1(\mathbf{Z}) = \mu_0^1(x, \mathbf{Z}).$$

The ratio nuisances are the followings:

$$\pi_0^1(X, \mathbf{Z}) = \frac{\mathbb{1}_x(X)}{P(X \mid \mathbf{Z}, S = 1)} \frac{P(\mathbf{Z} \mid S = 0)}{P(\mathbf{Z} \mid S = 1)}.$$

The representation for DML-UCA is

$$\mathbb{E}_P[\pi_0^1(X, \mathbf{Z})\{Y - \mu_0^1(X, \mathbf{Z})\} \mid S = 1] + \mathbb{E}_P[\breve{\mu}_0^1(\mathbf{Z}) \mid S = 0].$$

B.3 Off-policy evaluation

Let $\mathbf{V} = \{\mathbf{Z}, X, Y\}$ be a set of variables where $\mathbf{Z}$ is a covariate, X is a treatment and Y is an outcome. Let $\sigma^*(X \mid Z)$ denote the behavioral policy that an agent observed; i.e.,

$$(\mathbf{Z}, X, Y) \sim P(Y \mid X, \mathbf{Z})\sigma^*(X \mid \mathbf{Z})P(\mathbf{Z}). \tag{B.3}$$

Let $\sigma(X \mid \mathbf{Z})$ denote a policy to be evaluated. Then, the effect of the policy σ^* is given as

$$\mathbb{E}[Y \mid \sigma] \coloneqq \sum_{x, \mathbf{z}} \mathbb{E}_P[Y \mid x, \mathbf{z}]\sigma^*(x \mid \mathbf{z})P(\mathbf{z}). \tag{B.4}$$

The policy treatment effect in Eq. (B.4) can be represented as UCA as follow.

$$\mathbf{C}_1 \coloneqq \mathbf{Z}$$
$$\mathbf{R}_1 \coloneqq \{X\}$$
$$\sigma_{\mathbf{R}_1}^1 \coloneqq \sigma^*(X \mid Z)$$
$$\mathbf{S}_1 \coloneqq \{X\} \cup \mathbf{Z}.$$

Set $P^1(\mathbf{C}_1) \leftarrow P(\mathbf{Z})$, $\sigma^1_{\mathbf{R}_1}(\mathbf{R}_1 \mid \mathbf{Z}_1) \leftarrow \sigma(X \mid \mathbf{Z})$, and $P^2(Y \mid \mathbf{C}_1, \mathbf{R}_1) \leftarrow P(Y \mid X, \mathbf{Z})$. Then,

$$\begin{aligned}\Psi(\mathbf{P}; \boldsymbol{\sigma}) &\coloneqq \sum_{\mathbf{c},\mathbf{R}} \mathbb{E}_{P^2}[Y \mid \mathbf{c}_1, \mathbf{R}_1]\sigma^1_{\mathbf{R}_1} P^1(\mathbf{c}_1) \\ &= \sum_{x,\mathbf{z}} \mathbb{E}_P[Y \mid x, \mathbf{z}]\sigma^*(x \mid \mathbf{z})P(\mathbf{z}) \\ &= \mathbb{E}[Y \mid \sigma] \qquad \text{(Eq. (B.4))}.\end{aligned}$$

The regression nuisances are the followings:

$$\begin{aligned}\mu^1_0(\mathbf{C}^{(1)} \cup \mathbf{R}^{(1)}) &\coloneqq \mu^1_0(X, \mathbf{Z}) \coloneqq \mathbb{E}_P[Y \mid X, \mathbf{Z}] \\ \breve{\mu}^1_0(\mathbf{C}^{(1)}) &\coloneqq \breve{\mu}^1_0(\mathbf{Z}) \coloneqq \sum_x \mu^1_0(x, \mathbf{Z})\sigma^*(x \mid \mathbf{Z}).\end{aligned}$$

The ratio nuisances are the followings:

$$\pi^1_0(X, \mathbf{Z}) = \frac{\sigma^*(X \mid \mathbf{Z})}{P(X \mid \mathbf{Z})}.$$

The representation for DML-UCA is

$$\mathbb{E}_P[\pi^1_0(X, \mathbf{Z})\{Y - \mu^1_0(X, \mathbf{Z})\}] + \mathbb{E}_P[\breve{\mu}^1_0(\mathbf{Z})].$$

B.4 Treatment-treatment interactions

Let $\mathbf{V} = \{\mathbf{Z}, X, Y\}$ be a set of variables where $\mathbf{Z}$ is a covariate, X is a treatment and Y is an outcome. The estimand for treatment-treatment interaction discussed in Jung et al. (2023b) is

$$\mathbb{E}[Y \mid \mathrm{do}(x_1, x_2)] = \sum_{\mathbf{z}} \mathbb{E}[Y \mid \mathrm{do}(x_2), \mathbf{z}, x_1]P(\mathbf{z} \mid \mathrm{do}(x_1)), \tag{B.5}$$

which is an expectation of Y over a product of probability measure

$$P(Y \mid \mathbf{Z}, \mathrm{do}(x_2), X_1)P(\mathbf{Z} \mid \mathrm{do}(x_1))\mathbb{1}_{x_1}(X_1),$$

which satisfies an additivity. Therefore, $\mathbb{E}[Y \mid \mathrm{do}(x_1, x_2)]$ is UCA-expressible. Such reduction can be done since the probability measure satisfies additivity w.r.t. all conditional distributions and the policy $\mathbb{1}_{x_1}(X_1)$. Specifically, set

$$\begin{aligned}\mathbf{C}_1 &\coloneqq \mathbf{Z} \\ \mathbf{R}_1 &\coloneqq \{X_1\} \\ \mathbf{S}_1 &\coloneqq \{X_1\} \cup \mathbf{Z}.\end{aligned}$$

Also, set

$$\begin{aligned}P^1(\mathbf{C}_1) &\coloneqq P(\mathbf{Z} \mid \mathrm{do}(x_1)) \\ P^2(Y \mid \mathbf{C}_1 \cup \mathbf{R}_1) &\coloneqq P(Y \mid X_1, \mathbf{Z}, \mathrm{do}(x_2)) \\ \sigma^1_{\mathbf{R}_1} &\coloneqq \mathbb{1}_{x_1}(X_1).\end{aligned}$$

The regression nuisances are the followings:

$$\begin{aligned}\mu^1_0(\mathbf{C}^{(1)} \cup \mathbf{R}^{(1)}) &\coloneqq \mu^1_0(X_1, \mathbf{Z}) \coloneqq \mathbb{E}_P[Y \mid X_1, \mathbf{Z}, \mathrm{do}(x_2)] \\ \breve{\mu}^1_0(\mathbf{C}^{(1)}) &\coloneqq \mathbb{E}_P[Y \mid x_1, \mathbf{Z}, \mathrm{do}(x_2)].\end{aligned}$$

The ratio nuisances are the followings:

$$\pi_0^1(X, \mathbf{Z}) = \frac{\mathbb{1}_{x_1}(X_1)P(\mathbf{Z} \mid \mathrm{do}(x_1))}{P(X_1 \mid \mathbf{Z}, \mathrm{do}(x_2))P(\mathbf{Z} \mid \mathrm{do}(x_2))},$$

which can be estimated through the density estimation approach using the probabilistic classification method described in (Díaz et al., 2023, Sec. 5.4).

The representation for DML-UCA is

$$\mathbb{E}_P[\pi_0^1(X, \mathbf{Z})\{Y - \mu_0^1(X, \mathbf{Z})\} \mid \mathrm{do}(x_2)] + \mathbb{E}_P[\check{\mu}_0^1(\mathbf{Z}) \mid \mathrm{do}(x_1)].$$

C More Results

C.1 Formal definition of Partial influence function (PIF)

Definition C.1 (Partial influence function (PIF) (Pires and Branco, 2002)**).** *Let $g(\mathtt{P}^1, \cdots, \mathtt{P}^K)$ denote a K-multi-distribution functional. For the k-th component, let $\mathtt{P}_t^k := \mathtt{P}^k + t(\mathtt{Q}^k - \mathtt{P}^k)$ for $t \in [0, 1]$, where $\mathtt{Q}^k$ is an arbitrary distribution absolutely continuous w.r.t. $\mathtt{P}^k$. The k-th **partial influence function** is a function $\phi^k(\mathbf{V}; \boldsymbol{\eta}^i(\mathtt{P}^k), g_0)$ such that $\mathbb{E}_{\mathtt{P}^k}[\phi^k(\mathbf{V}; \eta^k(\mathtt{P}^k), g_0)] = 0$, $\mathbb{V}_{\mathtt{P}^k}[\phi^k(\mathbf{V}; \eta^k(\mathtt{P}^k), g_0)] < \infty$, and $\frac{\partial}{\partial t} g(\mathtt{P}^1, \cdots, \mathtt{P}_t^k, \cdots, \mathtt{P}^K)\big|_{t=0} = \mathbb{E}_{\mathtt{Q}^k}[\phi^k(\mathbf{V}; \boldsymbol{\eta}^k(\mathtt{P}^k), g_0)]$.*

C.2 Density Ratio Estimation

Two available approaches for estimating the density ratio are the followings. The first approach is to apply the Bayes rule for rewriting the density ratio into more tractable form. For example, consider the problem of estimating π_0^2 for FD, which is given as

$$\pi_0^2 \coloneqq \frac{P(Z \mid x, C)}{P(Z \mid X, C)}.$$

Suppose Z, C are high-dimensional random vectors, and X is a binary singleton variable. Then, $P(X \mid C)$ or $P(X \mid Z, C)$ are tractable to estimate compared to $P(Z \mid X, C)$, since estimating $P(X \mid \cdot)$ can be done using off-the-shelf probabilistic classification method. Here, π_0^2 can be written as a tractable form as follows:

$$\begin{aligned}
\pi_0^2 &\coloneqq \frac{P(Z \mid x, C)}{P(Z \mid X, C)} \\
&= \frac{P(Z, X, C)}{P(X \mid C)P(C)} \frac{P(x \mid C)P(C)}{P(Z, x, C)} \\
&= \frac{P(C)}{P(C)} \frac{P(Z, C)}{P(Z, C)} \frac{P(x \mid C)}{P(X \mid C)} \frac{P(X \mid Z, C)}{P(x \mid Z, C)} \\
&= \frac{P(x \mid C)}{P(X \mid C)} \frac{P(X \mid Z, C)}{P(x \mid Z, C)}.
\end{aligned}$$

The second approach is to recast the density ratio into the classification problem (Díaz et al., 2023, Sec. 5.4). For example, consider the ratio nuisance appeared in Treatment-treatment interactions:

$$\pi_0^1(X, \mathbf{Z}) = \frac{\mathbb{1}_{x_1}(X_1)P(\mathbf{Z} \mid \mathrm{do}(x_1))}{P(X_1 \mid \mathbf{Z}, \mathrm{do}(x_2))P(\mathbf{Z} \mid \mathrm{do}(x_2))}.$$

Here, $\frac{P(\mathbf{Z}|\mathrm{do}(x_1))}{P(\mathbf{Z}|\mathrm{do}(x_2))}$ can be estimated as a following procedure. Let $\mathcal{D}_1 \sim P(\mathbf{Z} \mid \mathrm{do}(x_1)$ and $\mathcal{D}_2 \sim P(\mathbf{Z} \mid \mathrm{do}(x_2)$ denote samples. Let $\mathcal{D}_0 \coloneqq \mathcal{D}_1 \cup \mathcal{D}_2$. Let λ denote an indicator such that $\lambda = 0$ means samples are from $\mathcal{D}_1$ and $\lambda = 1$ means they are from $\mathcal{D}_2$. Without loss of generality, $|\mathcal{D}_1| = |\mathcal{D}_2|$. Then,

$$\frac{P(\mathbf{Z} \mid \mathrm{do}(x_1))}{P(\mathbf{Z} \mid \mathrm{do}(x_2))} = \frac{P(\mathbf{Z} \mid \lambda = 0)}{P(\mathbf{Z} \mid \lambda = 1)} = \frac{P(\lambda = 1)}{P(\lambda = 0)} \frac{P(\lambda = 0 \mid \mathbf{Z})P(\mathbf{Z})}{P(\lambda = 1 \mid \mathbf{Z})P(\mathbf{Z})} = \frac{P(\lambda = 0 \mid \mathbf{Z})}{P(\lambda = 1 \mid \mathbf{Z})}.$$

Then, instead of estimating the density ratio explicitly as $\frac{P(\mathbf{Z}|\mathrm{do}(x_1))}{P(\mathbf{Z}|\mathrm{do}(x_2))}$, we can estimate the equivalent estimand $\frac{P(\lambda=0|\mathbf{Z})}{P(\lambda=1|\mathbf{Z})}$ using any off-the-shelf probabilistic classification method.

C.3 Analysis of non-UCA functionals

We consider three cases where a target estimand cannot be expressed through UCA:

1. **Case 1.** The target estimand is not in a form of the product (e.g., the target estimand is the quotient of sum-products of two conditional distributions).
2. **Case 2.** For a target estimand that is represented as the expectation of Y over the measure

$$\Psi'[\mathbf{P};\boldsymbol{\sigma}] \coloneqq P^{m+1}(Y \mid \mathbf{S}'_m)\prod_{i=1}^{m} P^i(\mathbf{C}_i \mid \mathbf{S}'_{m-1})\sigma^i_{\mathbf{R}_i}(\mathbf{R}_i \mid \mathbf{S}'_i \setminus \mathbf{R}_i),$$

   where $P^i(\mathbf{V}) = Q^i(\mathbf{V} \mid \mathbf{S}^b_{i-1} = \mathbf{s})$ for some distribution Q^i, suppose there exists $\mathbf{S}'_{i-1}$ such that $\mathbf{S}'_{i-1} \neq (\mathbf{C}^{(i-1)} \cup \mathbf{R}^{(i-1)}) \setminus \mathbf{S}^b_{i-1}$.
3. **Case 3.** $\mathbf{S}^b_i \cap \mathbf{C}^{\geq 2} \neq \emptyset$.

In this section, we will provide example functionals that cannot be expressed through UCA. Since an example for the first case is described (the napkin estimand where $P(y \mid \mathrm{do}(x)) = \frac{\sum_w P(y,x|r,w)P(w)}{\sum_w P(x|r,w)P(w)}$), we will only consider Case 2 and Case 3 in this section.

C.3.1 On Case 2

Here, we provide an example that the target estimand cannot be expressed through nested regression and empirical bifurcation when the estimand is within Case 2. Consider a following functional:

$$\sum_{a,x'_1,b} \mathbb{E}_P[Y \mid b, a, x'_1, x_2]P(b \mid x_1)P(a, x'_1). \tag{C.1}$$

This functional is an expectation of the probability measure Y over $P(Y \mid B, A, X_1, X_2)P(B \mid x_1)P(A, X_1)\mathbb{1}_{x_2}(X_2)$. Based on this probability measure, apply the following setting:

- $\mathbf{C}_1 = \{A, X_1\}$ and $\mathbf{C}_2 \coloneqq \{B\}$.
- $\mathbf{R}_1 = \emptyset$ and $\mathbf{R}_2 = \{X_2\}$ with $\sigma^2_{\mathbf{R}_2} = \mathbb{1}_{x_2}(X_2)$.
- $P^1(\mathbf{C}_1) \coloneqq P(A, X_1)$ with $\mathbf{S}^b_0 = \emptyset$.
- $P^2(\mathbf{C}_2 \mid \mathbf{S}'_1) = P(B \mid x_1)$ with $\mathbf{S}^b_1 = \{X_1\}$ and $\mathbf{S}'_1 \coloneqq \emptyset$.
- $P^3(Y \mid \mathbf{S}'_2) = P(Y \mid B, A, X_1, X_2)$ with $\mathbf{S}^b_2 = \emptyset$ and $\mathbf{S}'_2 \coloneqq \{B, A, X_1, X_2\}$.

Here, $\mathbf{S}'_1 = \emptyset \neq \mathbf{S}_1 \coloneqq (\mathbf{C}^{(1)} \cup \mathbf{R}^{(1)}) \setminus \mathbf{S}^b_1 = \{A\}$, and therefore, Eq. (C.1) is not within UCA-class.

Now, we will witness that the target estimand cannot be correctly represented through the nested regression and empirical bifurcation. Applying the nested regression, we have

$$\mu^2_0(\mathbf{S}'_2) = \mu^2_0(B, A, X_1, X_2) \coloneqq \mathbb{E}_P[Y \mid B, A, X_2, X_1]$$
$$\check{\mu}^2_0(\mathbf{S}'_2 \setminus \mathbf{R}_2) = \check{\mu}^2_0(B, A, X_1) \coloneqq \mathbb{E}_P[Y \mid B, A, x_2, X_1].$$

Then,

$$\mu^1_0(\mathbf{S}'_1) = \mu^1_0 = \mathbb{E}_P[\check{\mu}^2_0(B, A, X_1) \mid x_1] = \sum_{b,a} \mathbb{E}_P[Y \mid b, a, x_1, x_2]P(b \mid a, x_1)P(a \mid x_1).$$

This representation doesn't correctly represent the target estimand in Eq. (C.1) because of $P(b \mid a, x_1) \neq P(b \mid x_1)$ in general.

Even if the empirical bifurcation has been used to X_1 (i.e., the independent copy X_1' is used)

$$\mu_0^1(\mathbf{S}_1') = \mu_0^1 = \mathbb{E}_P[\check{\mu}_0^2(B, A, X_1') \mid x_1] = \sum_{b,a,x'} \mathbb{E}_P[Y \mid b, a, x_1', x_2] P(b \mid a, x_1) P(a \mid x_1) P(x').$$

Again, this representation doesn't correctly represent the target estimand because of $P(b \mid a, x_1) \neq P(b \mid x_1)$ in general.

C.3.2 On Case 3

Here, we provide an example that the target estimand cannot be expressed through nested regression and empirical bifurcation when the estimand is within Case 3. Consider the following functional:

$$\sum_{b,x',a} \mathbb{E}_P[Y \mid b, x', a] P(b \mid x, a) P(x' \mid a) P(a \mid x). \tag{C.2}$$

This functional is an expectation of Y over the probability measure $P(Y \mid B, X, A)P(B \mid x, A)P(X \mid A)P(A \mid x)$. Based on this probability measure, apply the following setting:

- $\mathbf{C}_1 = \{A\}$, $\mathbf{C}_2 \coloneqq \{X\}$ and $\mathbf{C}_3 \coloneqq \{B\}$.
- $\mathbf{R} = \mathbf{V} \setminus (\{Y\} \cup \mathbf{C}) = \emptyset$.
- $P^1(\mathbf{C}_1) \coloneqq P(A \mid x)$ with $\mathbf{S}_0^b = \{X\}$.
- $P^2(\mathbf{C}_2 \mid \mathbf{S}_1) = P(X \mid A)$ with $\mathbf{S}_1^b = \emptyset$ and $\mathbf{S}_1 \coloneqq \mathbf{C}^{(1)} \setminus \mathbf{S}_1^b = \{A\}$.
- $P^3(\mathbf{C}_3 \mid \mathbf{S}_2) = P(B \mid A, x)$ with $\mathbf{S}_2^b = \{X\}$ and $\mathbf{S}_2 \coloneqq \mathbf{C}^{(2)} \setminus \mathbf{S}_2^b = \{A\}$.
- $P^4(Y \mid \mathbf{S}_3) = P(Y \mid B, A, X)$ with $\mathbf{S}_3^b = \emptyset$ and $\mathbf{S}_3 \coloneqq \mathbf{C}^{(3)} \setminus \mathbf{S}_3^b = \{B, X, A\}$.

With such mapping, Eq. (C.2) can be represented as the expectation of Y over $\Psi[\mathbf{P}] \coloneqq P^4(Y \mid \mathbf{S}_3)P^3(\mathbf{C}_3 \mid \mathbf{S}_2)P^2(\mathbf{C}_2 \mid \mathbf{S}_1)P^1(\mathbf{C}_1)$.

Here, for $\{X\} = \mathbf{S}_0^b = \mathbf{S}_2^b$, and $\{X\} \cap \mathbf{C}^{\geq 2} \neq \emptyset$. That is, a variable X is summed by $\sum_{x'}$ with $P(x' \mid a)$ but fixed at $P(b \mid x, a)$ and $P(a \mid x)$.

Applying the nested regression, we have

$$\begin{aligned}
\mu_0^3(\mathbf{S}_3) = \mu_0^2(B, A, X) &\coloneqq \mathbb{E}_P[Y \mid B, A, X] \\
\check{\mu}_0^3(\mathbf{S}_3) = \check{\mu}_0^3(B, A, X) &\coloneqq \mathbb{E}_P[Y \mid B, A, X].
\end{aligned}$$

Then, consider $\mu_0^2(\mathbf{S}_2) = \mu_0^2(A) = \mathbb{E}_P[\check{\mu}_0^3(B, A, X) \mid A, x]$. It doesn't correctly specify the functional, since

$$\begin{aligned}
\mathbb{E}_P[\check{\mu}_0^3(B, A, X) \mid A, x] &= \mathbb{E}_P[\mu_0^3(B, A, X) \mid A, x] \\
&= \mathbb{E}_P[\mu_0^3(B, A, x) \mid A, x] \\
&= \sum_b \mathbb{E}_P[Y \mid b, A, x] P(b \mid A, x),
\end{aligned}$$

where X in $\mathbb{E}_P[Y \mid b, A, x]$ is fixed, instead of being summed. Therefore, the empirical bifurcation should be applied, and set $\mu_0^2(\mathbf{S}_2) = \mu_0^2(A) = \mathbb{E}_P[\check{\mu}_0^3(B, A, X') \mid A, x]$, where X' is an independent copy of X. This gives

$$\begin{aligned}
\mathbb{E}_P[\check{\mu}_0^3(B, A, X') \mid A, x] &= \mathbb{E}_P[\mu_0^3(B, A, X') \mid A, x] \\
&= \sum_{b,x'} \mathbb{E}_P[Y \mid b, A, x'] P(b \mid A, x) P(x' \mid A, x) \\
&= \sum_{b,x'} \mathbb{E}_P[Y \mid b, A, x'] P(b \mid A, x) P(x'),
\end{aligned}$$

where the last equation holds since X' is an independent copy that is independent with other variables. However, this functional doesn't correctly specify the target functional, in which $P(x' \mid a)$ is summed

instead of $P(x')$. In other words, the nested regression with the empirical bifurcation cannot correctly express this target functional.

To address this issue, a sampler of $P(X \mid A)$ is needed. Equipped with such sampler, let $X'' \sim P(X \mid A)$. Note that X'' is generated only depending on A. Then, set $\mu_0^2(\mathbf{S}_2) = \mu_0^2(A) = \mathbb{E}_P[\check{\mu}_0^3(B, A, X'') \mid A, x]$. This correctly specify the nuisance as follows:

$$\begin{aligned}\mathbb{E}_P[\check{\mu}_0^3(B, A, X'') \mid A, x] &= \mathbb{E}_P[\mu_0^3(B, A, X'') \mid A, x] \\ &= \sum_{b, x''} \mathbb{E}_P[Y \mid b, A, x''] P(b \mid A, x) P(x'' \mid A, x) \\ &= \sum_{b, x''} \mathbb{E}_P[Y \mid b, A, x''] P(b \mid A, x) P(x'' \mid A).\end{aligned}$$

Suppose $\mathbf{S}_i^b \cap \mathbf{C}^{\geq 2} \neq \emptyset$. Specifically, let $\mathbf{S}^k$ denote a non-empty set defined as $\mathbf{S}^k := \mathbf{S}_i^b \cap \mathbf{C}_k$ for some $k \geq 2$. Then, a sampler for $P^k(\mathbf{S}^k \mid \mathbf{S}_{k-1})$ is required to circumvent this issue.

D Proofs

D.1 Proof for Proposition 2

The proof for Proposition 2 is in the proof of Lemma D.1.

D.2 Proof for Proposition 3

$$\begin{aligned}&\mathbb{E}_{P^{i+1}}[\pi_0^m(\mathbf{C}^{(m)} \cup \mathbf{R}^{(m)})Y] \\ &= \mathbb{E}_{P^{i+1}}[\pi_0^m(\mathbf{C}^{(m)} \cup \mathbf{R}^{(m)})\mathbb{E}_{P^{i+1}}[Y \mid \mathbf{S}_m]] \\ &= \sum_{\mathbf{c} \cup \mathbf{r}} \mathbb{E}_{P^{i+1}}[Y \mid \mathbf{s}_m] \pi_0^m(\mathbf{c}^{(m)} \cup \mathbf{r}^{(m)}) P^{i+1}(\mathbf{s}_m) \\ &= \sum_{\mathbf{c} \cup \mathbf{r}} \mathbb{E}_{P^{i+1}}[Y \mid \mathbf{s}_m] \frac{1}{P^{i+1}(\mathbf{s}_m)} P^{i+1}(\mathbf{s}_m) \prod_{i=1}^{m} P^i(\mathbf{c}_i \mid \mathbf{s}_{i-1}) \sigma_{\mathbf{R}_i}^i(\mathbf{r}_i \mid \mathbf{s}_i \setminus \mathbf{r}_i) \\ &= \psi_0.\end{aligned}$$

D.3 Proof for Theorem 1

1. The sample-splitting takes $O((m+1)n_{\max})$.
2. For the fixed ℓ, learning $\hat{\mu}_\ell^i$ for $i = m, \cdots, 1$ takes $O(T_{\boldsymbol{\mu}} \times m)$. Therefore, learning all regression-nuisances takes $O(T_{\boldsymbol{\mu}} \times m \times L)$.
3. For the fixed ℓ, learning $\hat{\pi}_\ell^i$ for $i = 1, \cdots, m$ takes $O(T_{\boldsymbol{\pi}} \times m)$. Therefore, learning all ratio-nuisances takes $O(T_{\boldsymbol{\pi}} \times m \times L)$.
4. Evaluating the DML estimator in Eq. (10) takes $O((m+1)n_{\max})$.

In total, the time complexity is

$$\begin{aligned}&O((m+1)n_{\max}) + O(T_{\boldsymbol{\mu}} \times m \times L) + O(T_{\boldsymbol{\pi}} \times m \times L) + O((m+1)n_{\max}) \\ &= O(m \times \{n_{\max} + L \times (T_{\boldsymbol{\mu}} + T_{\boldsymbol{\pi}})\})\end{aligned}$$

D.4 Proof for Theorem 2

First, $\mu_0^i(\mathbf{S}_i) = \mu_0^i(\mathbf{C}^{(i)} \cup \mathbf{R}^{(i)})$ since $P^j(\mathbf{C}_j \mid \mathbf{S}_{j-1}) = P^j(\mathbf{C}_j \mid \mathbf{C}^{(j-1)} \cup \mathbf{R}^{(j-1)})$ for all $j = 1, \cdots, m+1$.

We first show the following:

$$\frac{\partial P_t^i}{\partial t}\frac{\partial}{\partial P_t^i}\Psi(P^1,\cdots,P_t^i,\cdots,P^{m+1};\boldsymbol{\sigma})\Big|_{t=0} = \mathbb{E}_{Q^i}[\varphi^i(\mathbf{C}^{(i)}\cup\mathbf{R}^{(i-1)};\eta_0^i,\psi_0)]. \tag{D.1}$$

For $i \neq m+1$, it holds as follows: (we will abbreviate $\sigma^k(\mathbf{R}_k \mid \mathbf{x}^{(k-1)}\cup\mathbf{c}^{(k)})$ as σ^k and P^k as $P^k(\mathbf{c}_k \mid \mathbf{c}^{(k-1)},\mathbf{x}^{(k-1)})$)

$$\begin{aligned}
&\frac{\partial P_t^i}{\partial t}\frac{\partial}{\partial P_t^i}\Psi(P^1,\cdots,P_t^i,\cdots,P^{m+1};\boldsymbol{\sigma})\Big|_{t=0}\\
&= \mathbb{E}_{Q^i}\left[\sum_{\mathbf{c}^{(m)},\mathbf{R}}\mu_0^m(\mathbf{c}^{(m)},\mathbf{x}^{(m)})\prod_{k\neq i}P^k\sigma^k\frac{\mathbb{1}_{\mathbf{c}^{(i-1)},\mathbf{x}^{(i-1)}}(\mathbf{C}^{(i-1)},\mathbf{R}^{(i-1)})}{P^i(\mathbf{C}^{(i-1)},\mathbf{R}^{(i-1)})}\{\mathbb{1}_{\mathbf{c}_i}(\mathbf{C}_i)-P^i(\mathbf{c}_i\mid\mathbf{C}^{(i-1)},\mathbf{R}^{(i-1)})\}\right]\\
&= \mathbb{E}_{Q^i}\left[\sum_{\mathbf{c}_i}\check{\mu}_0^i(\mathbf{c}_i,\mathbf{C}^{(i-1)},\mathbf{R}^{(i-1)})\prod_{k=1}^{i-1}P^k(\mathbf{C}_k\mid\mathbf{C}^{(k-1)},\mathbf{R}^{(k-1)})\sigma^k\frac{\mathbb{1}_{\mathbf{c}_i}(\mathbf{C}_i)-P^i(\mathbf{c}_i\mid\mathbf{C}^{(i-1)},\mathbf{R}^{(i-1)})}{P^i(\mathbf{C}^{(i-1)},\mathbf{R}^{(i-1)})}\right]\\
&= \mathbb{E}_{Q^i}\left[\{\check{\mu}_0^i(\mathbf{C}^{(i)},\mathbf{R}^{(i-1)})-\mu_0^{i-1}(\mathbf{C}^{(i-1)},\mathbf{R}^{(i-1)})\}\frac{\prod_{k=1}^{i-1}P^k(\mathbf{C}_k\mid\mathbf{C}^{(k-1)},\mathbf{R}^{(k-1)})\sigma^k(\mathbf{R}_k\mid\mathbf{R}^{(k-1)},\mathbf{C}^{(k)})}{P^i(\mathbf{C}^{(i-1)},\mathbf{R}^{(i-1)})}\right].
\end{aligned}$$

Therefore, Eq. (D.1) holds. Then,

$$\begin{aligned}
\frac{\partial}{\partial t}\Psi(\mathtt{P}^1,\cdots,\mathtt{P}_t^k,\cdots,\mathtt{P}^K)\Big|_{t=0} &= \sum_{i\in\mathcal{I}_k}\frac{\partial P_t^i}{\partial t}\frac{\partial}{\partial P_t^i}\Psi(P^1,\cdots,P_t^i,\cdots,P^{m+1};\boldsymbol{\sigma})\Big|_{t=0}\\
&= \sum_{i\in\mathcal{I}_k}\mathbb{E}_{Q^i}[\varphi^i(\mathbf{C}^{(i)},\mathbf{R}^{(i-1)};\eta_0^i,\psi_0)],
\end{aligned}$$

which completes the proof. ■

D.5 Proof for Theorem 3

Structure of the proof. The structure of the proof is the following.

- Theorem 3 will be proven based on Lemma D.5, Lemma D.7, and Lemma D.8.
- Lemma D.5 will be proven based on helper lemmas (D.1, D.2, D.3, D.4).

Therefore, we proceed the proof as follows:

1. We will prove helper lemmas (D.1, D.2, D.3, D.4).
2. Main lemmas in Lemma D.5, Lemma D.7, and Lemma D.8 will be proved based on the helper lemmas.
3. Berry-Essen's inequality (Berry, 1941) will be stated as a preliminary in Prop. D.1.
4. Theorem 3 will be proven based on the main lemmas and Berry-Essen's inequality.

Notation. First, $\mu_0^i(\mathbf{S}_i) = \mu_0^i(\mathbf{C}^{(i)}\cup\mathbf{R}^{(i)})$ since $P^j(\mathbf{C}_j \mid \mathbf{S}_{j-1}) = P^j(\mathbf{C}_j \mid \mathbf{C}^{(j-1)}\cup\mathbf{R}^{(j-1)})$ for all $j = 1,\cdots,m+1$. This equation leads to write $\pi_0^i(\mathbf{C}_1\cup\mathbf{S}^{(i)})$ as $\pi_0^i(\mathbf{C}^{(i)}\cup\mathbf{R}^{(i)})$.

We will use the following notation. For $i > 2$,

$$\omega_0^i \coloneqq \frac{\pi_0^i}{\pi_0^{i-1}}, \quad \text{and} \quad \hat{\omega}_\ell^i \coloneqq \frac{\hat{\pi}_\ell^i}{\hat{\pi}_\ell^{i-1}}, \tag{D.2}$$

and $\omega_0^1 \coloneqq \pi_0^1$ and $\hat{\omega}_\ell^1 \coloneqq \hat{\pi}_\ell^1$. Then,

$$\pi_0^i \coloneqq \omega_0^{(i)} \coloneqq \prod_{j=1}^{i} \omega_0^j, \tag{D.3}$$

$$\hat{\pi}_\ell^i \coloneqq \hat{\omega}_\ell^{(i)} \coloneqq \prod_{j=1}^{i} \hat{\omega}_\ell^j. \tag{D.4}$$

D.5.1 Helper lemmas

We first state and prove helper lemmas.

Lemma D.1.

$$\Psi(\mathbf{P};\boldsymbol{\sigma}) = \sum_{i=1}^{m} \mathbb{E}_{P^{i+1}}[\omega_0^{(i)}\{\breve{\mu}_0^{i+1} - \mu_0^i\}] + \mathbb{E}_{P^1}[\breve{\mu}_0^1]. \tag{D.5}$$

Proof of Lemma D.1. For $i = 1, \cdots, m$

$$\begin{aligned}
&\mathbb{E}_{P^{i+1}}[\omega^{(i)}(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\{\breve{\mu}_0^{i+1}(\mathbf{C}^{(i+1)},\mathbf{R}^{(i)}) - \mu_0^i(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\}]\\
&= \mathbb{E}_{P^{i+1}}[\omega^{(i)}(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\{\breve{\mu}_0^{i+1}(\mathbf{C}^{(i+1)},\mathbf{R}^{(i)}) - \mu_0^i(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\}]\\
&= \mathbb{E}_{P^{i+1}}[\omega^{(i)}(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\{\breve{\mu}_0^{i+1}(\mathbf{C}^{(i+1)},\mathbf{R}^{(i)}) - \mu_0^i(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\}]\\
&= \mathbb{E}_{P^{i+1}}[\mathbb{E}_{P^{i+1}}[\omega^{(i)}(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\{\breve{\mu}_0^{i+1}(\mathbf{C}^{(i+1)},\mathbf{R}^{(i)}) - \mu_0^i(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\}|\mathbf{C}^{(i)},\mathbf{R}^{(i)}]]\\
&= \mathbb{E}_{P^{i+1}}[\omega^{(i)}(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\mathbb{E}_{P^{i+1}}[\{\breve{\mu}_0^{i+1}(\mathbf{C}^{(i+1)},\mathbf{R}^{(i)})|\mathbf{C}^{(i)},\mathbf{R}^{(i)}] - \omega^{(i)}(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\mu_0^i(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\}]\\
&= \mathbb{E}_{P^{i+1}}[\omega^{(i)}(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\mu_0^i(\mathbf{C}^{(i)},\mathbf{R}^{(i)}) - \omega^{(i)}(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\mu_0^i(\mathbf{C}^{(i)},\mathbf{R}^{(i)})\}]\\
&= 0.
\end{aligned}$$

Therefore, it suffices to show that

$$\Psi(\mathbf{P};\boldsymbol{\sigma}) = \mathbb{E}_{P^1}[\breve{\mu}_0^1(\mathbf{C}_1)].$$

It can be shown that

$$\begin{aligned}
&\mathbb{E}_{P^1}[\breve{\mu}_0^1(\mathbf{C}_1)]\\
&= \sum_{\mathbf{c}_1} \breve{\mu}_0^1(\mathbf{c}_1)P^1(\mathbf{c}_1)\\
&= \sum_{\mathbf{c}_1,\mathbf{R}_1} \mu_0^1(\mathbf{c}_1,\mathbf{R}_1)P^1(\mathbf{c}_1)\sigma^1(\mathbf{R}_1 \mid \mathbf{c}_1)\\
&= \sum_{\mathbf{c}^{(2)},\mathbf{R}_1} \breve{\mu}_0^2(\mathbf{c}_2,\mathbf{R}_1)P^1(\mathbf{c}_1)P^2(\mathbf{c}_2 \mid \mathbf{c}_1,\mathbf{R}_1)\sigma^1(\mathbf{R}_1 \mid \mathbf{c}_1)\\
&= \sum_{\mathbf{c}^{(2)},\mathbf{x}^{(2)}} \mu_0^2(\mathbf{c}_2,\mathbf{R}_1)P^1(\mathbf{c}_1)P^2(\mathbf{c}_2 \mid \mathbf{c}_1,\mathbf{R}_1)\sigma^1(\mathbf{R}_1 \mid \mathbf{c}_1)\sigma^2(\mathbf{R}_2 \mid \mathbf{c}^{(2)},\mathbf{R}_1)\\
&\cdots\\
&= \sum_{\mathbf{c}^{(m)},\mathbf{R}} \mu_0^m(\mathbf{c}_m,\mathbf{R}) \prod_{i=1}^{m} P^m(\mathbf{c}_m \mid \mathbf{c}^{(m-1)},\mathbf{x}^{(m-1)}) \prod_{j=1}^{m} \sigma^j(\mathbf{R}_j \mid \mathbf{c}^{(j)},\mathbf{x}^{(j-1)})\\
&= \Psi(\mathbf{P};\boldsymbol{\sigma}).
\end{aligned}$$

□

Lemma D.2. *For $i = 2, \cdots, m$,*

$$\begin{aligned}&\mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\mu}_0^{i+1} - \hat{\mu}^i\}] + \mathbb{E}_{P^i}[\hat{\omega}^{(i-1)}\{\check{\hat{\mu}}^i - \hat{\mu}^{i-1}\}] \\ &= \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i-1)}\{\mu_0^i - \hat{\mu}^i\}\{\hat{\omega}^i - \omega_0^i\}] + \mathbb{E}_{P^i}[\hat{\omega}^{(i-1)}\{\check{\mu}_0^i - \hat{\mu}^{i-1}\}].\end{aligned} \tag{D.6}$$

Proof of Lemma D.2. We first rewrite Eq. (D.6) as follows:

$$\text{Eq. (D.6)} = \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\mu}_0^{i+1} - \hat{\mu}^i\}] \tag{D.7}$$

$$+ \mathbb{E}_{P^i}[\hat{\omega}^{(i-1)}\{\check{\hat{\mu}}^i - \check{\hat{\mu}}_0^i\}] \tag{D.8}$$

$$+ \mathbb{E}_{P^i}[\hat{\omega}^{(i-1)}\{\check{\hat{\mu}}_0^i - \hat{\mu}^{i-1}\}]. \tag{D.9}$$

Then, Eq. (D.8) can be represented as follow:

$$\begin{aligned}&\text{Eq. (D.8)} \\ &= \mathbb{E}_{P^i}[\hat{\omega}^{(i-1)}(\mathbf{C}^{(i-1)}, \mathbf{R}^{(i-1)})\{\check{\hat{\mu}}^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)}) - \check{\mu}_0^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)})\}] \\ &= \mathbb{E}_{P^{i+1}}\left[\hat{\omega}^{(i-1)}(\mathbf{C}^{(i-1)}, \mathbf{R}^{(i-1)})\frac{P^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)})}{P^{i+1}(\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)})}\{\check{\hat{\mu}}^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)}) - \check{\mu}_0^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)})\}\right] \\ &= \mathbb{E}_{P^{i+1}}\left[\hat{\omega}^{(i-1)}(\mathbf{C}^{(i-1)}, \mathbf{R}^{(i-1)})\frac{P^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)})}{P^{i+1}(\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)})}\mathbb{E}_{\sigma^i}[\mu^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i)}) - \mu_0^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i)})|\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)}]\right] \\ &= \mathbb{E}_{P^{i+1}}\left[\hat{\omega}^{(i-1)}(\mathbf{C}^{(i-1)}, \mathbf{R}^{(i-1)})\frac{P^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)})}{P^{i+1}(\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)})}\frac{\sigma^i(\mathbf{R}_i \mid \mathbf{C}^{(i)}, \mathbf{R}^{(i-1)})}{P^{i+1}(\mathbf{R}_i|\mathbf{C}^{(i)}, \mathbf{R}^{(i-1)})}\{\mu^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i)}) - \mu_0^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i)})\}\right] \\ &= \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i-1)}(\mathbf{C}^{(i-1)}, \mathbf{R}^{(i-1)})\omega_0^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i)})\{\mu^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i)}) - \mu_0^i(\mathbf{C}^{(i)}, \mathbf{R}^{(i)})\}].\end{aligned} \tag{D.10}$$

Therefore,

$$\begin{aligned}&\text{Eq. (D.7)} + \text{Eq. (D.8)} \\ &= \text{Eq. (D.7)} + \text{Eq. (D.10)} \\ &= \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\mu}_0^{i+1} - \hat{\mu}^i\}] + \text{Eq. (D.10)} \\ &= \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\mu_0^i - \hat{\mu}^i\}] + \text{Eq. (D.10)} \\ &= \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\mu_0^i - \hat{\mu}^i\}] + \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i-1)}\omega_0^i\{\hat{\mu}^i - \mu_0^i\}] \\ &= \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i-1)}\{\mu_0^i - \hat{\mu}^i\}\{\hat{\omega}^i - \omega_0^i\}].\end{aligned} \tag{D.11}$$

Finally,

$$\text{Eq. (D.6)} = \text{Eq. (D.11)} + \text{Eq. (D.9)}.$$

□

Lemma D.3. *Define the following*

$$\Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\omega}}) \coloneqq \sum_{i=1}^{m} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^i\}] + \mathbb{E}_{P^1}[\check{\hat{\mu}}^1] \tag{D.12}$$

$$\Phi(\boldsymbol{\mu_0}, \hat{\boldsymbol{\omega_0}}) \coloneqq \sum_{i=1}^{m} \mathbb{E}_{P^{i+1}}[\omega_0^{(i)}\{\check{\mu}_0^{i+1} - \mu_0^i\}] + \mathbb{E}_{P^1}[\check{\mu}_0^1]. \tag{D.13}$$

For $k = 3, \cdots, m$, the following holds:

$$\begin{aligned}
&\Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\omega}}) - \Phi(\boldsymbol{\mu_0}, \hat{\boldsymbol{\omega_0}}) \\
&= \sum_{r=k}^{m} \mathbb{E}_{P^{r+1}}[\hat{\omega}^{(r-1)}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}] + \mathbb{E}_{P^k}[\hat{\omega}^{(k-1)}\{\check{\mu}_0^k - \hat{\mu}^{k-1}\}] \\
&+ \sum_{i=1}^{k-2} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^i\}] + \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1].
\end{aligned}$$

Proof of Lemma D.3 The equation holds for $k = m$. It can be shown as follows:

$$\begin{aligned}
&\Phi(\{\hat{\mu}^i, \hat{\omega}^i : i = 1, \cdots, m\}) - \Phi(\{\mu_0^i, \omega_0^i : i = 1, \cdots, m\}) \\
&\overset{\text{Lemma D.1}}{=} \sum_{i=1}^{m} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^i\}] + \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1] \\
&= \mathbb{E}_{P^{m+1}}[\hat{\omega}^{(m)}\{\check{\mu}_0^{m+1} - \hat{\mu}^m\}] + \mathbb{E}_{P^m}[\hat{\omega}^{(m-1)}\{\check{\hat{\mu}}^m - \hat{\mu}^{m-1}\}] && \text{(D.14)} \\
&+ \sum_{i=1}^{m-2} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^i\}] + \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1]. && \text{(D.15)}
\end{aligned}$$

Then,

$$\begin{aligned}
&\text{Eq. (D.14)} \\
&\overset{\text{Lemma D.2}}{=} \mathbb{E}_{P^{m+1}}[\hat{\omega}^{(m-1)}\{\mu_0^m - \hat{\mu}^m\}\{\hat{\omega}^m - \omega_0^m\}] + \mathbb{E}_{P^m}[\hat{\omega}^{(m-1)}\{\check{\mu}_0^m - \hat{\mu}^{m-1}\}]. && \text{(D.16)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\Phi(\{\hat{\mu}^i, \hat{\omega}^i : i = 1, \cdots, m\}) - \Phi(\{\mu_0^i, \omega_0^i : i = 1, \cdots, m\}) \\
&= \text{Eq. (D.16)} + \sum_{i=1}^{m-2} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^i\}] + \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1]. && \text{(D.17)}
\end{aligned}$$

We make a following induction hypothesis. For any fixed $k + 1 \in \{m, m-1, \cdots, 4\}$, the following holds:

$$\begin{aligned}
&\Phi(\{\hat{\mu}^i, \hat{\omega}^i : i = 1, \cdots, m\}) - \Phi(\{\mu_0^i, \omega_0^i : i = 1, \cdots, m\}) && \text{(D.18)} \\
&= \sum_{r=k+1}^{m} \mathbb{E}_{P^{r+1}}[\hat{\omega}^{(r-1)}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}] && \text{(D.19)} \\
&+ \mathbb{E}_{P^{k+1}}[\hat{\omega}^{(k)}\{\check{\mu}_0^{k+1} - \hat{\mu}^k\}] && \text{(D.20)} \\
&+ \sum_{i=1}^{k-1} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^i\}] && \text{(D.21)} \\
&+ \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1]. && \text{(D.22)}
\end{aligned}$$

We note that this holds when $k = m - 1$, as shown in Eq. (D.17). We will now show that it will hold for k, too. First,

$$\text{Eq. (D.20)} + \text{Eq. (D.21)} \tag{D.23}$$

$$\begin{aligned}
&\mathbb{E}_{P^{k+1}}[\hat{\omega}^{(k)}\{\check{\mu}_0^{k+1} - \hat{\mu}^k\}] + \sum_{i=1}^{k-1} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^i\}] \\
&= \mathbb{E}_{P^{k+1}}[\hat{\omega}^{(k)}\{\check{\mu}_0^{k+1} - \hat{\mu}^k\}] + \mathbb{E}_{P^k}[\hat{\omega}^{(k-1)}\{\check{\hat{\mu}}^k - \hat{\mu}^{k-1}\}] + \sum_{i=1}^{k-2} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^i\}], \\
&\overset{\text{Lemma D.2}}{=} \mathbb{E}_{P^{k+1}}[\hat{\omega}^{(k-1)}\{\mu_0^k - \hat{\mu}^k\}\{\hat{\omega}^k - \omega_0^k\}] + \mathbb{E}_{P^k}[\hat{\omega}^{(k-1)}\{\check{\mu}_0^k - \hat{\mu}^{k-1}\}] + \sum_{i=1}^{k-2} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^i\}].
\end{aligned} \tag{D.24}$$

Therefore,

$$\begin{aligned}
&\Phi(\{\hat{\mu}^i, \hat{\omega}^i : i = 1, \cdots, m\}) - \Phi(\{\mu_0^i, \omega_0^i : i = 1, \cdots, m\}) \\
&= \text{Eq. (D.19)} + \text{Eq. (D.20)} + \text{Eq. (D.21)} + \text{Eq. (D.22)} \\
&= \text{Eq. (D.19)} + \text{Eq. (D.24)} + \text{Eq. (D.22)} \\
&= \sum_{r=k+1}^{m} \mathbb{E}_{P^{r+1}}[\hat{\omega}^{(r-1)}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}] + \mathbb{E}_{P^{k+1}}[\hat{\omega}^{(k-1)}\{\mu_0^k - \hat{\mu}^k\}\{\hat{\omega}^k - \omega_0^k\}] \\
&+ \mathbb{E}_{P^k}[\hat{\omega}^{(k-1)}\{\check{\mu}_0^k - \hat{\mu}^{k-1}\}] \\
&+ \sum_{i=1}^{k-2} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^i\}] \\
&+ \text{Eq. (D.22)}
\end{aligned}$$

This means that the induction hypothesis holds for k, too. This completes the proof. □

Lemma D.4.
$$\mathbb{E}_{P^2}[\hat{\omega}^1\{\check{\mu}_0^2 - \hat{\mu}^1\}] - \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1] = \mathbb{E}_{P^2}[\{\mu_0^1 - \hat{\mu}^1\}\{\hat{\omega}^1 - \omega_0^1\}].$$

Proof of Lemma D.4

$$\begin{aligned}
&\mathbb{E}_{P^1}[\hat{\mu}^1(\mathbf{C}_1) - \check{\mu}_0^1(\mathbf{C}_1)] \\
&= \mathbb{E}_{P^2}\left[\frac{P^1(\mathbf{C}_1)}{P^2(\mathbf{C}_1)}\{\check{\hat{\mu}}^1(\mathbf{C}_1) - \check{\mu}_0^1(\mathbf{C}_1)\}\right] \\
&= \mathbb{E}_{P^2}\left[\frac{P^1(\mathbf{C}_1)}{P^2(\mathbf{C}_1)}\mathbb{E}_{\sigma^1}[\hat{\mu}^1(\mathbf{C}_1, \mathbf{R}_1) - \mu_0^1(\mathbf{C}_1, \mathbf{R}_1)|\mathbf{C}_1]\right] \\
&= \mathbb{E}_{P^2}\left[\frac{P^1(\mathbf{C}_1)}{P^2(\mathbf{C}_1)}\mathbb{E}_{P^2}\left[\frac{\sigma^1(\mathbf{R}_1|\mathbf{C}_1)}{P^2(\mathbf{R}_1|\mathbf{C}_1)}\{\hat{\mu}^1(\mathbf{C}_1, \mathbf{R}_1) - \mu_0^1(\mathbf{C}_1, \mathbf{R}_1)\}\middle|\mathbf{C}_1\right]\right] \\
&= \mathbb{E}_{P^2}\left[\frac{P^1(\mathbf{C}_1)}{P^2(\mathbf{C}_1)}\frac{\sigma^1(\mathbf{R}_1|\mathbf{C}_1)}{P^2(\mathbf{R}_1|\mathbf{C}_1)}\{\hat{\mu}^1(\mathbf{C}_1, \mathbf{R}_1) - \mu_0^1(\mathbf{C}_1, \mathbf{R}_1)\}\right] \\
&= \mathbb{E}_{P^2}[\omega_0^1(\mathbf{C}_1, \mathbf{R}_1)\{\hat{\mu}^1(\mathbf{C}_1, \mathbf{R}_1) - \mu_0^1(\mathbf{C}_1, \mathbf{R}_1)\}].
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{P^2}[\hat{\omega}^1\{\check{\mu}_0^2 - \hat{\mu}^1\}] - \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1] \\
&= \mathbb{E}_{P^2}[\hat{\omega}^1\{\mu_0^1 - \hat{\mu}^1\}] - \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1] \\
&= \mathbb{E}_{P^2}[\hat{\omega}^1\{\mu_0^1 - \hat{\mu}^1\}] - \mathbb{E}_{P^2}[\omega_0^1\{\mu_0^1 - \hat{\mu}^1\}] \\
&= \mathbb{E}_{P^2}[\{\mu_0^1 - \hat{\mu}^1\}\{\hat{\omega}^1 - \omega_0^1\}].
\end{aligned}$$

□

D.5.2 Main lemmas

Based on the above helper lemmas in the box, the main lemma is stated and proven as follows:

Lemma D.5 (**Decomposition**-1)**.** *Define the following*

$$\Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\omega}}) \coloneqq \sum_{i=1}^{m} \mathbb{E}_{P^{i+1}}[\hat{\omega}^{(i)}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^{i}\}] + \mathbb{E}_{P^1}[\check{\hat{\mu}}^1] \tag{D.25}$$

$$\Phi(\boldsymbol{\mu_0}, \boldsymbol{\omega_0}) \coloneqq \sum_{i=1}^{m} \mathbb{E}_{P^{i+1}}[\omega_0^{(i)}\{\check{\mu}_0^{i+1} - \mu_0^{i}\}] + \mathbb{E}_{P^1}[\check{\mu}_0^1]. \tag{D.26}$$

The following decomposition holds:

$$\Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\omega}}) - \Phi(\boldsymbol{\mu_0}, \hat{\boldsymbol{\omega}}\boldsymbol{_0}) = \sum_{r=1}^{m} \mathbb{E}_{P^{r+1}}[\hat{\omega}^{(r-1)}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}]. \tag{D.27}$$

Proof of Lemma D.5

$$\begin{aligned}
&\Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\omega}}) - \Phi(\boldsymbol{\mu_0}, \hat{\boldsymbol{\omega}}\boldsymbol{_0}) \\
&\overset{\text{Lemma D.3}(k=3)}{=\!=} \sum_{r=3}^{m} \mathbb{E}_{P^{r+1}}[\hat{\omega}^{(r-1)}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}] \qquad \text{(D.28)} \\
&+ \mathbb{E}_{P^3}[\hat{\omega}^{(2)}\{\check{\mu}_0^3 - \hat{\mu}^2\}] + \mathbb{E}_{P^2}[\hat{\omega}^1\{\check{\mu}^2 - \mu^1\}] + \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1] \\
&\overset{\text{Lemma D.2}}{=\!=} \text{Eq. (D.28)} + \mathbb{E}_{P^3}[\hat{\omega}^1\{\mu_0^2 - \hat{\mu}^2\}\{\hat{\omega}^2 - \omega_0^2\}] + \mathbb{E}_{P^2}[\hat{\omega}^1\{\check{\mu}_0^2 - \hat{\mu}^1\}] + \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1] \\
&= \sum_{r=2}^{m} \mathbb{E}_{P^{r+1}}[\hat{\omega}^{(r-1)}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}] + \mathbb{E}_{P^2}[\hat{\omega}^1\{\check{\mu}_0^2 - \hat{\mu}^1\}] + \mathbb{E}_{P^1}[\check{\hat{\mu}}^1 - \check{\mu}_0^1] \\
&\overset{\text{Lemma D.4}}{=\!=} \sum_{r=2}^{m} \mathbb{E}_{P^{r+1}}[\hat{\omega}^{(r-1)}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}] + \mathbb{E}_{P^2}[\{\mu_0^1 - \hat{\mu}^1\}\{\hat{\omega}^1 - \omega_0^1\}] \\
&= \sum_{r=1}^{m} \mathbb{E}_{P^{r+1}}[\hat{\omega}^{(r-1)}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}].
\end{aligned}$$

□

Lemma D.6 (**Decomposition**-2)**.** *Define the following*

$$\Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}) \coloneqq \sum_{i=1}^{m} \mathbb{E}_{P^{i+1}}[\hat{\pi}^{i}\{\check{\hat{\mu}}^{i+1} - \hat{\mu}^{i}\}] + \mathbb{E}_{P^1}[\check{\hat{\mu}}^1] \tag{D.29}$$

$$\Phi(\boldsymbol{\mu_0}, \boldsymbol{\pi}_0) \coloneqq \sum_{i=1}^{m} \mathbb{E}_{P^{i+1}}[\pi_0^{i}\{\check{\mu}_0^{i+1} - \mu_0^{i}\}] + \mathbb{E}_{P^1}[\check{\mu}_0^1]. \tag{D.30}$$

The following decomposition holds:

$$\Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}) - \Phi(\boldsymbol{\mu_0}, \boldsymbol{\pi}_0) = \sum_{r=1}^{m} \mathbb{E}_{P^{r+1}}[\{\mu_0^r - \hat{\mu}^r\}\{\hat{\pi}^r - \pi_0^r\}] \tag{D.31}$$

$$+ \sum_{r=2}^{m} \mathbb{E}_{P^{r+1}}[\omega_0^r\{\mu_0^r - \hat{\mu}^r\}\{\hat{\pi}^{r-1} - \pi_0^{r-1}\}]. \tag{D.32}$$

Proof of Lemma D.6 Define A is the term satisfying

$$\Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}) - \Phi(\boldsymbol{\mu_0}, \boldsymbol{\pi}_0) = \sum_{r=1}^{m} \mathbb{E}_{P^{r+1}}[\{\mu_0^r - \hat{\mu}^r\}\{\hat{\pi}^r - \pi_0^r\}] + A. \tag{D.33}$$

From Lemma D.5 and the fact that $\Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}) = \Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\omega}})$ and $\Phi(\boldsymbol{\mu_0}, \boldsymbol{\pi}_0) = \Phi(\boldsymbol{\mu_0}, \boldsymbol{\omega}_0)$ by definition, we have

$$\Phi(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\pi}}) - \Phi(\boldsymbol{\mu_0}, \boldsymbol{\pi}_0) = \sum_{r=1}^{m} \mathbb{E}_{P^{r+1}}[\hat{\omega}^{(r-1)}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}] \tag{D.34}$$

$$= \sum_{r=1}^{m} \mathbb{E}_{P^{r+1}}[\hat{\pi}^{r-1}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}] \tag{D.35}$$

That is,

$$A = \sum_{r=1}^{m} \mathbb{E}_{P^{r+1}}[\hat{\pi}^{r-1}\{\mu_0^r - \hat{\mu}^r\}\{\hat{\omega}^r - \omega_0^r\}] - \sum_{r=1}^{m} \mathbb{E}_{P^{r+1}}[\{\mu_0^r - \hat{\mu}^r\}\{\hat{\pi}^r - \pi_0^r\}] \tag{D.36}$$

$$= \sum_{r=2}^{m} \left(\mathbb{E}_{P^{r+1}}[\hat{\pi}^r\{\mu_0^r - \hat{\mu}^r\}] - \mathbb{E}_{P^{r+1}}[\hat{\pi}^{r-1}\omega_0^r\{\mu_0^r - \hat{\mu}^r\}] + \mathbb{E}_{P^{r+1}}[\{\mu_0^r - \hat{\mu}^r\}\{\hat{\pi}^r - \pi_0^r\}]\right) \tag{D.37}$$

$$= \sum_{r=2}^{m} \mathbb{E}_{P^{i+1}}[\omega_0^r\{\mu_0^r - \hat{\mu}^r\}\{\pi_0^{r-1} - \hat{\pi}^{r-1}\}]. \tag{D.38}$$

This completes the proof. □

Lemma D.7 (Stochastic Equicontinuity). *Let $\mathcal{D} \overset{iid}{\sim} P$. Let $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$, where $n \coloneqq |\mathcal{D}_0|$. Let $\hat{f}$ be a function estimated from $\mathcal{D}_1$. Then, in probability greater than $1-\epsilon$ for any $\epsilon \in (0,1)$,*

$$\mathbb{E}_{\mathcal{D}_0 - P}\left[\left|\hat{f} - f\right|\right] \overset{w.p\ 1-\epsilon}{<} \frac{\|\hat{f} - f\|_P}{\sqrt{n\epsilon}}, \tag{D.39}$$

which implies that

$$\mathbb{E}_{\mathcal{D}_0 - P}[|\hat{f} - f|] = O_P\left(\frac{\|\hat{f} - f\|_P}{\sqrt{n}}\right).$$

Proof of Lemma D.7 This proof is from (Kennedy et al., 2020, Lemma 2). Since $\hat{f}$ is a function of $\mathcal{D}_1$, we will denote $\hat{f}_{\mathcal{D}_1}$. Define a following random variable of interest:

$$X \coloneqq \mathbb{E}_{\mathcal{D}_0 - P}[\hat{f}_{\mathcal{D}_1} - f].$$

Then, the conditional expectation of X given $\mathcal{D}_1$ is zero, since

$$\mathbb{E}_P\left[\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{\mathcal{D}_1}(\mathbf{V}_i) \,\middle|\, \mathcal{D}_1\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_P[\hat{f}_{\mathcal{D}_1}(\mathbf{V}_i) \mid \mathcal{D}_1] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_P[\hat{f}_{\mathcal{D}_1}(\mathbf{V}) \mid \mathcal{D}_1] = \mathbb{E}_P[\hat{f}_{\mathcal{D}_1}(\mathbf{V}) \mid \mathcal{D}_1],$$

where the third equality holds by the independence of $\mathcal{D}_0$ and $\mathcal{D}_1$. Therefore,

$$\begin{aligned}
\mathbb{E}_P[X \mid \mathcal{D}_1] &= \mathbb{E}_P[\mathbb{E}_{\mathcal{D}_0 - P}[\hat{f}_{\mathcal{D}_1} - f] \mid \mathcal{D}_1] \\
&= \mathbb{E}_P[\mathbb{E}_{\mathcal{D}_0}[\hat{f}_{\mathcal{D}_1} - f] \mid \mathcal{D}_1] - \mathbb{E}_P[\mathbb{E}_P[\hat{f}_{\mathcal{D}_1} - f] \mid \mathcal{D}_1] \\
&= \mathbb{E}_P[\mathbb{E}_P[\hat{f}_{\mathcal{D}_1} - f] \mid \mathcal{D}_1] - \mathbb{E}_P[\mathbb{E}_P[\hat{f}_{\mathcal{D}_1} - f] \mid \mathcal{D}_1] = 0.
\end{aligned}$$

Also,

$$\begin{aligned}\mathbb{V}_P[X \mid \mathcal{D}_1] &= \mathbb{V}_P[\mathbb{E}_{\mathcal{D}_0 - P}[\hat{f}_{\mathcal{D}_1} - f] \mid \mathcal{D}_1] \\ &= \mathbb{V}_P[\mathbb{E}_{\mathcal{D}_0}[\hat{f}_{\mathcal{D}_1} - f] \mid \mathcal{D}_1] \\ &= \frac{1}{n}\mathbb{V}_P[\hat{f}_{\mathcal{D}_1} - f \mid \mathcal{D}_1] \\ &\leq \frac{1}{n}\|\hat{f}_{\mathcal{D}_1} - f\|_P^2.\end{aligned}$$

By applying the (conditional-) Chevyshev's inequality,

$$P(|X - \mathbb{E}_P[X \mid \mathcal{D}_1]| \geq t \mid \mathcal{D}_1) \leq \frac{1}{t^2}\mathbb{V}_P[X \mid \mathcal{D}_1] \leq \frac{1}{nt^2}\|\hat{f}_{\mathcal{D}_1} - f\|_P^2.$$

Then,

$$\begin{aligned}P(|X| \geq t) &= P(|X - \mathbb{E}_P[X \mid \mathcal{D}_1]| \geq t) \\ &= \mathbb{E}_{P(\mathcal{D}_1)}[P(|X - \mathbb{E}_P[X \mid \mathcal{D}_1]| \geq t \mid \mathcal{D}_1)] \\ &\leq \frac{1}{nt^2}\|\hat{f}_{\mathcal{D}_1} - f\|_P^2.\end{aligned}$$

In other words, $X < t$ in probability greater than $1 - \frac{1}{nt^2}\|\hat{f}_{\mathcal{D}_1} - f\|_P^2$. If $t = \frac{\|\hat{f}_{\mathcal{D}_1} - f\|_P}{\sqrt{n\epsilon}}$, then $X < \frac{\|\hat{f}_{\mathcal{D}_1} - f\|_P}{\sqrt{n\epsilon}}$ in the probability greater than $1 - \epsilon$ for any $\epsilon \in (0, 1)$. □

Lemma D.8 (Combining concentration inequalities). *Suppose $P(A_k > t) \leq b_k/t^2$ for $k = 1, \cdots, K$. Then,*

$$P\left(\sum_{k=1}^{K} A_k \leq tK\right) \geq 1 - \frac{1}{t^2}\sum_{k=1}^{K} b_k.$$

Proof. The event $\sum_{k=1}^{K} A_k \leq tK$ includes the case where $A_k < t$ for $k = 1, \cdots, K$. Therefore,

$$\begin{aligned}P\left(\sum_{k=1}^{K} A_k \leq tK\right) &\geq P\left(A_1 \leq t \text{ and } \cdots \text{ and } A_K \leq t\right) \\ &= 1 - P\left(A_1 > t \text{ or } \cdots \text{ or } A_K > t\right) \\ &\geq 1 - \sum_{k=1}^{K} P\left(A_k > t\right) \\ &\geq 1 - \sum_{k=1}^{K} \frac{b_k}{t^2}.\end{aligned}$$

□

D.5.3 Preliminary Results

Proposition D.1 (Berry–Esseen's inequality (Berry, 1941; Esseen, 1942; Shevtsova, 2014)). *Suppose $\mathcal{D} = \{X_1, \cdots, X_n\}$ are independent and identically distributed random variables with $\mathbb{E}_P[X_i] = 0$, $\mathbb{E}_P[X_i^2] = \sigma^2$ and $\mathbb{E}_P[|X_i|^3] = \kappa^3$. Then, for all x and n,*

$$\left|P\left(\frac{\sqrt{n}}{\sigma_0}\mathbb{E}_{\mathcal{D}}[X] < x\right) - \Phi(x)\right| \leq \frac{0.4748\kappa^3}{\sigma^3\sqrt{n}}.$$

D.5.4 Proof of Theorem 3

By Lemma D.5, we decompose the error as follow:

$$\hat{\psi} - \psi_0 = \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}^k - \mathrm{P}^k}[\phi_0^k] \tag{D.40}$$

$$+ \frac{1}{L} \sum_{\ell=1}^{L} \sum_{k=1}^{K} \mathbb{E}_{\mathcal{D}_\ell^k - \mathrm{P}^k}[\hat{\phi}_\ell^k - \phi_0^k] \tag{D.41}$$

$$+ \frac{1}{L} \sum_{\ell=1}^{L} \sum_{i=1}^{m} \mathbb{E}_{P^{i+1}}[\hat{\omega}_\ell^{(i-1)}\{\mu_0^i - \hat{\mu}_\ell^i\}\{\hat{\omega}_\ell^i - \omega_0^i\}]. \tag{D.42}$$

By Lemma D.7,

$$\mathrm{P}^k\left(\left|\mathbb{E}_{\mathcal{D}_\ell^k - \mathrm{P}^k}[\hat{\phi}_\ell^k - \phi_0^k]\right| > t\right) = \mathrm{P}^k\left(\frac{1}{L}\left|\mathbb{E}_{\mathcal{D}_\ell^k - \mathrm{P}^k}[\hat{\phi}_\ell^k - \phi_0^k]\right| > Lt\right) \leq \frac{1}{t^2} \frac{\|\hat{\phi}_\ell^k - \phi_0^k\|_{\mathrm{P}^k}^2}{|\mathcal{D}_\ell^k|}. \tag{D.43}$$

By Lemma D.8,

$$\mathrm{P}^k\left(\frac{1}{L}\sum_{\ell=1}^{L}\left|\mathbb{E}_{\mathcal{D}_\ell^k - \mathrm{P}^k}[\hat{\phi}_\ell^k - \phi_0^k]\right| \leq Lt\right) \geq 1 - \frac{1}{t^2}\sum_{\ell=1}^{L}\frac{\|\hat{\phi}_\ell^k - \phi_0^k\|_{\mathrm{P}^k}^2}{|\mathcal{D}_\ell^k|}. \tag{D.44}$$

Equivalently, by choosing $t = \sqrt{\frac{1}{\epsilon}\sum_{\ell=1}^{L}\frac{\|\hat{\phi}_\ell^k - \phi_0^k\|_{\mathrm{P}^k}^2}{|\mathcal{D}_\ell^k|}}$,

$$\frac{1}{L}\sum_{\ell=1}^{L}\left|\mathbb{E}_{\mathcal{D}_\ell^k - \mathrm{P}^k}[\hat{\phi}_\ell^k - \phi_0^k]\right| \overset{\text{w.p } 1-\epsilon}{\leq} \sqrt{\frac{L^2}{\epsilon}\sum_{\ell=1}^{L}\frac{\|\hat{\phi}_\ell^k - \phi_0^k\|_{\mathrm{P}^k}^2}{|\mathcal{D}_\ell^k|}}. \tag{D.45}$$

Define

$$A^k := \mathbb{E}_{\mathcal{D}^k - \mathrm{P}^k}[\phi_0^k] \tag{D.46}$$

$$B^k := \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_{\mathcal{D}_\ell^k - \mathrm{P}^k}[\hat{\phi}_\ell^k - \phi_0^k] \tag{D.47}$$

$$C^k := \frac{1}{L}\sum_{\ell=1}^{L}\left|\mathbb{E}_{\mathcal{D}_\ell^k - \mathrm{P}^k}[\hat{\phi}_\ell^k - \phi_0^k]\right| \tag{D.48}$$

$$\Delta_k := \sqrt{\frac{L^2}{\epsilon}\sum_{\ell=1}^{L}\frac{\|\hat{\phi}_\ell^k - \phi_0^k\|_{\mathrm{P}^k}^2}{|\mathcal{D}_\ell^k|}}. \tag{D.49}$$

Here,

$$R^k := A^k + B^k. \tag{D.50}$$

Then,

$$\mathtt{P}^k\left(R^k < x\right) \tag{D.51}$$

$$= \mathtt{P}^k\left(\mathbb{E}_{\mathcal{D}^k - \mathtt{P}^k}[\phi_0^k] + \frac{1}{L}\sum_{\ell=1}^{L}\mathbb{E}_{\mathcal{D}_\ell^k - \mathtt{P}^k}[\hat{\phi}_\ell^k - \phi_0^k] < x\right) \tag{D.52}$$

$$= \mathtt{P}^k\left(A_k + B_k < x\right) \tag{D.53}$$

$$= \mathtt{P}^k\left(A_k < x - B_k\right) \tag{D.54}$$

$$\leq \mathtt{P}^k\left(A_k < x + C_k\right) \tag{D.55}$$

$$\overset{\text{w.p } 1-\epsilon}{\leq} \mathtt{P}^k\left(A_k < x + \Delta_k\right). \tag{D.56}$$

Then,

$$\left|\mathtt{P}^k\left(A_k < x + \Delta_k\right) - \Phi(x)\right| \tag{D.57}$$

$$= \left|\mathtt{P}^k\left(A_k < x + \Delta_k\right) - \Phi(x + \Delta_k) + \Phi(x + \Delta_k) - \Phi(x)\right| \tag{D.58}$$

$$\leq \left|\mathtt{P}^k\left(A_k < x + \Delta_k\right) - \Phi(x + \Delta_k)\right| + \left|\Phi(x + \Delta_k) - \Phi(x)\right| \tag{D.59}$$

$$\leq \frac{0.4748\kappa_0^3}{\rho_{k,0}^3\sqrt{|\mathcal{D}^k|}} + \left|\Phi(x + \Delta_k) - \Phi(x)\right| \qquad \text{(Prop. D.1)} \tag{D.60}$$

$$= \frac{0.4748\kappa_0^3}{\rho_{k,0}^3\sqrt{|\mathcal{D}^k|}} + \left|\Phi'(x')\Delta_k\right| \qquad \text{(Mean-value theorem)} \tag{D.61}$$

$$\leq \frac{0.4748\kappa_0^3}{\rho_{k,0}^3\sqrt{|\mathcal{D}^k|}} + \frac{1}{\sqrt{2\pi}}\Delta_k. \tag{D.62}$$

This completes the proof. ■

D.6 Proof for Corollary 3

By Cauchy-Schwartz' inequality,

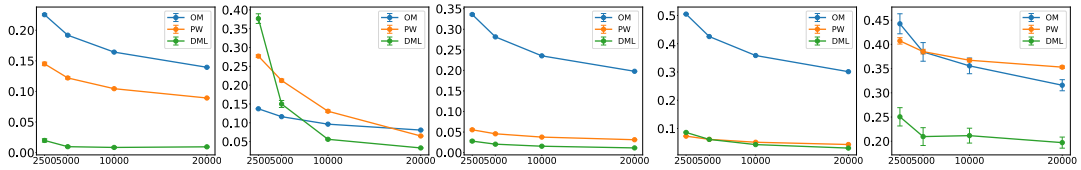
$$\frac{1}{L}\sum_{\ell=1}^{L}\sum_{i=1}^{m}\mathbb{E}_{P^{i+1}}[\hat{\omega}_\ell^{(i-1)}\{\mu_0^i - \hat{\mu}_\ell^i\}\{\hat{\omega}_\ell^i - \omega_0^i\}] \leq \frac{1}{L}\sum_{\ell=1}^{L}\sum_{i=1}^{m}O_{P^{i+1}}\left(\|\mu_0^i - \hat{\mu}_\ell^i\|\|\omega_0^i - \hat{\omega}_\ell^i\|\right). \tag{D.63}$$

Given assumption, the upper bound in Eq. (13) converges at $o_{\mathtt{P}^k}(1/\sqrt{|\mathcal{D}_\ell^k|})$, we conclude that R^k converges in distribution to $\texttt{normal}(0, \rho_{k,0}^2)$.

E More Experiments

In this section, we demonstrate the DML-UCA estimator through examples for the ETT, S-admissibility, FD, Verma's equation, and Ctf-DE described in Sec. 2. For each example, the proposed estimator is constructed using a dataset $\mathtt{D}^k$ following a distribution $\mathtt{P}^k$. Our goal is to provide empirical evidence of the fast convergence behavior of the proposed estimator compared to competing baseline estimators. We consider two standard baselines in the literature: the 'regression-based estimator (reg)' only uses the regression nuisance parameters $\boldsymbol{\mu}$, and the 'ratio-based estimator (ratio)' that only uses the ratio nuisance parameters $\boldsymbol{\pi}$, while our DML-UCA estimator ('dml') uses both. Details of the regression-based ('reg') and the ratio-based ('ratio') estimators are provided in Sec. A. Details of experimental setting is provided in Sec. F. In this experiments, we set all variables other than the treatment variable X as continuous.

We compare DML-UCA estimator to the regression-based estimator ('reg') and the ratio-based estimator ('ratio'). In particular, we use $\hat{\psi}^{\text{est}}$ for est $\in \{\text{reg}, \text{pw}, \text{dml}\}$ to denote the regression-



(a) (b) (c) (d) (e)

Figure E.3: **(a)** ETT in Sec. B, **(b)** Transportability (S-admissibility) in Sec. B, **(c)** Front-door in Example 1, **(d)** Verma in Example 2, **(e)** Ctf-DE in Example 3.

based, probability-weighting, and DML-UCA estimators. We assess the quality of the estimators by computing the *average absolute error* AAE^{est} which is defined as follow. For the ETT and Ctf-DE, $\text{AAE}^{\text{est}} := |\hat{\psi}^{\text{est}} - \psi_0|$, where $\psi_0 := \mathbb{E}[Y_{X=0} \mid X = 1]$ for the ETT and $\psi_0 := \mathbb{E}[Y_{X=0,W_{X=1}} \mid X = 2]$ for the Ctf-DE. For the other examples, $\text{AAE}^{\text{est}} := \frac{1}{\text{domqin}(X)} \sum_{x \in \text{domain}(X)} |\hat{\psi}^{\text{est}}(x) - \psi_0(x)|$ where $\psi_0(x) := \mathbb{E}[Y \mid \text{do}(x)]$, $\hat{\psi}^{\text{est}}(x)$ is an estimator for $\psi_0(x)$ and $\text{dom}(X)$ is a cardinality of the domain of X. Nuisance functions are estimated using XGBoost (Chen and Guestrin, 2016). We ran 100 simulations for each number of samples $n = \{2500, 5000, 10000, 20000\}$ and drew the AAE plot. We evaluate the AAE^{est} in the presence of the 'converging noise ϵ' as in Sec. 4.

Statistical Robustness. The AAE plots for all scenarios are presented in Fig. E.3. For all examples, all the estimators ('reg', 'pw', 'dml') converge as the sample size grows. Furthermore, the proposed DML-UCA estimator outperforms the other two estimators by achieving fast convergence. This result corroborates the robustness property in Thm. 3, which implies that DML-UCA converges faster than the other counterparts.

F Details in Experiments

As described in Sec. 4, we used the XGBoost (Chen and Guestrin, 2016) as a model for estimating nuisances. We implemented the model using Python. In modeling nuisance using the XGBoost, we used the command `xgboost.XGBClassifier(eval_metric='logloss')`[1] to use the XGBoost. We tuned the parameters for each examples to empirically guarantee the convergence of the regression and ratio nuisances. For each examples, the same parameters are used globally for implementing DML-UCA, regression-based estimator, ratio-based estimator, or other competing estimators (Fulcher et al., 2019; Jung et al., 2021a).

Now, we present the structural causal models (SCMs) utilized for generating the dataset. Furthermore, we include a segment of the code employed to generate the dataset.

F.1 FD (Fig. 1a) for Simulation in Fig. 2a

We define the following structural causal models:

$$
\begin{aligned}
U &\sim \texttt{normal}(0.5, 0, 5), \\
U_{Z_i} &\sim \texttt{normal}(0, 1), \text{ for } i = 1, \cdots, d_Z \\
C_i &:= f_{C_i}(U), \text{ where } \mathbf{C} := \{C_i : i = 1, \cdots, d_C\} \\
X &:= f_X(\mathbf{C}, U), \\
Z_i &:= f_{Z_i}(\mathbf{C}, X), \text{ where } \mathbf{Z} := \{Z_i : i = 1, \cdots, d_Z\} \\
Y &:= f_Y(\mathbf{C}, Z, U),
\end{aligned}
$$

[1] Detailed parametrization of parameters including learning rates, maximum depth of the trees, etc. are explained in `https://xgboost.readthedocs.io/en/stable/python/python_api.html#module-xgboost.training`

where

$$
\begin{aligned}
f_{C_i}(U) &\coloneqq \left\lfloor \frac{1}{1+\exp(0.25U_Z + 2U - 1)} \right\rfloor, \\
f_X(\mathbf{C}, U) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp(2\mathbf{C}^\intercal\mathbf{1} - 1 + U)}\right), \\
f_{Z_i}(\mathbf{C}, X) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp(2X - 1 + 0.5\mathbf{C}^\intercal\mathbf{1} + U_{Z_i})}\right) \\
f_Y(\mathbf{C}, Z, U) &\coloneqq \frac{1}{1+\exp((1/d_C)\mathbf{C}^\intercal\mathbf{1} + (1/d_Z)(2\mathbf{Z}^\intercal\mathbf{1} - 1) + 2U)}.
\end{aligned}
$$

The parameterization for XGBoost used in $\boldsymbol{\mu}$ called (`mu_params`) and $\boldsymbol{\pi}$ called (`pi_params`) is the following:

```
    mu_params = {
'booster': 'gbtree',
'eta': 0.3,
'gamma': 0,
'max_depth': 10,
'min_child_weight': 1,
'subsample': 1.0,
'colsample_bytree': 1,
'lambda': 0.0,
'alpha': 0.0,
'objective': 'reg:squarederror',
'eval_metric': 'rmse',
'n_jobs': 4
}

pi_params = {
'booster': 'gbtree',
'eta': 0.3,
'gamma': 0,
'max_depth': 10,
'min_child_weight': 1,
'subsample': 0.0,
'colsample_bytree': 1,
'objective': 'binary:logistic',
'eval_metric': 'logloss',
'reg_lambda': 0.0,
'reg_alpha': 0.0,
'nthread': 4
}
```

F.2 Verma (Fig. 1b) for Simulation in Fig. 2b

We define the following structural causal models:

$$
\begin{aligned}
U_{XB} &\sim \texttt{normal}(1, 0, 5), \\
U_{AY} &\sim \texttt{normal}(-1, 0, 5), \\
U_A &\sim \texttt{normal}(0, 1) \\
U_B &\sim \texttt{normal}(0, 1) \\
X &\coloneqq f_X(U_X B) \\
A_i &\coloneqq f_{A_i}(X, U_{AY}), \text{ for } i = 1, \cdots, d_A \\
B_i &\coloneqq f_{B_i}(X, U_{XB}), \text{ for } i = 1, \cdots, d_B \\
Y &\coloneqq f_Y(\mathbf{B}, U_{AY}),
\end{aligned}
$$

where

$$
\begin{aligned}
f_X(U_X B) &\coloneqq \texttt{Binary}\left(\frac{1}{1 + \exp(2U_{XB} - 1)}\right), \\
f_{A_i}(X, U_{AY}) &\coloneqq \texttt{Binary}\left(\frac{1}{1 + \exp(2X - 1 + U_A + U_{AY})}\right) \\
f_{B_i}(X, U_{XB}) &\coloneqq \texttt{Binary}\left(\frac{1}{1 + \exp(2\mathbf{A}^\intercal\mathbf{1} - 1 + U_B + 0.5U_{XB})}\right) \\
f_Y(\mathbf{B}, U_{AY}) &\coloneqq \frac{1}{1 + \exp(2\mathbf{B}^\intercal\mathbf{1} - 1 + 0.5U_{AY})}.
\end{aligned}
$$

The parameterization for XGBoost used in $\boldsymbol{\mu}$ called (`mu_params`) and $\boldsymbol{\pi}$ called (`pi_params`) is the following:

```
    mu_params = {
'booster': 'gbtree',
'eta': 0.35,
'gamma': 0,
'max_depth': 6,
'min_child_weight': 1,
'subsample': 1.0,
'colsample_bytree': 1,
'lambda': 0.0,
'alpha': 0.0,
'objective': 'reg:squarederror',
'eval_metric': 'rmse',
'n_jobs': 4  # Assuming you have 4 cores
}

pi_params = {
'booster': 'gbtree',
'eta': 0.1,
'gamma': 0,
'max_depth': 10,
'min_child_weight': 1,
'subsample': 0.0,
'colsample_bytree': 1,
'objective': 'binary:logistic',  # Change as per your objective
'eval_metric': 'logloss',  # Change as per your needs
'reg_lambda': 0.0,
'reg_alpha': 0.0,
'nthread': 4
```

}

F.3 Example estimand (Fig. 1e) for Simulation in Fig. 2c

We define the following structural causal models:

$$
\begin{aligned}
U_{X_1,Z} &\sim \texttt{normal}(1, 0, 5),\\
U_{X_1,Y} &\sim \texttt{normal}(-1, 0, 5),\\
U_{Z,Y} &\sim \texttt{normal}(0.5, 0.5)\\
U_R &\sim \texttt{normal}(0, 0.5)\\
U_Z &\sim \texttt{normal}(0, 0.5)\\
U_{X_2} &\sim \texttt{normal}(0, 0.5)\\
X_1 &\coloneqq f_{X_1}(U_{X_1,Z}, U_{X_1,Y})\\
Z_i &\coloneqq f_{Z_i}(X_1, U_{X1,Z}, U_{Z,Y}), \text{ for } i = 1, \cdots, d_Z\\
R_i &\coloneqq f_{R_i}(X_1),, \text{ for } i = 1, \cdots, d_R\\
Y &\coloneqq f_Y(\mathbf{B}, U_{AY}),
\end{aligned}
$$

where

$$
\begin{aligned}
f_{X_1}(U_{X_1,Z}, U_{X_1,Y}) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp(2U_{X_1,Z} - U_{X_1,Y} - 1)}\right),\\
f_{R_i}(X_1) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp(2X_1 - 1 + U_R)}\right)\\
f_{Z_i}(X_1, U_{X1,Z}, U_{Z,Y}) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp(4X_1 - 1 + U_Z + U_{X_1,Z} + U_{Z,Y})}\right)\\
f_{X_2}(\mathbf{Z}, X_1) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp((2X_1 - 1)\mathbf{Z}^\intercal\mathbf{1} - U_{X_2})}\right),\\
f_Y(\mathbf{R}, X_2, U_{X_1,Y}, U_{Z,Y}) &\coloneqq \frac{1}{1+\exp((1/dR)\mathbf{R}^\intercal\mathbf{1} + 2X_2 - 1 + 2(U_{X_1,Y} + U_{Z,Y}))}.
\end{aligned}
$$

The parameterization for XGBoost used in $\boldsymbol{\mu}$ called (`mu_params`) and $\boldsymbol{\pi}$ called (`pi_params`) is the following:

```
    mu_params = {
'booster': 'gbtree',
'eta': 0.3,
'gamma': 0,
'max_depth': 8,
'min_child_weight': 1,
'subsample': 0.8,
'colsample_bytree': 0.8,
'lambda': 0.0,
'alpha': 0.0,
'objective': 'reg:squarederror',
'eval_metric': 'rmse',
'n_jobs': 4  # Assuming you have 4 cores
}

pi_params = {
'booster': 'gbtree',
'eta': 0.1,
'gamma': 0,
'max_depth': 10,
```

```
'min_child_weight': 1,
'subsample': 0.75,
'colsample_bytree': 0.75,
'objective': 'binary:logistic',  # Change as per your objective
'eval_metric': 'logloss',  # Change as per your needs
'reg_lambda': 0.0,
'reg_alpha': 0.0,
'nthread': 4
}
```

F.4 ETT in Sec. B for Simulation in Fig. E.3a

We define the following structural causal models:

$$
\begin{aligned}
U_X &\sim \texttt{normal}(0,1) \\
U_Y &\sim 0.5\, textttnormal(0,1) \\
\mathbf{Z} &\sim 0.25\texttt{normal}(0,1,dZ), \\
X &\coloneqq f_X(\mathbf{Z}) \\
Y &\coloneqq f_Y(X,\mathbf{Z})
\end{aligned}
$$

where

$$
\begin{aligned}
f_X(\mathbf{Z}) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp(2\mathbf{Z}^\intercal\mathbf{1}-1+U_X)}\right) \\
f_Y(\mathbf{Z},X) &\coloneqq \frac{1}{1+\exp(\mathbf{Z}^\intercal\mathbf{1}(2X-1)+U_Y)}.
\end{aligned}
$$

The parameterization for XGBoost used in $\boldsymbol{\mu}$ called (`mu_params`) and $\boldsymbol{\pi}$ called (`pi_params`) is the following:

```
    mu_params = {
'booster': 'gbtree',
'eta': 0.5,
'gamma': 0,
'max_depth': 15,
'min_child_weight': 1,
'subsample': 0.8,
'colsample_bytree': 1,
'lambda': 0,
'alpha': 0,
'objective': 'reg:squarederror',
'eval_metric': 'rmse',
'n_jobs': 4  # Assuming you have 4 cores
}

pi_params = {
'booster': 'gbtree',
'eta': 0.3,
'gamma': 0,
'max_depth': 10,
'min_child_weight': 1,
'subsample': 1,
'colsample_bytree': 1,
'objective': 'binary:logistic',  # Change as per your objective
```

```
'eval_metric': 'logloss',  # Change as per your needs
'reg_lambda': 1,
'reg_alpha': 0,
'nthread': 4
}
```

F.5 Transportability in Sec. B for Simulation in Fig. E.3b

We define the following structural causal models:

$$
\begin{aligned}
U_X &\sim \texttt{normal}(0,1) \\
U_Y &\sim 0.5\, texttttnormal(0,1) \\
\mathbf{Z} &\sim 0.25\texttt{normal}(0,0.5,dZ) + S\texttt{normal}(0.1,0.5,dZ) \\
X &\coloneqq f_X(\mathbf{Z}) \\
Y &\coloneqq f_Y(X,\mathbf{Z})
\end{aligned}
$$

where

$$
\begin{aligned}
f_X(\mathbf{Z}) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp((1/dZ)(2\mathbf{Z}^\intercal\mathbf{1}-1)+U_X)}\right) \\
f_Y(\mathbf{Z},X) &\coloneqq \frac{1}{1+\exp(\mathbf{Z}^\intercal\mathbf{1}(2X-1)+U_Y)}.
\end{aligned}
$$

The parameterization for XGBoost used in $\boldsymbol{\mu}$ called (`mu_params`) and $\boldsymbol{\pi}$ called (`pi_params`) is the following:

```
    mu_params = {
'booster': 'gbtree',
'eta': 0.3,
'gamma': 0,
'max_depth': 15,
'min_child_weight': 1,
'subsample': 0.8,
'colsample_bytree': 1,
'lambda': 0,
'alpha': 0,
'objective': 'reg:squarederror',
'eval_metric': 'rmse',
'n_jobs': 4  # Assuming you have 4 cores
}

pi_params = {
'booster': 'gbtree',
'eta': 0.1,
'gamma': 0,
'max_depth': 10,
'min_child_weight': 1,
'subsample': 1,
'colsample_bytree': 1,
'objective': 'binary:logistic',  # Change as per your objective
'eval_metric': 'logloss',  # Change as per your needs
'reg_lambda': 1,
'reg_alpha': 0,
'nthread': 4
```

}

F.6 FD with continuous mediators for Simulation in Fig. E.3c

We define the following structural causal models:

$$\begin{aligned}
\mathbf{U_C} &\sim \texttt{normal}(0, 1, d_C) \\
U &\sim \texttt{normal}(0, 1) \\
\mathbf{C} &\coloneqq f_{\mathbf{C}}(U) \\
X &\coloneqq f_X(U, \mathbf{C}) \\
Z &\coloneqq f_Z(X, \mathbf{C}) \\
Y &\coloneqq f_Y(U, Z, \mathbf{C})
\end{aligned}$$

where

$$\begin{aligned}
f_{\mathbf{C}}(U) &\coloneqq 0.25\mathbf{U_C} + 2U - 1 \\
f_X(U, \mathbf{C}) &\coloneqq \texttt{Binary}\left(\frac{1}{1 + \exp((2\mathbf{C}^\intercal\mathbf{1} - 1) + U)}\right) \\
f_Z(X, \mathbf{C}) &\coloneqq \frac{1}{1 + \exp(0.1\mathbf{C}^\intercal\mathbf{1}(2X - 1) + X)} \\
f_Y(\mathbf{Z}, X) &\coloneqq \frac{1}{1 + \exp(\mathbf{C}^\intercal\mathbf{1} + (2Z - 1) + U)}.
\end{aligned}$$

The parameterization for XGBoost used in $\boldsymbol{\mu}$ called (`mu_params`) and $\boldsymbol{\pi}$ called (`pi_params`) is the following:

```
    mu_params = {
'booster': 'gbtree',
'eta': 0.01,
'gamma': 0,
'max_depth': 10,
'min_child_weight': 1,
'subsample': 1.0,
'colsample_bytree': 1,
'lambda': 0.0,
'alpha': 0.0,
'objective': 'reg:squarederror',
'eval_metric': 'rmse',
'n_jobs': 4
}

pi_params = {
'booster': 'gbtree',
'eta': 0.3,
'gamma': 0,
'max_depth': 20,
'min_child_weight': 1,
'subsample': 0.0,
'colsample_bytree': 1,
'objective': 'binary:logistic',
'eval_metric': 'logloss',
'reg_lambda': 0.0,
'reg_alpha': 0.0,
'nthread': 4
```

}

F.7 Verma's equation with continuous mediators for Simulation in Fig. E.3d

We define the following structural causal models:

$$
\begin{aligned}
U_{XB} &\sim \texttt{normal}(1, 0, 5), \\
U_{AY} &\sim \texttt{normal}(-1, 0, 5), \\
X &\coloneqq f_X(U_X B) \\
A &\coloneqq f_A(X, U_{AY}) \\
B &\coloneqq f_B(X, U_{XB}) \\
Y &\coloneqq f_Y(B, U_{AY}),
\end{aligned}
$$

where

$$
\begin{aligned}
f_X(U_X B) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp(2U_{XB}-1)}\right), \\
f_A(X, U_{AY}) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp(2X-1+0.5U_A Y)}\right) \\
f_B(X, U_{XB}) &\coloneqq \texttt{Binary}\left(\frac{1}{1+\exp(2A-1+0.5U_X B)}\right) \\
f_Y(B, U_{AY}) &\coloneqq \frac{1}{1+\exp(2B-1+0.5U_A Y)}.
\end{aligned}
$$

The parameterization for XGBoost used in $\boldsymbol{\mu}$ called (`mu_params`) and $\boldsymbol{\pi}$ called (`pi_params`) is the following:

```
mu_params = {
'booster': 'gbtree',
'eta': 0.35,
'gamma': 0,
'max_depth': 6,
'min_child_weight': 1,
'subsample': 1.0,
'colsample_bytree': 1,
'lambda': 0.0,
'alpha': 0.0,
'objective': 'reg:squarederror',
'eval_metric': 'rmse',
'n_jobs': 4  # Assuming you have 4 cores
}

pi_params = {
'booster': 'gbtree',
'eta': 0.1,
'gamma': 0,
'max_depth': 10,
'min_child_weight': 1,
'subsample': 0.0,
'colsample_bytree': 1,
'objective': 'binary:logistic',  # Change as per your objective
'eval_metric': 'logloss',  # Change as per your needs
'reg_lambda': 0.0,
'reg_alpha': 0.0,
'nthread': 4}
```

F.8 Ctf-DE in Example 3 for Simulation in Fig. E.3e

We define the following structural causal models:

$$
\begin{aligned}
U &\sim \texttt{normal}(0, 2),\\
X &\coloneqq f_X(U)\\
Z &\coloneqq f_Z(U)\\
W &\coloneqq f_W(X, Z)\\
Y &\coloneqq f_Y(X, Z, W),
\end{aligned}
$$

where

$$
\begin{aligned}
f_X(U) &\coloneqq \begin{cases} 0 \text{ if } \frac{1}{1+\exp(2U_{XB}-1)} < 0.5 \\ 1 \text{ if } \leq 0.5 \frac{1}{1+\exp(2U_{XB}-1)} < 0.8 \\ 2 \text{ if } \leq 0.8 \frac{1}{1+\exp(2U_{XB}-1)}. \end{cases}\\
f_Z(U) &\coloneqq \frac{1}{1+\exp(-U+1)}\\
f_W(X, Z) &\coloneqq \frac{1}{1+\exp(X-1+Z)}\\
f_Y(Z, X, W) &\coloneqq \frac{1}{1+\exp(3X-1+0.1Z+0.1W+W(X-1))}.
\end{aligned}
$$

The parameterization for XGBoost used in $\boldsymbol{\mu}$ called (`mu_params`) and $\boldsymbol{\pi}$ called (`pi_params`) is the following:

```
mu_params = {
'booster': 'gbtree',
'eta': 0.3, # vab
'gamma': 0.0,
'max_depth': 6, #vb (same as va)
'min_child_weight': 1,
'subsample': 1.0,
'colsample_bytree': 1,
'lambda': 0.0,
'alpha': 0.0,
'objective': 'reg:squarederror',
'eval_metric': 'rmse',
'n_jobs': 4  # Assuming you have 4 cores
}

pi_params = {
'booster': 'gbtree',
'eta': 0.05,
'gamma': 0,
'max_depth': 10,
'min_child_weight': 1,
'subsample': 1.0,
'colsample_bytree': 1,
'objective': 'multi:softprob',  # Change as per your objective
'num_class': 3,
'eval_metric': 'mlogloss',  # Change as per your needs
'reg_lambda': 0.0,
'reg_alpha': 0.0,
'nthread': 4
}
```