Tight Partial Identification of Causal Effects with Marginal Distribution of Unmeasured Confounders

Anonymous Authors[1]

Abstract

Partial identification (PI) presents a significant challenge in causal inference due to the incomplete measurement of confounders. Given that obtaining auxiliary variables of confounders is not always feasible and relies on untestable assumptions, researchers are encouraged to explore the internal information of latent confounders without external assistance. However, these prevailing PI results often lack precise mathematical measurement from observational data or assume that the information pertaining to confounders falls within extreme scenarios. In our paper, we reassess the significance of the marginal confounder distribution in PI. We refrain from imposing additional restrictions on the marginal confounder distribution, such as entropy or mutual information. Instead, we establish the closed-form tight PI for any possible $\mathbb{P}(U)$ in the discrete case. Furthermore, we establish the if and only if criteria for discerning whether the marginal confounder information leads to non-vanilla PI regions. This reveals a fundamental negative result wherein the marginal confounder information minimally contributes to PI as the confounder's cardinality increases. Our theoretical findings are supported by experiments.

1. Introduction

Estimating causal effect is important in a wide range of fields, including medicine (Castro et al., 2020), economics (Hicks et al., 1980), and education (Peng & Knowles, 2003). Due to the existence of latent confounders, the causal effect is usually not identifiable just from observational distribution. For example, when there exists a latent confounder that affects observed random variables X and Y via the causal diagram as described in Figure 1, the existence of two paths $U \to X$ and $U \to Y$ may affect the judgment of the direct causal effect from X to Y. This is also related to the famous "Simpson's paradox" (Pearl, 2014).

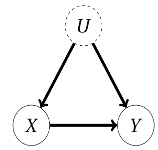

FIGURE 1: *We consider the fundamental causal graph involving treatment X, outcome Y, and confounders U. Our focus lies in achieving the tight PI of causal queries using only the information from the marginal distribution $P(U)$ in conjunction with the observed $P(X, Y)$.*

In the absence of unobserved confounders, it is well known that causal effect is "identifiable" (Robins, 1987). Taking Figure 1 again for instance, when the joint distribution of random variables $\{X, Y, U\}$ can be observed, the causal effect from X to Y could be fully recovered according to the famous "back-door criteria" (Pearl et al., 2000). When unmeasured confounders exist, Tian & Pearl (2002) established the if and only if criteria for the identification of causal queries. When it is not satisfied, one can at most identify a region where the true causal effect belongs, which is commonly known as the *Partial identification* (PI).

When only the marginal distribution $\mathbb{P}(X, Y)$ is accessible, and the causal diagram follows Figure 1, the tight PI region of causal estimand is provided by Tian & Pearl (2000), which is also known as the so-called "vanilla bound". To achieve an identification region tighter than just vanilla bound, existing methods can be split into two categories. The first category resorts to external auxiliary variables. For example, Balke & Pearl (1997) proposed the famous "Balke-Pearl" bound via auxiliary instrument variables, which is further extended by Kitagawa (2009) to the continuous case. Ghassami et al. (2023) generalized the traditional double-negative control method, which took advantage of the treatment and outcome confounding proxy variables to construct valid PI bounds. Besides, Gabriel et al. (2022) selected the outcome-dependent samples for assistance.

Since the auxiliary variable is not always obtainable, the

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

second category instead focuses on only using additional side information of latent confounders to improve identification. The most fundamental strategy is to directly exploit the marginal probability distribution of latent confounders $\mathbb{P}(U)$ (Li & Pearl, 2022; Li et al., 2023; Jiang et al., 2023). Unfortunately, these methods can only handle specific extreme cases of confounders (i.e., cases where the information entropy of $\mathbb{P}(U)$ is rather small) and do not have any tightness guarantee. Other series of work shifts to characterize the association between confounder U and $\{X, Y\}$ relying on additional hyper-parameters and customized measures, e.g. sensitivity analysis (Dorn et al., 2021), information-theoretic method (Janzing et al., 2013; Geiger et al., 2014).

Hence, we consider the problem of finding the *tight* PI region of causal effects with the assistance of general marginal distribution information of latent confounders. Moreover, we do not impose any restriction on $\mathbb{P}(U)$. Throughout this paper, we focus on the simplest structure in Figure 1, which is the same as the settings considered by the closely related literature (Li & Pearl, 2022; Li et al., 2023; Jiang et al., 2023). We explore the closed-form solution of the tight PI region using $\mathbb{P}(X, Y)$ and $\mathbb{P}(U)$ and discuss further the intuitions from it. On this basis, we derive theoretical negative results that enhance the scientific paradigm for PI through the constrained optimization approach. In sum, our contributions are as follows:

- We develop the *tight* PI region of casual queries with the marginal distribution of unmeasured confounders without additional restrictions.
- We establish the *if and only if* criteria for $\mathbb{P}(U)$ so that the tight PI is stricter than the vanilla bound[1]. We fundamentally indicate that as the confounder's cardinality increases, it is less likely that the additional information of $\mathbb{P}(U)$ can provide an identification region tighter than just vanilla bound.
- We conduct synthetic and real-world experiments to quantify the information loss of the traditional entropy-based optimization for PI in various settings, compared with our proposed PI region.

The rest of the paper is organized as follows. In Section 2, we review the basic framework and corresponding literature on PI. In Section 3, we establish the closed-form tight identification region for different causal quantities with binary confounders. In Section 4, we extend our result to the categorical case and establish the above-mentioned *if and only if* condition. We generalize such conditions to new measurements to quantify identification improvement by using $\mathbb{P}(U)$. In Section 5, we illustrate our simulation and real-world experiment results. We conclude this paper with a discussion in Section 6.

[1] Here "vanilla bound" refers to the identification region when the confounder information is completely unknown (Tian & Pearl, 2000).

2. Framework, notation and literature review

Framework and notations In this paper, we assume that there exists a latent variable U such that the causal relation between X, Y, U follows the causal diagram in Figure 1. We assume that both X, Y are binary variables taking values in $\{0, 1\}$; and that U is a discrete random variables taking values from $\{0, \ldots, d_u - 1\}$ for some positive integer $d_u \geq 2$. To describe causality, we adopt the do-calculus framework (Pearl, 1995), i.e., that $\mathbb{P}(Y = 1 \mid do(X = 0))$ denotes the probability that Y is equal to 1 had we assigned X to be equal to 0. For simplicity we write $\mathbb{P}(y \mid do(x))$ as $\mathbb{P}(Y = y \mid do(X = x))$ and write $\mathbb{P}(x, y), \mathbb{P}(u)$ as $\mathbb{P}(Y = y, X = x)$ and $\mathbb{P}(U = u)$, respectively. Our goal in this paper is to calculate the identification region of the $\mathbb{P}(y \mid do(x))$ and average treatment effect $\mathbb{E}(Y \mid do(X = 1)) - \mathbb{E}(Y \mid do(X = 0)) \equiv \mathbb{P}(Y = 1 \mid do(X = 1)) - \mathbb{P}(Y = 1 \mid do(X = 0))$ with the assistance of background information of the marginal distribution of latent confounders, i.e., $\mathbb{P}(U)$.

When the information $\mathbb{P}(U)$ is not accessible, Tian & Pearl (2000) shows that

$$\mathbb{P}(y \mid do(x)) = \mathbb{P}(x, y) + \sum_u \frac{\mathbb{P}(y, u, x)}{\mathbb{P}(u, x)} \mathbb{P}(u, \neg x) \tag{1}$$

belongs to $[\mathbb{P}(x, y), \mathbb{P}(x, y) + \mathbb{P}(\neg x)]$, where since $x \in \{0, 1\}$, we write $\neg x \equiv 1 - x$, i.e., that $\mathbb{P}(\neg x) \equiv \mathbb{P}(X = 1 - x)$. We denote $\mathbb{P}(x, y)$ and $\mathbb{P}(x, y) + \mathbb{P}(\neg x)$ as the "vanilla lower bound" and "vanilla upper bound" of $\mathbb{P}(y \mid do(x))$.

Stepping forward, we follow Robins (1989) to denote $-\mathbb{P}(X = 1, Y = 0) - \mathbb{P}(X = 0, Y = 1)$ and $\mathbb{P}(X = 1, Y = 1) + \mathbb{P}(X = 0, Y = 0)$ as the 'vanilla lower bound of ATE' and 'vanilla upper bound of ATE'.

Literature review In observational studies, partial identification (PI) indeed originates from point-wise identification, for which additional auxiliary variables and assumptions are required. Wright (1928) proposed instrument variable (IV) to estimate causal effect via regression with linear model assumption. Kuroki & Pearl (2014) established sufficient conditions under which the proxy variables could help point-wisely restore the causal effect. It was subsequently developed into double negative control (Nagasawa, 2018; Shi et al., 2020; Singh, 2020; Cui et al., 2023; Tchetgen et al., 2020; Deaner, 2018; Kallus et al., 2021; Miao et al., 2018; Qi et al., 2023), and currently further simplified to be single proxy control (Tchetgen et al., 2023; Park & Tchetgen, 2023; Xu & Gretton, 2023). Informally speaking, these two methodologies both require the confounder proxies to be informative enough, namely, the transition matrix from the confounders to proxies is left-reversible in the discrete case.

To avoid being constrained to particular contexts as above, researchers are encouraged to weaken these assumptions to further explore PI (Manski, 1990; Tamer, 2010; Kline & Tamer, 2023). Geiger & Meek (2013) theoretically illustrated the feasibility of transforming PI into an optimisation problem. Following our introduction, two categories are divided. Correspondingly, the first auxiliary-based category inherits and generalizes the above point-wise identification as IV-based PI (Balke & Pearl, 1997; Swanson et al., 2018; Kitagawa, 2009; Zhang & Bareinboim, 2021a), negative control-based PI (Ghassami et al., 2023), outcome-dependent sampling PI (Gabriel et al., 2022) .

The second category is the most relevant to ours and, therefore, warrants further in-depth discussion. Removing untenable auxiliary variables and untestable assumptions brings out a greater challenge for PI optimization. Pioneering works started with rough qualitative analyses. Geiger et al. (2014) proved that $\mathbb{P}(y \mid do(x))$ in Eqn 1 is bounded by so-called "back-door dependence", which is measured by mutual information between U and X. Such information-theoretic concepts could help bound various causal quantities (Janzing et al., 2013) whereas are practically constrained by external hyper-parameters, e.g., sensitivity analysis (Kallus et al., 2019; Marmarelis et al., 2023; Dorn et al., 2021; Christopher Frey & Patil, 2002), or parametric machine learning models (Hu et al., 2021; Balazadeh Meresht et al., 2022). For simplicity and generalization, people currently revisit Figure 1 and directly utilize the marginal confounder information (Schuster et al., 2015; Dawid et al., 2017; Mueller et al., 2021). Taking advantage of Single world intervention graphs (Richardson & Robins, 2013), Jiang et al. (2023) surrogated $\mathbb{P}(U)$ information into entropy [2] $H(U)$ and provided a state-of-the-art entropy-based valid PI region of Eqn 1:

$$\left\{ \sum_{x'=0,1} b_{yx'} P\left(x'\right) : \mathbf{b} \in \mathcal{B} \cap \mathcal{B}_U \right\} \text{ (Jiang et al., 2023).}$$

Here $\mathbf{b} := \{b_{ij}\}_{i,j\in\{0,1\}}$, and $\mathcal{B}, \mathcal{B}_U$ represent the linear constraint and the entropy constraint, respectively.[3] In view of the non-convex feasible set, Li et al. (2023) further simplified it to a closed-form valid PI bound in the binary case, which degenerates linearly with sufficiently small $H(U)$:

$$\Big[\mathbb{P}(y \mid x) - c_l H(U), \mathbb{P}(y \mid x) + c_u H(U)\Big] \text{ (Li et al., 2023).}$$

[2]Entropy $H(U) := -\sum_{u\in U} \mathbb{P}(u) log(\mathbb{P}(u))$.

[3]$\mathcal{B} := \{\mathbf{b} : \forall i, j, b_{ij} \in [0, 1], \sum_{y',x'} b_{y'x'} P\left(x'\right) = 1; \forall y', b_{y'x} = P\left(y' \mid x\right)\}$. Moreover, they set $\mathcal{B}_U := \{\mathbf{b} : \sum_{y',x'} b_{y'x'} P\left(x'\right) \log \left(b_{y'x'} / \sum_{x'} b_{y'x'} P\left(x'\right)\right) \leq H(U)\}$.

c_l, c_u are positive constants. A comprehensive analysis of its optimal form will be discussed in Section 3. Both of these results argued that confounders with sufficiently small entropy could help construct non-vanilla PI but not guarantee tightness.

By this motivation, a research gap arises: what is the general tight PI region, and when would it be non-vanilla conditioning on any possible $\mathbb{P}(U)$, instead of other surrogates like entropy?[4] Although it could be approximated via advanced optimisation programming (Duarte et al., 2023), the testing on each specific $\mathbb{P}(U)$ is completely empirical. Even worse, the computational complexity grows exponentially with the cardinality of U due to the exhaustive branch-and-bound searching (Duarte et al., 2023). We address this research gap by exploring the closed-form tight PI and its mathematical insight without any additional imposed restrictions.

3. Tight partial identification with binary confounder

For simplicity of illustration, in this section, we first consider the tight PI region for $d_u = 2$. In Theorem 3.3, we showcase the tight bound of $\mathbb{P}(y \mid do(x))$ with prior knowledge of $\mathbb{P}(U)$, and then Theorem 3.5 generalizes Theorem 3.3 to the ATE case. Furthermore, Corollary 3.4 provides a further investigation of Theorem 3.3 when $\mathbb{P}(U = 1)$ or $\mathbb{P}(U = 0)$ is close to zero, i.e., $H(U)$ is small.

Throughout this article, we invoke the following assumption on the marginal distribution $\mathbb{P}(X, Y)$ and $\mathbb{P}(U)$.

Assumption 3.1 (Positivity). $\forall x, y \in \{0, 1\}, \mathbb{P}(x, y) > 0.$

Assumption 3.2. U is a discrete random variable taking values in $\{0, \ldots, d_u - 1\}$. Moreover, there does not exist a u' such that $\mathbb{P}(U = u') = 1$.

The reason why we impose the additional constraint $\mathbb{P}(U = u') \neq 1$ is that once it is violated, U will become deterministic so that there is no latent confounding anymore, and the causal conclusion becomes trivial.

3.1. Identification of interventional probability

In this section, we discuss the tight PI region of interventional probability $\mathbb{P}(y \mid do(x))$ when U is a binary random variable. Our goal is to derive a closed-form solution for the tight identification region. Taking the lower bound for example, it can be obtained by seeking a joint distribution $\mathbb{P}(X, Y, U)$ that minimizes (1) while still compatible with the marginal probabilities $\mathbb{P}(X, Y)$ and $\mathbb{P}(U)$. In other words, we can obtain the lower bound by solving the follow-

[4]Jiang et al. (2023) first established a sufficiency criteria upon what is the greatest entropy $H(U)$ (so-called "entropy threshold") to cause non-vanilla PI. Stepping forward, we formalize an if and only if criteria upon each possible $\mathbb{P}(U)$ in Section 4.

ing optimization program:

$$\min \frac{\mathbb{P}(x,y,U=0)}{\mathbb{P}(x,U=0)}\mathbb{P}(U=0)+\frac{\mathbb{P}(x,y,U=1)}{\mathbb{P}(x,U=1)}\mathbb{P}(U=1),$$
such that $\mathbb{P}(X,Y,U)$ is compatible with the observed marginal distributions $\mathbb{P}(U),\mathbb{P}(X,Y)$. (2)

Apparently, this is a non-trivial non-convex fractional optimization problem where one has to minimize the objective function by varying the denominators $\mathbb{P}(x,U=0)$ and $\mathbb{P}(x,U=1)$. Nevertheless, when we assume that beyond $\mathbb{P}(U)$, $\mathbb{P}(X,U)$ is also known a priori, then we just need to optimize the numerators, which breaks down to a linear optimization problem. Consequently, we can derive that with $\mathbb{P}(X,U)$ known, (2) is equivalent to $\min_{t\in\{0,1\}}\{\max(\mathcal{S}_t)\}$ where

$$\mathcal{S}_t := \left\{\frac{\mathbb{P}(x,y)}{\mathbb{P}(x,U=t)}\mathbb{P}(U=t),\right. \left.\frac{\mathbb{P}(x,y)-\mathbb{P}(x,U=t)}{\mathbb{P}(x)-\mathbb{P}(x,U=t)}(1-\mathbb{P}(U=t))+\mathbb{P}(U=t)\right\}. \tag{3}$$

Due to the min-max operation, (3) is not straightforward to be optimized directly. Fortunately, we have found that this function is piece-wise monotone. This allows us to derive a closed-form identification region by exhaustively examining the boundary points of each piece. Our closed-form identification strategy is presented in Theorem 3.3, which is also a tight identification strategy (Appendix B).

Theorem 3.3 (Identification of interventional probability)**.** *Suppose we are under Assumptions 3.1 and 3.2 with $d_u=2$ and the distribution $\mathbb{P}(U)$ observable. The tight identification region of the interventional probability $\mathbb{P}(y\mid do(x))$ is given by*

$$\left[\min_{t\in\{0,1\}}\mathcal{LB}\left(\mathbb{P}(U=t)\right),\max_{t\in\{0,1\}}\mathcal{UB}\left(\mathbb{P}(U=t)\right)\right].$$

Here $\mathcal{LB}(\cdot),\mathcal{UB}(\cdot)$ are two piece-wise linear functions defined as

$$\begin{cases}\frac{\mathbb{P}(x,y)-t}{\mathbb{P}(x)-t}(1-t)+t & t\in(0,\mathbb{P}(x,y)]\\ \mathbb{P}(x,y) & t\in(\mathbb{P}(x,y),\mathbb{P}(x)]\\ \mathbb{P}(y\mid x)t & t\in(\mathbb{P}(x),1)\end{cases} \quad \textit{and}$$
$$\begin{cases}\mathbb{P}(y\mid x)(1-t)+t & t\in(0,\mathbb{P}(\neg x)]\\ \mathbb{P}(x,y)+\mathbb{P}(\neg x) & t\in(\mathbb{P}(\neg x),1-\mathbb{P}(x,\neg y)]\\ \frac{\mathbb{P}(x,y)t}{\mathbb{P}(x)-(1-t)} & t\in(1-\mathbb{P}(x,\neg y),1)\end{cases}, \tag{4}$$

respectively.

Informally, then, Theorem 3.3 means that when

$$\{\mathbb{P}(U=0),\mathbb{P}(U=1)\}\cap[\mathbb{P}(x,y),\mathbb{P}(x)]\neq\emptyset \text{ and}$$
$$\{\mathbb{P}(U=0),\mathbb{P}(U=1)\}\cap[\mathbb{P}(\neg x),1-\mathbb{P}(x,\neg y)]\neq\emptyset, \tag{5}$$

the tight identification region of the interventional probability is no different from the "vanilla bound". When both $\mathbb{P}(U=0)$ and $\mathbb{P}(U=1)$ are outside of this region, one can expect a more nontrivial identification bound. Figure 2 provides a visualization of the lower bound in Theorem 3.3. Apparently, as $\mathbb{P}(U=1)$ varies from 0 to 1, the lower bound will change in different trends, depending on the marginal distribution of the observed variables (X,Y).

In the scenario where a practitioner does not know the exact value of $\mathbb{P}(U)$, but just that it belongs to a class of distribution $\mathcal{P}$, it's straightforward that the tight identification region of the interventional probability $\mathbb{P}(y\mid do(x))$ is given by

$$\bigcup_{\mathbb{P}(U)\in\mathcal{P}}\left[\min_{t\in\{0,1\}}\mathcal{LB}\left(\mathbb{P}(U=t)\right),\max_{t\in\{0,1\}}\mathcal{UB}\left(\mathbb{P}(U=t)\right)\right].$$

We now consider a special scenario where $\mathcal{P}(\varepsilon):=\{\mathbb{P}(U):\mathbb{P}(U=0)\le\varepsilon \text{ or } \mathbb{P}(U=1)\le\varepsilon\}(\varepsilon\le1/2)$, or equivalently, $H(U)\le-\varepsilon log(\varepsilon)-(1-\varepsilon)log(1-\varepsilon)$, which has been considered before by Li et al. (2023); Jiang et al. (2023). We strengthen the previous result via extreme analysis to a tight PI region:

Corollary 3.4. *Suppose we are under Assumption 3.1. Suppose $d_u=2$ and we are given a $\mathcal{P}(\varepsilon)$ with $\varepsilon\le\min\{\mathbb{P}(x),\mathbb{P}(\neg x)\}$, then the tight identification region of $\mathbb{P}(y\mid do(x))$ is given by*

$$\Big[\min\{\mathcal{LB}(\varepsilon),\mathcal{LB}(1-\varepsilon)\},\max\{\mathcal{UB}(\varepsilon),\mathcal{UB}(1-\varepsilon)\}\Big]. \tag{6}$$

If additionally $\varepsilon\le\min\{\mathbb{P}(x,y),\mathbb{P}(x,\neg y),\mathbb{P}(\neg x)\}$, it can be simplified as

$$[\mathbb{P}(y\mid x)-\varepsilon\mathscr{E}_y(\varepsilon),\mathbb{P}(y\mid x)+\varepsilon\mathscr{E}_{\neg y}(\varepsilon)],\ \textit{where}$$
$$\mathscr{E}_{y'}(\varepsilon)=\max\left\{\frac{\mathbb{P}(x,\neg y')\mathbb{P}(\neg x)}{(\mathbb{P}(x)-\varepsilon)\mathbb{P}(x)},\mathbb{P}(y'\mid x)\right\},y'\in\{0,1\}. \tag{7}$$

In Li et al. (2023), the authors shows that when $\varepsilon\in[0,\mathbb{P}(x))$, a valid identification region is $[\mathcal{LB}_{li},\mathcal{UB}_{li}]$, where

$$\mathcal{LB}_{li}:=\mathbb{P}(y\mid x)-\frac{\mathbb{P}(x)+1}{\mathbb{P}(x)}\varepsilon,$$
$$\mathcal{UB}_{li}=\mathbb{P}(y\mid x)+\frac{\mathbb{P}(x)+1}{\mathbb{P}(x)}\varepsilon+\frac{\varepsilon^2}{\mathbb{P}(x)[\mathbb{P}(x)-\varepsilon]}. \tag{8}$$

When $\mathbb{P}(x)\le\mathbb{P}(\neg x)$, the range of ε considered by Li et al. (2023) is exactly the same as the one considered in the corollary above, and our results show that Li's identification region is strictly looser than the tight identification region for any $\varepsilon\in(0,\mathbb{P}(x))$. When $\mathbb{P}(x)>\mathbb{P}(\neg x)$, compared to our work, Li et al. (2023) additionally considered the region $\varepsilon\in(\mathbb{P}(\neg x),\mathbb{P}(x))$. However, under this regime, one can immediately have $[\mathbb{P}(x,y),\mathbb{P}(x,y)+\mathbb{P}(\neg x)]\subseteq[\mathcal{LB}_{li},\mathcal{UB}_{li}]$, i.e., Li's bound is looser than the vanilla bound without any information about $\mathbb{P}(U)$. Proof and more quantitative analysis are shown in Appendix C.

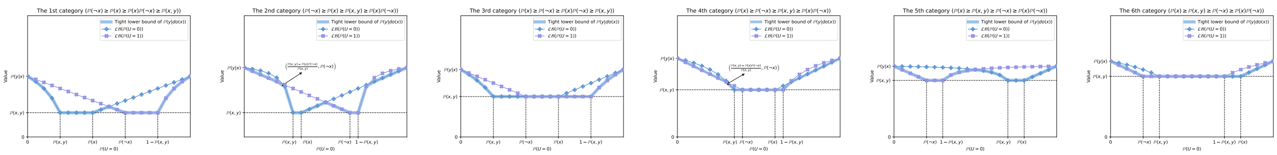

FIGURE 2: *The visualization of Theorem 3.3. We take the lower bound for instance and the upper bound is in the same vein. It could be categorized into six types in total according to the order of* $\{\mathbb{P}(\neg x), \mathbb{P}(x), \mathbb{P}(x)\mathbb{P}(\neg x), \mathbb{P}(x,y)\}$. *As illustrated, the tight lower bound is vanilla if and only if* $\{\mathbb{P}(U=0), \mathbb{P}(U=1)\} \cap [\mathbb{P}(x,y), \mathbb{P}(x)] \neq \emptyset$.

3.2. Identification of average treatment effect

In this section, we further provide the tight identification region of the average treatment effect (ATE) given prior information of $\mathbb{P}(U)$. We also discuss the constraints on the observed marginal distributions so that the resulting identification bound does not degenerate into vanilla.

Theorem 3.5 (Identification of average treatment effect)**.** *Consider the same setup as Theorem 3.3, then the tight identification region of* ATE *is given by*

$$\left[\min_{t\in\{0,1\}}\{-\mathcal{B}(\mathbb{P}(U=t);0,1)\}, \max_{t\in\{0,1\}}\mathcal{B}(\mathbb{P}(U=t);1,1)\right],$$

where $\mathcal{B}(t;x,y) :=$

$$\begin{cases} \Big(-\mathbb{P}(y\mid\neg x)+\frac{\mathbb{P}(x,y)}{\mathbb{P}(x)-t}\Big)(1-t) & t\in(0,p_0] \\ -\mathbb{P}(y\mid\neg x)(1-t)-\mathbb{P}(x,\neg y)+1 & t\in(p_0,p_1] \\ -\mathbb{P}(\neg x,y)+\mathbb{P}(y\mid x)t+(1-t) & t\in(p_1,p_2] \\ \Big(-\frac{\mathbb{P}(\neg x,y)-(1-t)}{\mathbb{P}(\neg x)-(1-t)}+\mathbb{P}(y\mid x)\Big)t & t\in(p_2,1) \end{cases}.$$

Here $p_0=\mathbb{P}(x,\neg y)$, $p_1=\mathbb{P}(x)$, $p_2=1-\mathbb{P}(\neg x,y)$.

The proof is deferred to Appendix D. When

$$\{\mathbb{P}(U=0),\mathbb{P}(U=1)\}\cap\{\mathbb{P}(X=1)\}\neq\emptyset, \quad (9)$$

the tight bound will degenerate into the vanilla bound $[-\mathbb{P}(X=0,Y=1)+\mathbb{P}(X=1,Y=0), \mathbb{P}(X=0,Y=0)+\mathbb{P}(X=1,Y=1)]$, i.e., the tight identification region of ATE provided no prior knowledge about $\mathbb{P}(U)$; otherwise, it is *always* tighter, which is quite inconsistent with the degeneration requirement for interventional probability in (5)[5]. It should be noticed that Theorem 3.5 is not a simple composition of the results of Theorem 3.3; in other words, the lower (upper) tight bound of ATE cannot simply be equated to the difference between the lower (upper) tight bound of $\mathbb{P}(Y=1\mid do(x')), x'=0,1$. In fact, the tightness of the two may not be simultaneously reached.

So far, we have provided the tight identification bound for the interventional probability and average treatment effect when $d_u=2$; we also provide the if and only if conditions so that the tight identification bound does not degenerate into a vanilla bound. This raises an interesting question: is it possible to extend these results into the multivariate case with $d_u\geq 3$? We answer this question in the next section.

[5]Counter-intuitively, (9) and (5) will exhibit consistency under multi-value settings, which will be illustrated in Section 4.

4. Tight partial identification with multi-valued confounder

In this section, we consider the identification of casual queries with multi-valued confounders, namely, $d_u\geq 3$.

4.1. The if and only if condition of degeneration to the vanilla identification region

In Section 3, we have shown that under the setting $d_u=2$, when the prior distribution of U lies in the region characterized by (5), the tight identification region given the prior information in fact has no improvement compared to the vanilla bound. Moreover, such characterization in (5) is "if and only if", in the sense that when (5) is violated, then the tight identification region is definitely tighter than the vanilla bound. In Theorem 4.1 (Appendix E) we further extend such "if and only if" characterization to the multi-valued U.

Theorem 4.1. *Suppose Assumptions 3.1-3.2 hold. The tight lower bound of the interventional probability* $\mathbb{P}(y\mid do(x))$ *given prior knowledge of* $\mathbb{P}(U)$ *is equal to the vanilla lower bound if and only if* $\mathbb{P}(U)$ *belongs to* $\mathcal{P}^L :=$

$$\big\{\mathbb{P}(U) : \exists\, \mathcal{U}\subseteq\mathbb{R} \text{ s.t. } \mathbb{P}(U\in\mathcal{U})\in[\mathbb{P}(x,y),\mathbb{P}(x)]\big\}.$$

Analogously, the if and only if condition for the degeneration of the upper bound is when $\mathbb{P}(U)$ *belongs to* $\mathcal{P}^U :=$

$$\big\{\mathbb{P}(U) : \exists\, \mathcal{U}\subseteq\mathbb{R} \text{ s.t. } \mathbb{P}(U\in\mathcal{U})\in[\mathbb{P}(\neg x),1-\mathbb{P}(x,\neg y)]\big\}.$$

Thus the tight identification region of $\mathbb{P}(y\mid do(x))$ *given prior knowledge of* $\mathbb{P}(U)$ *is equal to the vanilla bound if and only if* $\mathbb{P}(U)\in\mathcal{P}:=\mathcal{P}^L\cap\mathcal{P}^U$.

Compared with Jiang et al. (2023), Theorem 4.1 constructed an *if and only if* criterion upon the non-vanilla region with each possible $\mathbb{P}(U)$, instead of seeking other conservative sufficiency conditions based on the information entropy constraint of U. In other words, the oracle non-vanilla region derived from Theorem 4.1 reveals the ground truth and surrogates the previous result as the special case. More importantly, to demonstrate its simplicity and essence, as a corroboration, we further show that ATE also possesses a consistent form of the decision criterion:

Theorem 4.2. *Suppose Assumptions 3.1-3.2 hold. The if and only if conditions for the tight upper and lower bounds of the average treatment effect to degenerate into vanilla bounds are when $\mathbb{P}(U)$ belongs to*

$$\mathcal{P}^L_{\text{ATE}} := \{\mathbb{P}(U) : \exists\, \mathcal{U}_0, \mathcal{U}_1 \subseteq \mathbb{R} \text{ with } \mathcal{U}_0 \cap \mathcal{U}_1 = \emptyset, \text{ s.t. } \forall z \in \{0,1\}, \mathbb{P}(U \in \mathcal{U}_z) \in \mathcal{I}_{z,z}\},$$

where $\mathcal{I}_{x',y'} := [\mathbb{P}(X = x', Y = y'), \mathbb{P}(X = x')]$ *for* $x', y' \in \{0,1\}$ *and*

$$\mathcal{P}^U_{\text{ATE}} := \{\mathbb{P}(U) : \exists\, \mathcal{U}_0, \mathcal{U}_1 \subseteq \mathbb{R} \text{ with } \mathcal{U}_0 \cap \mathcal{U}_1 = \emptyset, \text{ s.t. } \forall z \in \{0,1\}, \mathbb{P}(U \in \mathcal{U}_z) \in \mathcal{I}_{\neg z,z}\},$$

respectively. Thus, the identification region of the average treatment effect is vanilla if and only if $\mathbb{P}(U) \in \mathcal{P}_{\text{ATE}} := \mathcal{P}^L_{\text{ATE}} \cap \mathcal{P}^U_{\text{ATE}}$.

Theorems 4.1 and 4.2 (Appendix E and Appendix F) provide the if and only if conditions for the tight identification regions of interventional probabilities and average treatment effects to degenerate into vanilla bounds. It shows that whether the tight identification region is vanilla depends on the existence of a subspace of confounder $\mathcal{U} \subseteq \mathbb{R}$ (or two disjoint subspaces $\mathcal{U}_0, \mathcal{U}_1 \subseteq \mathbb{R}$) whose probability measure $\mathbb{P}(U \in \mathcal{U})$ (or $\mathbb{P}(U \in \mathcal{U}_0)$, $\mathbb{P}(U \in \mathcal{U}_1)$) locates in the desired interval. Moreover, such interval is constructed by the observational probability $\mathbb{P}(X, Y)$. When $d_u = 2$, one can easily prove that the conditions displayed in Theorems 4.1 and 4.2 break down to the intervals in (5) and (9).

According to Theorems 4.1 and 4.2, the identification regions $\{\mathcal{P}^L, \mathcal{P}^U, \mathcal{P}, \mathcal{P}_{\text{ATE}}, \mathcal{P}^L_{\text{ATE}}, \mathcal{P}^U_{\text{ATE}}\}$ have properties in the following corollary (Appendix G):

Corollary 4.3. *Suppose Assumptions 3.1-3.2 hold and $d_u \geq 3$. We have*

(i) $\mathcal{P} = \mathcal{P}^L \cap \mathcal{P}^U \neq \emptyset$ *and* $\mathcal{P}_{\text{ATE}} = \mathcal{P}^L_{\text{ATE}} \cap \mathcal{P}^U_{\text{ATE}} \neq \emptyset$;

(ii) $\mathcal{P}^L_{\text{ATE}} \subsetneqq \mathcal{P}^L, \mathcal{P}^U_{\text{ATE}} \subsetneqq \mathcal{P}^U$, *and therefore* $\mathcal{P}_{\text{ATE}} \subsetneqq \mathcal{P}$.

We provide instances in Appendix N. Informally then, it means that i) for any choice of observational distribution $\mathbb{P}(X, Y)$, there always exists some specifications of $\mathbb{P}(U)$ so that its induced identification regions of interventional probability and average treatment effect are no different from vanilla bound, and ii) the identification of average treatment effect is strictly less likely to degenerate into the vanilla case than the interventional probability. To better understand the relationships among these sets, in Figure 3 we provide some visualizations of their relationships via Venn diagrams.

Corollary 4.3 naturally leads to following the question: how large the "volume" of $\mathcal{P}$ and $\mathcal{P}_{\text{ATE}}$ is relative to the entire probability space $\Omega := \{\mathbb{P}(U) : \sum_t \mathbb{P}(U = t) = 1, \mathbb{P}(U = t) \geq 0\}$? To understand this we now consider a Bayesian flavoured setup where the d_u-dimensional parameters $(\mathbb{P}(U = 0), \mathbb{P}(U = 1), \ldots \mathbb{P}(U = d_u - 1))$ is sampled uniformly at random from the $d_u - 1$ probability simplex, and the problem then is transformed to analyzing the probability that the induced $\mathbb{P}(U)$ falls into the "non-vanilla" region $(\mathcal{P})^c$ and $(\mathcal{P}_{\text{ATE}})^c$. First, we have the following theoretical result (Appendix H):

Proposition 4.4. *Assuming that the parameters* $(\mathbb{P}(U = 0), \mathbb{P}(U = 1), \ldots, \mathbb{P}(U = d_u - 1))$ *is sampled uniformly at random from the $(d_u - 1)$-simplex, the probability that $\mathbb{P}(U)$ falls into $(\mathcal{P})^c, (\mathcal{P}_{\text{ATE}})^c$ are all monotonically non-increasing with the increasing d_u; and are at most* $d_u \left(1 - \min_{y' \in \{0,1\}} \mathbb{P}(x, y')\right)^{d_u - 1}$ *and* $d_u \left(1 - \min_{x', y' \in \{0,1\}} \mathbb{P}(x', y')\right)^{d_u - 1}$, *respectively.*

Informally then, it means that the probability that $\mathbb{P}(U)$ falls into the region with a non-trivial bound decreases exponentially with the increment of d_u. Since it just represents some upper bound which is not necessarily tight, we further visualize in Figure 4 how the probability $\mathbb{P}(U) \in \mathcal{P}$ and $\mathbb{P}(U) \in \mathcal{P}_{\text{ATE}}$ vary the increment of d_u under different choices of $\mathbb{P}(X, Y)$. Consistent with the upper bound described in Proposition 4.4, the probability of having a non-trivial $\mathbb{P}(U)$ will tend to zero as d_u goes to infinity; moreover, the probability of having a non-trivial $\mathbb{P}(U)$ are all consistently below 0.1 for d_u no smaller than 10, regardless of the choice of $\mathbb{P}(X, Y)$. This indicates that the marginal information $\mathbb{P}(U)$ is usually more useful with a relatively small d_u. When d_u is, say greater than 10, a practitioner usually cannot expect an informative distribution of the latent confounder with non-vanilla PI regions.

4.2. Closed-form identification region

As indicated in Proposition 4.4 and Figure 4, the provided marginal information $\mathbb{P}(U)$ becomes more likely to lie in the trivial region with increasing d_u. Hence, at a global level, the confounder marginal information shows limited assistance to PI, especially when d_u is relatively large. In this section, we discuss the PI of the causal estimands when $\mathbb{P}(U)$ does not belong to $\mathcal{P}$. First, in Theorem 4.5 (Appendix I) we provide a *closed-form* formulation of the *tight identification* bound of the interventional probability for any $d_u \geq 2$.

To formally describe the new theorem, We first write $\{p_{\min}(\mathcal{I}, \mathcal{I}'), p_{\max}(\mathcal{I}, \mathcal{I}')\}$ as the minimum and the max-

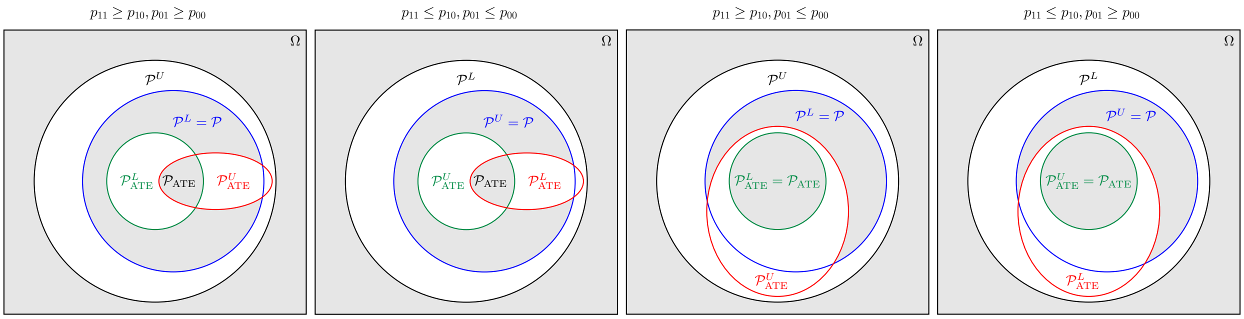

FIGURE 3: *The affiliation relationship of the sets $\mathcal{P}^L, \mathcal{P}^U, \mathcal{P}, \mathcal{P}_{\text{ATE}}, \mathcal{P}^L_{\text{ATE}}$ and $\mathcal{P}^U_{\text{ATE}}$ under different constraints of $\mathbb{P}(X,Y)$; the constraint is displayed at the top of each figure, where for example $\mathbb{P}(X=0,Y=1)$ is denoted as p_{01}. With a slight abuse of notation, $\mathcal{P}^L, \mathcal{P}^U$ and $\mathcal{P}$ correspond to the identification region of the interventional probability $\mathbb{P}(Y=1 \mid do(X=1))$. The whole space of $\mathbb{P}(U)$ is denotes as $\Omega := \{\mathbb{P}(U) : \sum_t \mathbb{P}(U=t) = 1, \mathbb{P}(U=t) \geq 0\}$. Corollary 4.3 guarantees that the gray region is non-empty (contains at least one legitimate $\mathbb{P}(U)$ in Ω).*

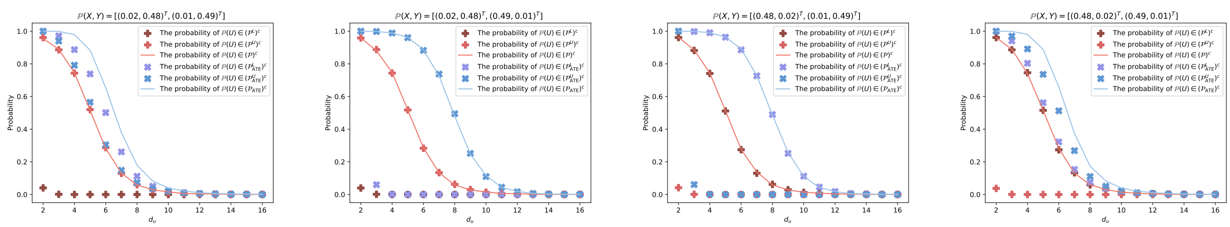

FIGURE 4: *The probability that $\mathbb{P}(U)$ satisfies the if and only if condition given by Theorem 4.1 with varying d_u. Here $\mathbb{P}(U)$ is uniformly sampled on the $(d_u-1)-$ probability simplex via 10^6 Monte Carlo simulations. There are four types of observed data which are recorded as $\mathbb{P}(X,Y) = [(\mathbb{P}(X=1,Y=1), \mathbb{P}(X=1,Y=0))^T, (\mathbb{P}(X=0,Y=1), \mathbb{P}(X=0,Y=0))^T], x=y=1$. The probability of $\mathbb{P}(U) \in (\mathcal{P})^c$ and $\mathbb{P}(U) \in (\mathcal{P}_{\text{ATE}})^c$ both monotonically decreases to zero with increasing d_u, and the degeneration rate of $\mathbb{P}(U) \in (\mathcal{P})^c$ is lower than that of $\mathbb{P}(U) \in (\mathcal{P}_{\text{ATE}})^c$.*

imum of the set $\{\mathbb{P}(U \in \mathcal{U}) : \mathcal{U} \subseteq \mathcal{I}, \mathbb{P}(U \in \mathcal{U}) \in \mathcal{I}'\}$. If this set is empty, we let $p_{\min}(\mathcal{I},\mathcal{I}') = -\infty$ and $p_{\max}(\mathcal{I},\mathcal{I}') = +\infty$. Armed with this notation, in Theorem 4.5, we provide the closed-form solution of the tight identification region of the interventional probability.

The computation of $p_{\min}(\mathcal{I},\mathcal{I}'), p_{\max}(\mathcal{I},\mathcal{I}')$ refer to the famous Subset-Sum Problem (SSP) (J Kleinberg, 2006) in theoretical computer science. Widely-existed SSP algorithms (J Kleinberg, 2006) could approximately extract the extreme subset-sum larger (smaller) than a given threshold (Appendix L). In this sense, $\{p_{\min}(\mathcal{I},\mathcal{I}'), p_{\max}(\mathcal{I},\mathcal{I}')\}$ can be viewed as constants that can be adequately approximated.

Theorem 4.5. *Suppose Assumptions 3.1-3.2 hold, and $\mathbb{P}(U)$ with $d_u \geq 2$ is observable. The tight identification region of the interventional probability $\mathbb{P}(y \mid do(x))$ is given by $\left[\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)), \mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U))\right]$, where*

$$\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) := \begin{cases} \mathcal{B}'(\mathbb{P}(U); x, y) & \mathbb{P}(U) \in (\mathcal{P}^L)^c \\ \mathbb{P}(x,y) & \mathbb{P}(U) \in \mathcal{P}^L \end{cases}, \tag{10}$$

$$\mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) := \begin{cases} 1 - \mathcal{B}'(\mathbb{P}(U); x, \neg y) & \mathbb{P}(U) \in (\mathcal{P}^U)^c \\ \mathbb{P}(x,y) + \mathbb{P}(\neg x) & \mathbb{P}(U) \in \mathcal{P}^U \end{cases}, \tag{11}$$

and $\mathcal{B}'(\mathbb{P}(U); x', y'), \{x', y'\} \in \{0,1\}$ is defined as:

$$\mathcal{B}'(\mathbb{P}(U); x', y') = \min \bigcup_{t \in \mathcal{A}} \Big\{ s + \frac{\mathbb{P}(x',y') - s}{\mathbb{P}(x') - s} \mathbb{P}(U=t) : s \in \{p_{\min}(\mathcal{I}_t, \mathcal{I}_t'), p_{\max}(\mathcal{I}_t, \mathcal{I}_t')\} \neq \emptyset \Big\}.$$

Here $\mathcal{A} := \{u : \mathbb{P}(U=u) \geq \mathbb{P}(x', \neg y')\} \neq \emptyset$, $\mathcal{I}_t := \mathbb{R}/\{t\}$ and $\mathcal{I}_t' := \big[0 \vee \big(\mathbb{P}(x') - \mathbb{P}(U=t)\big), \mathbb{P}(x',y')\big]$.

To better understand how $\mathbb{P}(U)$ helps improve identification, we introduce a new measure indicating the "distance" between the set $\{\mathbb{P}(U \in \mathcal{U}) : \mathcal{U} \in \mathbb{R}\}$ and the interval $\mathcal{I}$:

$$D\big(\mathbb{P}(U), \mathcal{I}\big) := \min |\mathbb{P}(U \in \mathcal{U}) - t|, s.t. \mathcal{U} \subseteq \mathbb{R}, t \in \mathcal{I}.$$

The following theorem shows that the identification improvement with prior knowledge of $\mathbb{P}(U)$ can be bounded by quantities depending on this new measure:

Proposition 4.6. *Consider a $\mathbb{P}(X,Y)$ and $\mathbb{P}(U)$; write*

$$\alpha_{x'} = 1/\mathbb{P}(x') \quad \& \quad \beta_{x',y'} = (\mathbb{P}(\neg x') \vee \mathbb{P}(x',y'))/\mathbb{P}(x, \neg y'),$$

$$\Delta_{x',y'} = D(\mathbb{P}(U), [\mathbb{P}(x', y'), \mathbb{P}(x')]),$$

where $x', y' \in \{0, 1\}$, *then we have that*

$$\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) - \mathbb{P}(x, y) \in [\alpha_x \Delta^2_{x,y}, \beta_{x,y}\Delta_{x,y}];$$

$$\mathbb{P}(x, y) + \mathbb{P}(\neg x) - \mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) \in [\alpha_x \Delta^2_{x,\neg y}, \beta_{x,\neg y}\Delta_{x,\neg y}],$$

Details show in Appendix J. Remark that when $\mathbb{P}(U)$ is in the set $\mathcal{P}^L$, then $D(\mathbb{P}(U), [\mathbb{P}(x, y), \mathbb{P}(x)])$ is always equal to zero. Informally then, it means that the theoretical improvement taking into account prior knowledge of $\mathbb{P}(U)$ depends on the distance between $\{\mathbb{P}(U \in \mathcal{U}) : \mathcal{U} \in \mathbb{R}\}$ and the intervals $[\mathbb{P}(x, y), \mathbb{P}(x)]$ or $[\mathbb{P}(x, \neg y), \mathbb{P}(x)]$.

Moving forward, we now consider the identification region of ATE. Inheriting the definition of $D\big(\mathbb{P}(U), \mathcal{I}\big)$, we use $D_{\text{ATE}}\big(\mathbb{P}(U), \{\mathcal{I}, \mathcal{I}'\}\big)$ to represent

$$\min\Big(|\mathbb{P}(U \in \mathcal{U}_0) - t_0| + |\mathbb{P}(U \in \mathcal{U}_1) - t_1|\Big),$$
$$s.t.\, \mathcal{U}_0, \mathcal{U}_1 \subseteq \mathbb{R}, \mathcal{U}_0 \cap \mathcal{U}_1 = \emptyset, t_0 \in \mathcal{I}, t_1 \in \mathcal{I}'.$$

Recalling the definitions of $\mathcal{I}_{x',y'}$ in Theorem 4.2 and $\alpha_{x'}, \beta_{x',y'}$ in Proposition 4.6, we now have the following result, which can be used for the construction of a valid bound of ATE:

Proposition 4.7. *The lower tight identification bound of average treatment effect* $\underline{\text{ATE}}$ *is controlled by* $\underline{\text{ATE}} - \text{ATE}^{\text{L}}_{\text{vanilla}} \in$

$$[(\alpha_1 \Delta^2_{1,1} + \alpha_0 \Delta^2_{0,0}) \vee (\Delta^2_{\text{ATE}}/d_u), (\beta_{1,1} + \beta_{0,0})\Delta_{\text{ATE}}],$$

where $\Delta_{\text{ATE}} = D_{\text{ATE}}(\mathbb{P}(U), \{\mathcal{I}_{0,0}, \mathcal{I}_{1,1}\})$. *Analogously, the upper tight identification bound of average treatment effect* $\overline{\text{ATE}}$ *is controlled by* $\text{ATE}^{\text{U}}_{\text{vanilla}} - \overline{\text{ATE}} \in$

$$[(\alpha_0 \Delta^2_{0,1} + \alpha_1 \Delta^2_{1,0}) \vee (\Delta^2_{\text{ATE}}/d_u), (\beta_{0,1} + \beta_{1,0})\Delta_{\text{ATE}}],$$

where $\Delta_{\text{ATE}} = D_{\text{ATE}}(\mathbb{P}(U), \{\mathcal{I}_{0,1}, \mathcal{I}_{1,0}\})$.

$\Delta_{x',y'}, x', y' \in \{0, 1\}$ is identified as above. Proposition 4.7 (Appendix K) indicates that for ATE, the difference between the tight PI bound and the vanilla bound is controlled by $D_{\text{ATE}}(\mathbb{P}(U), \{\mathcal{I}_{0,0}, \mathcal{I}_{1,0}\})$ and $D_{\text{ATE}}(\mathbb{P}(U), \{\mathcal{I}_{0,1}, \mathcal{I}_{1,1}\})$ between the linear and squared convergence rate.

It contributes a powerful theoretical bound rather than a natural composite of Proposition 4.6 upon both the upper and lower bounds of $\mathbb{P}(Y = 1 \mid do(x')), x' = 0, 1$. Such enhancement is due to our new measure being stricter: for any given $\{\mathcal{I}, \mathcal{I}'\}$, $D_{\text{ATE}}\big(\mathbb{P}(U), \{\mathcal{I}, \mathcal{I}'\}\big) \geq D\big(\mathbb{P}(U), \mathcal{I}\big) \vee D\big(\mathbb{P}(U), \mathcal{I}'\big)$ always holds through their definitions.

5. Auxiliary experiments

After theoretically proving the optimality, we focus on highlighting two additional key observations inspired by our PI, which have not been empirically validated in the previous literature: i) traditional information-theoretic PI bounds indeed lose information; ii) for the only bound mentioned in the main text that guarantees validity but not tightness (valid ATE bound in Proposition 4.7), we verify that its efficiency still significantly surpasses the competitive baseline and guides decision making.

For our first goal, it is visualized that given confounder information, our tight PI could grasp more non-vanilla cases than entropy-based methods, especially when entropy is not sufficiently small as the previous (Li et al., 2023; Jiang et al., 2023); for our second goal, we conduct experiments on INSURANCE dataset (Binder et al., 1997) and the ADULT dataset (Dua & Graff, 2017). Our result shows that even without a tightness guarantee, our PI bounds of ATE (Proposition 4.7) still provide more reliable information than the previous methods of separately estimating the upper and lower bounds of $p(Y = 1|x'), x' = 0, 1$ (Jiang et al., 2023) and could guide decision making. We refer readers for detailed design in Appendix N due to space limitations.

6. Discussion

In this paper, we focus on the PI of causal estimands with marginal confounder information; in particular, we have developed a closed-from tight identification region with causal structure following Figure 1, allowing the latent confounders to follow arbitrary distribution; we also establish the if and only if conditions for the identification region to be tighter than the vanilla bound. Such if and only if criteria establish the intrinsic equivalence between classical causal queries and subset-sum algorithms in theoretical computer science. We indicate that latent confounder information may not be very helpful in aiding PI when the cardinality of confounders is relatively large. We also develop in our manuscript several metrics to evaluate the improvement brought out by $\mathbb{P}(U)$ compared to without such information.

We believe this is not only of theoretical interest but also provides important guidance for practitioners on whether to collect information of $\mathbb{P}(U)$ when such information is not directly accessible in the first place. Our theory shows that practitioners are not recommended to spend much energy on collecting distributional information of $\mathbb{P}(U)$ when the cardinality of U is relatively large (e.g., larger than 10 according to our simulations).

Our paper has opened up several research directions; one is to extend the current result to the more complex causal graph; another is to consider PI of more complex counterfactual queries (e.g., Pearl (2022)) based on the subset-sum setting. Moreover, it would be of interest to combine our tight PI framework to facilitate other auxiliary-based PI methods. We will leave these possibilities for future work.

7. Impact Statements

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Balazadeh Meresht, V., Syrgkanis, V., and Krishnan, R. G. Partial identification of treatment effects with implicit generative models. *Advances in Neural Information Processing Systems*, 35:22816–22829, 2022.

Balke, A. and Pearl, J. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pp. 46–54. Elsevier, 1994.

Balke, A. and Pearl, J. Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176, 1997.

Binder, J., Koller, D., Russell, S., and Kanazawa, K. Adaptive probabilistic networks with hidden variables. *Machine Learning*, 29:213–244, 1997.

Castro, D. C., Walker, I., and Glocker, B. Causality matters in medical imaging. *Nature Communications*, 11(1):3673, 2020.

Chickering, D. M. and Meek, C. Finding optimal bayesian networks. *arXiv preprint arXiv:1301.0561*, 2012.

Christopher Frey, H. and Patil, S. R. Identification and review of sensitivity analysis methods. *Risk analysis*, 22 (3):553–578, 2002.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, pp. 1–12, 2023.

Dawid, A. P., Musio, M., and Murtas, R. The probability of causation. *Law, Probability and Risk*, 16(4):163–179, 2017.

Deaner, B. Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*, 2018.

Dorn, J., Guo, K., and Kallus, N. Doubly-valid/doubly-sharp sensitivity analysis for causal inference with unmeasured confounding. *arXiv preprint arXiv:2112.11449*, 2021.

Dua, D. and Graff, C. Uci machine learning repository (2020). *URL http://archive. ics. uci. edu/ml*, 2017.

Duarte, G., Finkelstein, N., Knox, D., Mummolo, J., and Shpitser, I. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, (just-accepted):1–25, 2023.

Gabriel, E. E., Sachs, M. C., and Sjölander, A. Causal bounds for outcome-dependent sampling in observational studies. *Journal of the American Statistical Association*, 117(538):939–950, 2022.

Geiger, D. and Meek, C. Quantifier elimination for statistical problems. *arXiv preprint arXiv:1301.6698*, 2013.

Geiger, P., Janzing, D., and Schölkopf, B. Estimating causal effects by bounding confounding. In *UAI*, pp. 240–249, 2014.

Ghassami, A., Shpitser, I., and Tchetgen, E. T. Partial identification of causal effects using proxy variables. *arXiv preprint arXiv:2304.04374*, 2023.

Guo, W., Yin, M., Wang, Y., and Jordan, M. Partial identification with noisy covariates: A robust optimization approach. In *Conference on Causal Learning and Reasoning*, pp. 318–335. PMLR, 2022.

Hicks, J. et al. *Causality in economics*. Australian National University Press, 1980.

Hu, Y., Wu, Y., Zhang, L., and Wu, X. A generative adversarial framework for bounding confounded causal effects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12104–12112, 2021.

J Kleinberg, E. T. *Algorithm Design*. 2006.

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Schölkopf, B. Quantifying causal influences. *The annals of statistics*, 2013.

Jiang, Z., Wei, L., and Kocaoglu, M. Approximate causal effect identification under weak confounding. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Kallus, N., Mao, X., and Zhou, A. Interval estimation of individual-level causal effects under unobserved confounding. In *The 22nd international conference on artificial intelligence and statistics*, pp. 2281–2290. PMLR, 2019.

Kallus, N., Mao, X., and Uehara, M. Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*, 2021.

Kitagawa, T. Identification region of the potential outcome distributions under instrument independence. *Journal of Econometrics*, 2009.

Kline, B. and Tamer, E. Recent developments in partial identification. *Annual Review of Economics*, 15:125–150, 2023.

Kuroki, M. and Pearl, J. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437, 2014.

Li, A. and Pearl, J. Bounds on causal effects and application to high dimensional data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5773–5780, 2022.

Li, A., Mueller, S., and Pearl, J. Epsilon-identifiability of causal quantities. *arXiv preprint arXiv:2301.12022*, 2023.

Manski, C. F. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Marmarelis, M. G., Haddad, E., Jesson, A., Jahanshad, N., Galstyan, A., and Ver Steeg, G. Partial identification of dose responses with hidden confounders. In *Uncertainty in Artificial Intelligence*, pp. 1368–1379. PMLR, 2023.

Masten, M. A. and Poirier, A. Identification of treatment effects under conditional partial independence. *Econometrica*, 86(1):317–351, 2018. doi: https://doi.org/10.3982/ECTA14481. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA14481.

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

Mueller, S., Li, A., and Pearl, J. Causes of effects: Learning individual responses from population data. *arXiv preprint arXiv:2104.13730*, 2021.

Nagasawa, K. Treatment effect estimation with noisy conditioning variables. *arXiv preprint arXiv:1811.00667*, 2018.

Park, C. and Tchetgen, E. T. Single proxy synthetic control. *arXiv preprint arXiv:2307.16353*, 2023.

Pearl, J. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Pearl, J. Comment: understanding simpson's paradox. *The American Statistician*, pp. 68(1), 8–13, 2014.

Pearl, J. Probabilities of causation: three counterfactual interpretations and their identification. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 317–372. 2022.

Pearl, J. et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.

Peng, K. and Knowles, E. D. Culture, education, and the attribution of physical causality. *Personality and Social Psychology Bulletin*, 29(10):1272–1284, 2003.

Qi, Z., Miao, R., and Zhang, X. Proximal learning for individualized treatment regimes under unmeasured confounding. *Journal of the American Statistical Association*, pp. 1–14, 2023.

Richardson, T. S. and Robins, J. M. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Robins, J. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of chronic diseases*, 40:139S–161S, 1987.

Robins, J. M. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pp. 113–159, 1989.

Rothman, K. J., Greenland, S., and Lash, T. L. *Modern epidemiology, pages 345–380. Lippincott Williams & Wilkins, Philadelphia, PA, 3rd edition.* Lippincott Williams & Wilkins, 2008.

Rubin, D. B. The bayesian bootstrap. *The annals of statistics*, pp. 130–134, 1981.

Schuster, T., Pang, M., and Platt, R. W. On the role of marginal confounder prevalence–implications for the high-dimensional propensity score algorithm. *Pharmacoepidemiology and drug safety*, 24(9):1004–1007, 2015.

Shi, X., Miao, W., Nelson, J. C., and Tchetgen Tchetgen, E. J. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):521–540, 2020.

Singh, R. Kernel methods for unobserved confounding: Negative controls, proxies, and instruments. *arXiv preprint arXiv:2012.10315*, 2020.

Swanson, S. A., Hernán, M. A., Miller, M., Robins, J. M., and Richardson, T. S. Partial identification of the average treatment effect using instrumental variables: review of methods for binary instruments, treatments, and outcomes. *Journal of the American Statistical Association*, 113(522):933–947, 2018.

Tamer, E. Partial identification in econometrics. *Annu. Rev. Econ.*, 2(1):167–195, 2010.

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Tchetgen, E. T., Park, C., and Richardson, D. Single proxy control. *arXiv preprint arXiv:2302.06054*, 2023.

Tian, J. and Pearl, J. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

Tian, J. and Pearl, J. A general identification condition for causal effects. In *Aaai/iaai*, pp. 567–573, 2002.

Wright, P. G. *The tariff on animal and vegetable oils*. Number 26. Macmillan, 1928.

Xu, L. and Gretton, A. Kernel single proxy control for deterministic confounding. *arXiv preprint arXiv:2308.04585*, 2023.

Zhang, J. and Bareinboim, E. Bounding causal effects on continuous outcome. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 12207–12215, 2021a.

Zhang, J. and Bareinboim, E. Non-parametric methods for partial identification of causal effects. *Columbia CausalAI Laboratory Technical Report*, 2021b.

SUPPLEMENT TO "TIGHT PARTIAL IDENTIFICATION OF CAUSAL EFFECT VIA CONFOUNDER INFORMATION"

Appendix A supplements a review of previous literature on PI. It confirms the originality of our tight PI region.

Appendix B contains the complete proof of Theorem 3.3, including the validity and sufficiency parts.

Appendix C proves Corollary 3.4, where we establish the tight PI region for the small entropy confounders.

Appendix D is for Theorem 3.5, which extends Theorem 3.3 from interventional probability to the ATE case.

Appendix E-F prove the IFF condition of falling into the vanilla case for interventional probability and ATE in multi-valued confounders, respectively.

Appendix H further analyzes the degeneration property after proposing the IFF condition as above, which is summarized as Corollary 4.4.

Appendix I-J justify Theorem 4.5 and Proposition 4.6 in the main text. Then Appendix K illustrates the valid identification region of ATE and its changing trend under the given marginal distribution of confounders.

In addition, Appendix L and Appendix M showcase the auxiliary lemma and algorithms that are presented in the above analysis and the main text, and Appendix N provides auxiliary experiment results.

A. Review of partial identification

TABLE 1: *The summary of previous causal effect identification. In our paper, we are the first to construct the closed-form tight PI of causal effects solely via marginal confounder information without additional hyper-parameters or auxiliary variables. Noteworthy, Duarte et al. (2023) claimed a tight bound without closed-form, and they usually achieve non-tight bound during approximation in practice.*

Literature	Model		Result		External variables/assumptions
	Hyperparametric	Non-hyperparametric	Point identification	Partial identification	
(Balke & Pearl, 1994) (Kitagawa, 2009)	✖	✔	✖	✔(tight)	Instrument variables
(Kuroki & Pearl, 2014) (Rothman et al., 2008) (Miao et al., 2018)	✖	✔	✔	✖	Negative control
(Ghassami et al., 2023)	✔	✖	✖	✔(valid)	
(Nagasawa, 2018) (Shi et al., 2020) (Singh, 2020) (Cui et al., 2023) (Tchetgen et al., 2020) (Deaner, 2018) (Kallus et al., 2021) (Qi et al., 2023)	✔	✖	✔	✖	
(Gabriel et al., 2022)	✖	✔	✖	✔(tight)	Outcome-dependent Sampling
(Li et al., 2023) (Geiger et al., 2014) (Zhang & Bareinboim, 2021b) (Duarte et al., 2023)	✖	✔	✖	✔(valid)	Confounder information
(Jiang et al., 2023)	✔	✖	✖	✔(valid)	
(Guo et al., 2022) (Masten & Poirier, 2018)	✔	✖	✖	✔(tight)	
Ours	✖	✔	✖	✔(tight)	

B. The proof of Theorem 3.3

Since we consider the binary scenario of confounders U, for ease of presentation we simply write $\mathbb{P}(U = t)$ as $\mathbb{P}(u_t)$ and $\mathbb{P}(U = t, x, y)$ as $\mathbb{P}(u_t, x, y)$ for $t \in \{0, 1\}$. Moreover, RHS (LHS) denotes the 'right (left) hand side'.

Lemma B.1. *We have*

$$\mathbb{P}(y \mid do(x)) \in [\min_{t=\{0,1\}}\{\max(\mathcal{S}_t)\}, \max_{t=\{0,1\}}\{\min(\mathcal{S}_t)\}],$$

where

$$\mathcal{S}_t = \left\{\frac{\mathbb{P}(x,y)}{\mathbb{P}(U=t,x)}\mathbb{P}(U=t), \frac{\mathbb{P}(x,y)-\mathbb{P}(U=t,x)}{\mathbb{P}(x)-\mathbb{P}(U=t,x)}\mathbb{P}(U=1-t)+\mathbb{P}(U=t)\right\}. \tag{12}$$

For brevity, we add the notation for the elements of $\mathcal{S}_t$ before the proof:

$$\begin{aligned}
\mathcal{S}_0 &= \left\{\underbrace{\frac{\mathbb{P}(x,y)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0)}_{T_{00}}, \underbrace{\frac{\mathbb{P}(x,y)-\mathbb{P}(u_0,x)}{\mathbb{P}(x)-\mathbb{P}(u_0,x)}\mathbb{P}(u_1)+\mathbb{P}(u_0)}_{T_{01}}\right\}. \\
\mathcal{S}_1 &= \left\{\underbrace{\frac{\mathbb{P}(x,y)}{\mathbb{P}(u_1,x)}\mathbb{P}(u_1)}_{T_{10}}, \underbrace{\frac{\mathbb{P}(x,y)-\mathbb{P}(u_1,x)}{\mathbb{P}(x)-\mathbb{P}(u_1,x)}\mathbb{P}(u_0)+\mathbb{P}(u_1)}_{T_{11}}\right\},
\end{aligned} \tag{13}$$

Proof of Lemma B.1 We first consider the lower bound, it is equal to prove $\mathbb{P}(y \mid do(x)) \geq \min\{\max(\mathcal{S}_0), \max(\mathcal{S}_1)\}$ It suffices if we can prove that $\mathbb{P}(y \mid do(x))$ is no smaller than $\min\{T_{00}, T_{10}\}$, $\min\{T_{00}, T_{11}\}$, $\min\{T_{01}, T_{10}\}$ and $\min\{T_{01}, T_{11}\}$, respectively. Below, we prove them one by one. Specifically, we prove them by contradiction.

- $\mathbb{P}(y \mid do(x)) \geq \min\{T_{00}, T_{10}\}$: Suppose in contradiction $\mathbb{P}(y \mid do(x)) < \min\{T_{00}, T_{10}\}$, then

  $$\mathbb{P}(y \mid do(x)) < \mathbb{P}(u_0 \mid x, y)T_{00} + \mathbb{P}(u_1 \mid x, y)T_{10}.$$

  Now expanding T_{00}, T_{10}, we have

  $$\mathbb{P}(y \mid do(x)) < \frac{\mathbb{P}(x,y,u_0)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \frac{\mathbb{P}(x,y,u_1)}{\mathbb{P}(u_1,x)}\mathbb{P}(u_1) = \mathbb{P}(y \mid do(x)),$$

  which raises a contradiction.

- $\mathbb{P}(y \mid do(x)) \geq \min\{T_{00}, T_{11}\}$: Suppose in contradiction $\mathbb{P}(y \mid do(x)) < \min\{T_{00}, T_{11}\}$, then

  $$\mathbb{P}(y \mid do(x)) < \mathbb{P}(\neg y \mid x, u_1)T_{00} + \mathbb{P}(y \mid x, u_1)T_{11}.$$

  Expanding T_{00} and T_{11}, we have

  $$\begin{aligned}
  \mathbb{P}(y \mid do(x)) &< \mathbb{P}(\neg y \mid x, u_1)\frac{\mathbb{P}(x,y)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \mathbb{P}(y \mid x, u_1)\frac{\mathbb{P}(x,y)-\mathbb{P}(u_1,x)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \mathbb{P}(y \mid x, u_1)\mathbb{P}(u_1) \\
  &= \frac{\mathbb{P}(x,y)-\mathbb{P}(x,y,u_1)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \mathbb{P}(y \mid x, u_1)\mathbb{P}(u_1) \\
  &= \frac{\mathbb{P}(x,y,u_0)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \mathbb{P}(y \mid x, u_1)\mathbb{P}(u_1) = \mathbb{P}(y \mid do(x)),
  \end{aligned}$$

  which raises a contradiction.

- $\mathbb{P}(y \mid do(x)) \geq \min\{T_{01}, T_{10}\}$: This can be directly obtained based on the duality of u_0, u_1 and $\mathbb{P}(y \mid do(x)) \geq \min\{T_{00}, T_{11}\}$.

- $\mathbb{P}(y \mid do(x)) \geq \min\{T_{01}, T_{11}\}$: Suppose in contradiction $\mathbb{P}(y \mid do(x)) < \min\{T_{01}, T_{11}\}$, then

$$\begin{aligned}
\mathbb{P}(y \mid do(x)) <& \mathbb{P}(u_1 \mid x, \neg y)T_{01} + \mathbb{P}(u_0 \mid x, \neg y)T_{11} \\
=& \mathbb{P}(u_1)\left[\frac{\mathbb{P}(x,y) - \mathbb{P}(u_0,x)}{\mathbb{P}(u_1,x)}\mathbb{P}(u_1 \mid x, \neg y) + \mathbb{P}(u_0 \mid x, \neg y)\right] \\
&+ \mathbb{P}(u_0)\left[\frac{\mathbb{P}(x,y) - \mathbb{P}(u_1,x)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0 \mid x, \neg y) + \mathbb{P}(u_1 \mid x, \neg y)\right] \\
=& \mathbb{P}(u_1)\left[\frac{\mathbb{P}(x,y) - \mathbb{P}(x)}{\mathbb{P}(u_1,x)}\mathbb{P}(u_1 \mid x, \neg y) + 1\right] + \mathbb{P}(u_0)\left[\frac{\mathbb{P}(x,y) - \mathbb{P}(x)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0 \mid x, \neg y) + 1\right] \\
=& \frac{\mathbb{P}(x,y,u_0)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \frac{\mathbb{P}(x,y,u_1)}{\mathbb{P}(u_1,x)}\mathbb{P}(u_1) = \mathbb{P}(y \mid do(x)).
\end{aligned}$$

Using an analogous analysis, we can prove the upper bound as well. It suffices if we can prove that $\mathbb{P}(y \mid do(x))$ is no larger than $\max\{T_{00}, T_{10}\}$, $\max\{T_{00}, T_{11}\}$, $\max\{T_{01}, T_{10}\}$ and $\max\{T_{01}, T_{11}\}$, respectively. Below we prove them one by one. Specifically, we prove them by contradiction.

- $\mathbb{P}(y \mid do(x)) \leq \max\{T_{00}, T_{10}\}$: Suppose in contradiction $\mathbb{P}(y \mid do(x)) > \max\{T_{00}, T_{10}\}$, then

$$\mathbb{P}(y \mid do(x)) > \mathbb{P}(u_0 \mid x, y)T_{00} + \mathbb{P}(u_1 \mid x, y)T_{10}.$$

  Now expanding T_{00}, T_{10}, we have

$$\mathbb{P}(y \mid do(x)) > \frac{\mathbb{P}(x,y,u_0)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \frac{\mathbb{P}(x,y,u_1)}{\mathbb{P}(u_1,x)}\mathbb{P}(u_1) = \mathbb{P}(y \mid do(x)),$$

  which raises a contradiction.

- $\mathbb{P}(y \mid do(x)) \leq \max\{T_{00}, T_{11}\}$: Suppose in contradiction $\mathbb{P}(y \mid do(x)) > \max\{T_{00}, T_{11}\}$, then

$$\mathbb{P}(y \mid do(x)) > \mathbb{P}(\neg y \mid x, u_1)T_{00} + \mathbb{P}(y \mid x, u_1)T_{11}.$$

  Expanding T_{00} and T_{11}, we have

$$\begin{aligned}
\mathbb{P}(y \mid do(x)) &> \mathbb{P}(\neg y \mid x, u_1)\frac{\mathbb{P}(x,y)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \mathbb{P}(y \mid x, u_1)\frac{\mathbb{P}(x,y) - \mathbb{P}(u_1,x)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \mathbb{P}(y \mid x, u_1)\mathbb{P}(u_1) \\
&= \frac{\mathbb{P}(x,y) - \mathbb{P}(x,y,u_1)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \mathbb{P}(y \mid x, u_1)\mathbb{P}(u_1) \\
&= \frac{\mathbb{P}(x,y,u_0)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \mathbb{P}(y \mid x, u_1)\mathbb{P}(u_1) = \mathbb{P}(y \mid do(x)),
\end{aligned}$$

  which raises a contradiction.

- $\mathbb{P}(y \mid do(x)) \leq \max\{T_{01}, T_{10}\}$: This can be directly obtained based on the duality of u_0, u_1 and $\mathbb{P}(y \mid do(x)) \leq \max\{T_{00}, T_{11}\}$.

- $\mathbb{P}(y \mid do(x)) \leq \max\{T_{01}, T_{11}\}$: Suppose in contradiction $\mathbb{P}(y \mid do(x)) > \max\{T_{01}, T_{11}\}$, then

$$\begin{aligned}
\mathbb{P}(y \mid do(x)) >& \mathbb{P}(u_1 \mid x, \neg y)T_{01} + \mathbb{P}(u_0 \mid x, \neg y)T_{11} \\
=& \mathbb{P}(u_1)\left[\frac{\mathbb{P}(x,y) - \mathbb{P}(u_0,x)}{\mathbb{P}(u_1,x)}\mathbb{P}(u_1 \mid x, \neg y) + \mathbb{P}(u_0 \mid x, \neg y)\right] \\
&+ \mathbb{P}(u_0)\left[\frac{\mathbb{P}(x,y) - \mathbb{P}(u_1,x)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0 \mid x, \neg y) + \mathbb{P}(u_1 \mid x, \neg y)\right] \\
=& \mathbb{P}(u_1)\left[\frac{\mathbb{P}(x,y) - \mathbb{P}(x)}{\mathbb{P}(u_1,x)}\mathbb{P}(u_1 \mid x, \neg y) + 1\right] + \mathbb{P}(u_0)\left[\frac{\mathbb{P}(x,y) - \mathbb{P}(x)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0 \mid x, \neg y) + 1\right] \\
=& \frac{\mathbb{P}(x,y,u_0)}{\mathbb{P}(u_0,x)}\mathbb{P}(u_0) + \frac{\mathbb{P}(x,y,u_1)}{\mathbb{P}(u_1,x)}\mathbb{P}(u_1) = \mathbb{P}(y \mid do(x)).
\end{aligned}$$

Putting together, we obtain the desired result $\mathbb{P}(y \mid do(x)) \leq \max\{\min\{T_{00}, T_{01}\}, \min\{T_{10}, T_{11}\}\}$. Combined with the lower bound and the upper bound, Lemma. B.1 has been proved.

■

After preparation, here we start the main proof of Theorem 3.3. Briefly, we do transformation on the above bound $[\min_{t=\{0,1\}}\{\max(\mathcal{S}_t)\}, \max_{t=\{0,1\}}\{\min(\mathcal{S}_t)\}]$ via observational data, and then demonstrate the tightness via construction.

Proof of Theorem 3.3 **(VALIDITY)** We first prove the lower bound. Exploiting Lemma B.1, it suffices to provide a lower bound of $\min\{\max\{T_{00}, T_{01}\}, \max\{T_{10}, T_{11}\}\}$ using the marginal probabilities $\mathbb{P}(x, y)$ and $\mathbb{P}(u)$.

We first consider $\max\{T_{10}, T_{11}\}$. For T_{10}, using $\mathbb{P}(u_1) \geq \mathbb{P}(u_1, x)$ and $\mathbb{P}(x) \geq \mathbb{P}(u_1, x)$, we have

$$\frac{\mathbb{P}(u_1)}{\mathbb{P}(u_1, x)} \geq \max\left\{1, \frac{\mathbb{P}(u_1)}{\mathbb{P}(x)}\right\},$$

which leads to

$$T_{10} = \frac{\mathbb{P}(x, y)}{\mathbb{P}(u_1, x)}\mathbb{P}(u_1) \geq \max\left\{\mathbb{P}(x, y), \mathbb{P}(y \mid x)\mathbb{P}(u_1)\right\} = \begin{cases} \mathbb{P}(x, y) & \mathbb{P}(u_1) \in (0, \mathbb{P}(x)] \\ \mathbb{P}(y \mid x)\mathbb{P}(u_1) & \mathbb{P}(u_1) \in (\mathbb{P}(x), 1] \end{cases}. \tag{14}$$

For T_{11}, we hope to use T_{11}'s information to construct $\max\{T_{10}, T_{11}\}$, and hence enhance the above piecewise lower estimate (14). Notice that when $\mathbb{P}(u_1) \leq \mathbb{P}(x, y)$, we have

$$\frac{\mathbb{P}(x, y) - \mathbb{P}(u_1, x)}{\mathbb{P}(x) - \mathbb{P}(u_1, x)} = 1 - \frac{\mathbb{P}(x, \neg y)}{\mathbb{P}(x) - \mathbb{P}(u_1, x)} \geq 1 - \frac{\mathbb{P}(x, \neg y)}{\mathbb{P}(x) - \mathbb{P}(u_1)} = \frac{\mathbb{P}(x, y) - \mathbb{P}(u_1)}{\mathbb{P}(x) - \mathbb{P}(u_1)},$$

which leads to

$$T_{11} = \frac{\mathbb{P}(x, y) - \mathbb{P}(u_1, x)}{\mathbb{P}(x) - \mathbb{P}(u_1, x)}\mathbb{P}(u_0) + \mathbb{P}(u_1) \geq \begin{cases} \frac{\mathbb{P}(x,y)-\mathbb{P}(u_1)}{\mathbb{P}(x)-\mathbb{P}(u_1)}(1 - \mathbb{P}(u_1)) + \mathbb{P}(u_1) & \mathbb{P}(u_1) \in (0, \mathbb{P}(x, y)] \\ -\infty & \mathbb{P}(u_1) \in (\mathbb{P}(x, y), 1) \end{cases} \tag{15}$$

The combination of (14)-(15) leads to a valid lower bound of $\max\{T_{10}, T_{11}\}$ as below:

$$\max\{T_{10}, T_{11}\} \geq \begin{cases} \frac{\mathbb{P}(x,y)-\mathbb{P}(u_1)}{\mathbb{P}(x)-\mathbb{P}(u_1)}(1 - \mathbb{P}(u_1)) + \mathbb{P}(u_1) & \mathbb{P}(u_1) \in (0, \mathbb{P}(x, y)] \\ \mathbb{P}(x, y) & \mathbb{P}(u_1) \in (\mathbb{P}(x, y), \mathbb{P}(x)] \\ \mathbb{P}(y \mid x)\mathbb{P}(u_1) & \mathbb{P}(u_1) \in (\mathbb{P}(x), 1) \end{cases} \tag{16}$$

Compared with the lower estimate of T_{10} (14), (16) further divided the case of $\mathbb{P}(u_1) \in [0, \mathbb{P}(x)]$ into the cases of $\mathbb{P}(u_1) \in [0, \mathbb{P}(x, y)]$ and $\mathbb{P}(u_1) \in [\mathbb{P}(x, y), \mathbb{P}(x)]$ in a more detailed manner.

Thanks to the duality between $\mathbb{P}(u_0)$ and $\mathbb{P}(u_1)$, we also have

$$\max\{T_{00}, T_{01}\} \geq \begin{cases} \frac{\mathbb{P}(x,y)-\mathbb{P}(u_0)}{\mathbb{P}(x)-\mathbb{P}(u_0)}(1 - \mathbb{P}(u_0)) + \mathbb{P}(u_0) & \mathbb{P}(u_0) \in (0, \mathbb{P}(x, y)] \\ \mathbb{P}(x, y) & \mathbb{P}(u_0) \in (\mathbb{P}(x, y), \mathbb{P}(x)] \\ \mathbb{P}(y \mid x)\mathbb{P}(u_0) & \mathbb{P}(u_0) \in (\mathbb{P}(x), 1) \end{cases} \tag{17}$$

In light of (16) and (17) together, we prove the validity of the lower bound of Theorem 3.3.

We now consider the upper bound $\max\{\min\{T_{00}, T_{01}\}, \min\{T_{10}, T_{11}\}\}$. Analogously, we first consider $\min\{T_{10}, T_{11}\}$. For T_{11}, we present two different upper bounds for all $\mathbb{P}(u_1) \in [0, 1]$. Firstly,

$$T_{11} = \frac{\mathbb{P}(x, y) - \mathbb{P}(u_1, x)}{\mathbb{P}(x) - \mathbb{P}(u_1, x)}\mathbb{P}(u_0) + \mathbb{P}(u_1) = \frac{-\mathbb{P}(x, \neg y)}{\mathbb{P}(u_0, x)}\mathbb{P}(u_0) + 1 \leq 1 - \mathbb{P}(x, \neg y) = \mathbb{P}(x, y) + \mathbb{P}(\neg x), \tag{18}$$

and secondly, due to $[\mathbb{P}(x, y) - \mathbb{P}(u_1, x)]\mathbb{P}(x) \leq \mathbb{P}(x, y)[\mathbb{P}(x) - \mathbb{P}(u_1, x)]$, we have

$$T_{11} = \frac{\mathbb{P}(x, y) - \mathbb{P}(u_1, x)}{\mathbb{P}(x) - \mathbb{P}(u_1, x)}\mathbb{P}(u_0) + \mathbb{P}(u_1) \leq \frac{\mathbb{P}(x, y)}{\mathbb{P}(x)}\mathbb{P}(u_0) + \mathbb{P}(u_1). \tag{19}$$

Combining (18)-(19) together, we can obtain a piecewise function of the form:

$$T_{11} \leq \begin{cases} \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)}(1-\mathbb{P}(u_1)) + \mathbb{P}(u_1) & \mathbb{P}(u_1) \in (0, \mathbb{P}(\neg x)] \\ \mathbb{P}(x,y) + \mathbb{P}(\neg x) & \mathbb{P}(u_1) \in (\mathbb{P}(\neg x), 1] \end{cases} . \tag{20}$$

For T_{10}, when $\mathbb{P}(u_1) \in [1-\mathbb{P}(x,\neg y), 1]$, namely when $\mathbb{P}(u_0) \in [0, \mathbb{P}(x, \neg y)]$, we have

$$\frac{\mathbb{P}(x,y)}{\mathbb{P}(x) - \mathbb{P}(x,u_0)} \leq \frac{\mathbb{P}(x,y)}{\mathbb{P}(x) - \mathbb{P}(u_0)}.$$

Hence the lower estimate of T_{10} can be constructed as

$$T_{10} = \mathbb{P}(x,y) + \frac{\mathbb{P}(x,y)}{\mathbb{P}(x) - \mathbb{P}(x,u_0)}\mathbb{P}(u_1, \neg x) \leq \begin{cases} \mathbb{P}(x,y) + \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)-\mathbb{P}(u_0)}\mathbb{P}(\neg x) & \mathbb{P}(u_1) \in (1-\mathbb{P}(x,\neg y), 1] \\ +\infty & \mathbb{P}(u_1) \in (0, 1-\mathbb{P}(x,\neg y)] \end{cases} \tag{21}$$

Combined with (21) and (20), we have

$$\min\{T_{10}, T_{11}\} \leq \begin{cases} \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)}(1-\mathbb{P}(u_1)) + \mathbb{P}(u_1) & \mathbb{P}(u_1) \in (0, \mathbb{P}(\neg x)] \\ \mathbb{P}(x,y) + \mathbb{P}(\neg x) & \mathbb{P}(u_1) \in (\mathbb{P}(\neg x), 1-\mathbb{P}(x,\neg y)] \\ \mathbb{P}(x,y) + \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)-\mathbb{P}(u_0)}\mathbb{P}(\neg x) & \mathbb{P}(u_1) \in (1-\mathbb{P}(x,\neg y), 1). \end{cases} \tag{22}$$

With again the duality between $\mathbb{P}(u_0)$ and $\mathbb{P}(u_1)$, we get

$$\min\{T_{00}, T_{01}\} \leq \begin{cases} \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)}(1-\mathbb{P}(u_0)) + \mathbb{P}(u_0) & \mathbb{P}(u_0) \in (0, \mathbb{P}(\neg x)] \\ \mathbb{P}(x,y) + \mathbb{P}(\neg x) & \mathbb{P}(u_0) \in (\mathbb{P}(\neg x), 1-\mathbb{P}(x,\neg y)] \\ \mathbb{P}(x,y) + \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)-\mathbb{P}(u_1)}\mathbb{P}(\neg x) & \mathbb{P}(u_0) \in (1-\mathbb{P}(x,\neg y), 1). \end{cases} \tag{23}$$

In light of both (22) and (23), we prove the validity of the upper bound.

(TIGHTNESS) We now prove that our identification strategy is tight; we first consider the tightness of the lower bound. We would like to prove that given any marginal distributions $\mathbb{P}(X,Y)$ and $\mathbb{P}(U)$, there exist two joint distributions of the three random variables so that their corresponding $\mathbb{P}(y \mid do(x))$'s are equal to $\mathcal{UB}(\mathbb{P}(u_0))$ and $\mathcal{UB}(\mathbb{P}(u_1))$, respectively. Furthermore, we will prove each point between the lower and upper bound of Theorem 3.3 is compatible with a joint distribution $\mathbb{P}(X,Y,U)$.

Recall that given any $\mathbb{P}(X,Y)$ and $\mathbb{P}(U)$, its corresponding $\mathbb{P}(y \mid do(x))$ is defined as

$$\mathbb{P}(y \mid do(x)) := \frac{\mathbb{P}(x,y,u_0)}{\mathbb{P}(x,y,u_0) + \mathbb{P}(x,\neg y,u_0)}\mathbb{P}(u_0) + \frac{\mathbb{P}(x,y,u_1)}{\mathbb{P}(x,y,u_1) + \mathbb{P}(x,\neg y,u_1)}\mathbb{P}(u_1). \tag{24}$$

We consider three categories: $\mathbb{P}(U=0) \in (0, \mathbb{P}(x,y)]$, $\mathbb{P}(U=0) \in (\mathbb{P}(x,y), \mathbb{P}(x)]$ and $\mathbb{P}(U=0) \in (\mathbb{P}(x), 1)$.

Case I: $\big(\mathbb{P}(U=0) \in (0, \mathbb{P}(x,y)]\big)$ We simply set

$$\mathbb{P}(x,y,u_0) = \mathbb{P}(u_0),\ \mathbb{P}(x,\neg y,u_0) = \mathbb{P}(\neg x,y,u_0) = \mathbb{P}(\neg x,\neg y,u_0) = 0;$$

and set $\mathbb{P}(x',y',u_1) = \mathbb{P}(x',y') - \mathbb{P}(x',y',u_0)$ for $x', y' \in \{0,1\}$. Apparently, through such construction all the joint probabilities are non-negative; moreover, they are compatible with the marginal distributions of (X,Y) and U, respectively. Moreover, with such construction, we have $\mathbb{P}(y \mid do(x))$ in (24) is equivalent to

$$\mathbb{P}(y \mid do(x)) = \frac{\mathbb{P}(u_0)}{\mathbb{P}(u_0)+0}\mathbb{P}(u_0) + \frac{\mathbb{P}(x,y) - \mathbb{P}(u_0)}{\mathbb{P}(x,y) + \mathbb{P}(x,\neg y) - \mathbb{P}(u_0)}\mathbb{P}(u_1) = \frac{\mathbb{P}(x,y) - \mathbb{P}(u_0)}{\mathbb{P}(x) - \mathbb{P}(u_0)}\mathbb{P}(u_1) + \mathbb{P}(u_0),$$

which is equivalent to $\mathcal{LB}(\mathbb{P}(u_0))$.

Case II: $\big(\mathbb{P}(U=0) \in (\mathbb{P}(x,y), \mathbb{P}(x))\big)$ We set

$$\mathbb{P}(x,y,u_0) = \mathbb{P}(x,y), \mathbb{P}(x,\neg y,u_0) = \mathbb{P}(u_0) - \mathbb{P}(x,y), \mathbb{P}(\neg x,y,u_0) = \mathbb{P}(\neg x,\neg y,u_0) = 0,$$

and set $\mathbb{P}(x',y',u_1) = \mathbb{P}(x',y') - \mathbb{P}(x',y',u_0)$ for $x',y' \in \{0,1\}$. Apparently, such construction is still non-negative and compatible with the observed marginal probabilities. With such construction, $\mathbb{P}(y \mid do(x))$ in (24) is equivalent to

$$\mathbb{P}(y \mid do(x)) = \frac{\mathbb{P}(x,y)}{\mathbb{P}(x,y) + \mathbb{P}(u_0) - \mathbb{P}(x,y)}\mathbb{P}(u_0) + \frac{0}{0 + \mathbb{P}(x) - \mathbb{P}(u_0)}\mathbb{P}(u_1) = \mathbb{P}(x,y).$$

This matches $\mathcal{LB}(\mathbb{P}(u_0))$.

Case III: $\big(\mathbb{P}(U=0) \in [\mathbb{P}(x), 1)\big)$ We set

$$\mathbb{P}(u_0,x) = \mathbb{P}(x), \mathbb{P}(u_1,x) = 0, \mathbb{P}(u_0,\neg x) = \mathbb{P}(u_0) - \mathbb{P}(x), \mathbb{P}(u_1,\neg x) = \mathbb{P}(u_1).$$

In this case, we instead construct directly the conditional probability as follows:

$$\mathbb{P}(y \mid u_0,x) = \mathbb{P}(y \mid x), \mathbb{P}(y \mid u_1,x) = 0, \mathbb{P}(y \mid u_0,\neg x) = \mathbb{P}(y \mid \neg x), \mathbb{P}(y \mid u_1,\neg x) = \mathbb{P}(y \mid \neg x),$$

and set $\mathbb{P}(\neg y \mid u^{'}, x^{'}) = 1 - \mathbb{P}(y \mid u^{'}, x^{'}), u^{'}, x^{'} \in \{0,1\}$. Apparently, with such construction, all the joint probabilities are nonnegative and compatible with the observed $\mathbb{P}(X,Y)$ and $\mathbb{P}(U)$. With such construction, the $\mathbb{P}(y \mid do(x))$ is equivalent to

$$\mathbb{P}(y \mid do(x)) = \mathbb{P}(y \mid u_0,x)\mathbb{P}(u_0) + \mathbb{P}(y \mid u_1,x)\mathbb{P}(u_1) = \mathbb{P}(y \mid x)\mathbb{P}(u_0),$$

which matches $\mathcal{LB}(\mathbb{P}(u_0))$.

According to the duality between $\mathbb{P}(u_0)$ and $\mathbb{P}(u_1)$, we can establish compatible joint probabilities to achieve the bounds given by $\mathcal{LB}(\mathbb{P}(u_1))$, thereby proving that the lower bound $\min\{\mathcal{LB}(\mathbb{P}(u_0)), \mathcal{LB}(\mathbb{P}(u_1))\}$ is matched with compatible $\mathbb{P}(X,Y,U)$.

We now turn to the upper bound. Again, we consider three cases, which are $\mathbb{P}(U=0) \in (0, \mathbb{P}(\neg x)]$, $\mathbb{P}(U=0) \in (\mathbb{P}(\neg x), 1 - \mathbb{P}(x,\neg y)]$ and $\mathbb{P}(U=0) \in (1 - \mathbb{P}(x,\neg y), 1)$.

Case I: $\big(\mathbb{P}(U=0) \in (0, \mathbb{P}(\neg x)]\big)$ Mimicking the construction of **Case III** of the lower bound, we set

$$\mathbb{P}(u_0,x) = 0, \mathbb{P}(u_1,x) = \mathbb{P}(x), \mathbb{P}(u_0,\neg x) = \mathbb{P}(u_0), \mathbb{P}(u_1,\neg x) = \mathbb{P}(u_1) - \mathbb{P}(x).$$

On this basis, analogous to **Case III** of the lower bound, the conditional probability is constructed as follows:

$$\mathbb{P}(y \mid u_0,x) = 1, \mathbb{P}(y \mid u_1,x) = \mathbb{P}(y \mid x), \mathbb{P}(y \mid u_0,\neg x) = \mathbb{P}(y \mid \neg x), \mathbb{P}(y \mid u_1,\neg x) = \mathbb{P}(y \mid \neg x),$$

and set $\mathbb{P}(\neg y \mid u^{'}, x^{'}) = 1 - \mathbb{P}(y \mid u^{'}, x^{'}), u^{'}, x^{'} \in \{0,1\}$. It can be verified that the non-negativity and compatibility of the joint probabilities under such construction also hold. Hence $\mathbb{P}(y \mid do(x))$ can be computed as

$$\mathbb{P}(y \mid do(x)) = \mathbb{P}(y \mid u_0,x)\mathbb{P}(u_0) + \mathbb{P}(y \mid u_1,x)\mathbb{P}(u_1) = \mathbb{P}(u_0) + \mathbb{P}(y \mid \neg x)\mathbb{P}(u_1),$$

which matches $\mathcal{UB}(\mathbb{P}(U=0))$.

Case II & III: For the tightness of the upper bound in these two cases, we instead set

$$\mathbb{P}(x,y,u_0) = \mathbb{P}(u_0) - \mathbb{P}(\neg x),\ \mathbb{P}(\neg x,y,u_0) = \mathbb{P}(\neg x,y),\ \mathbb{P}(x,\neg y,u_0) = 0,\ \mathbb{P}(\neg x,\neg y,u_0) = \mathbb{P}(\neg x,\neg y)$$

and

$$\mathbb{P}(x,y,u_0) = \mathbb{P}(x,y),\ \mathbb{P}(\neg x,y,u_0) = \mathbb{P}(\neg x,y),\ \mathbb{P}(x,\neg y,u_0) = \mathbb{P}(x,\neg y) - \mathbb{P}(u_1),\ \mathbb{P}(\neg x,\neg y,u_0) = \mathbb{P}(\neg x,\neg y),$$

respectively. Then we can use an analogous argument to prove that their induced $\mathbb{P}(y \mid do(x))$ matches the one given by $\mathcal{UB}(\mathbb{P}(u_0))$. Using again the duality between $\mathbb{P}(u_0)$ and $\mathbb{P}(u_1)$, we can use an analogous argument to prove the upper bound $\max\{\mathcal{UB}\left(\mathbb{P}(u_0)\right), \mathcal{UB}\left(\mathbb{P}(u_1)\right)\}$ can be achieved with matched $\mathbb{P}(X,Y,U)$.

Now we have proved that for each specification of $\mathbb{P}(X, Y)$ and $\mathbb{P}(U)$, there exists a compatible joint distribution so that its induced $\mathbb{P}(y \mid do(x))$ is equal to the lower and upper bound. Below we further prove that for each o between the two bounds, there exists a legitimate joint distribution with its corresponding $\mathbb{P}(y \mid do(x)) = o$. We first consider the case where $\mathbb{P}(u_0)$ or $\mathbb{P}(u_1)$ is equal to $\mathbb{P}(x)$. Without loss of generality, we just consider $\mathbb{P}(u_0) = \mathbb{P}(x)$. Then our proposed identification region is equal to $[\mathbb{P}(x, y), \mathbb{P}(x, y) + \mathbb{P}(\neg x)]$. Then we construct

$$\mathbb{P}(u_0, x) = \mathbb{P}(u_0), \mathbb{P}(u_1, x) = 0, \mathbb{P}(u_0, \neg x) = 0, \mathbb{P}(u_1, \neg x) = \mathbb{P}(\neg x).$$

Moreover, we set $\mathbb{P}(y \mid u_0, x) = \frac{\mathbb{P}(x,y)}{\mathbb{P}(u_0)}, \mathbb{P}(y \mid u_1, x) = \varepsilon, \mathbb{P}(y \mid u_0, \neg x) = 0, \mathbb{P}(y \mid u_1, \neg x) = \mathbb{P}(y \mid \neg x), \varepsilon \in [0, 1]$, and $\mathbb{P}(\neg y \mid u', x') = 1 - \mathbb{P}(y \mid u', x'), \forall u', x' \in \{0, 1\}$. Apparently, one can verify that this construction is compatible with the observed marginal distributions. On this basis, we have

$$\mathbb{P}(y \mid do(x)) = \mathbb{P}(y \mid u_0, x)\mathbb{P}(u_0) + \mathbb{P}(y \mid u_1, x)\mathbb{P}(u_1) = \mathbb{P}(x, y) + \varepsilon\mathbb{P}(\neg x).$$

By varying $\varepsilon \in [0, 1]$, all values in $[\mathbb{P}(x, y), \mathbb{P}(x, y) + \mathbb{P}(\neg x)]$ is achievable, which proves the desired result.

Now we further consider the more general case where $\mathbb{P}(u_0), \mathbb{P}(u_1) \neq \mathbb{P}(x)$. Given any fixed $\varepsilon > 0$, let

$$\mathcal{LB}_\varepsilon(t) := \begin{cases} \frac{\mathbb{P}(x,y)-t}{\mathbb{P}(x)-t}(1-t) + t & t \in (0, \mathbb{P}(x, y)] \\ \frac{\mathbb{P}(x,y)-\varepsilon}{t-\varepsilon}t + \frac{\varepsilon}{\varepsilon+\mathbb{P}(x)-t}(1-t) & t \in (\mathbb{P}(x, y), \mathbb{P}(x)] \\ \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)-\varepsilon}t & t \in (\mathbb{P}(x), 1) \end{cases}$$

and

$$\mathcal{UB}_\varepsilon(t) := \begin{cases} \frac{\mathbb{P}(x,y)-\varepsilon}{\mathbb{P}(x)-\varepsilon}(1-t) + t & t \in (0, \mathbb{P}(\neg x)] \\ \frac{t-\mathbb{P}(\neg x)}{t-\mathbb{P}(\neg x)+\varepsilon}t + \frac{\mathbb{P}(x,y)+\mathbb{P}(\neg x)-t}{1-t-\varepsilon}(1-t) & t \in (\mathbb{P}(\neg x), 1-\mathbb{P}(x, \neg y)] \\ \frac{\mathbb{P}(x,y)t}{\mathbb{P}(x)-(1-t)} & t \in (1-\mathbb{P}(x, \neg y), 1). \end{cases}$$

Apparently as $\varepsilon \to 0$, $\mathcal{LB}_\varepsilon(t)$ and $\mathcal{UB}_\varepsilon(t)$ converges to $\mathcal{LB}(t)$ and $\mathcal{UB}(t)$ when $t \in \{\mathbb{P}(u_0), \mathbb{P}(u_1)\}$, respectively. To prove that for each point $o \in (\min_{t\in\{0,1\}} \mathcal{LB}(\mathbb{P}(U = t)), \max_{t\in\{0,1\}} \mathcal{UB}(\mathbb{P}(U = t)))$, there exists a legitimate joint probability so that $\mathbb{P}(y \mid do(x)) = o$, we just need to prove that there exists a constant $\varepsilon_0 > 0$ sufficiently small such that for all $\varepsilon \in (0, \varepsilon_0]$,

$$\Big[\min_{t\in\{0,1\}} \mathcal{LB}_\varepsilon(\mathbb{P}(U = t)), \max_{t\in\{0,1\}} \mathcal{UB}_\varepsilon(\mathbb{P}(U = t))\Big] \tag{25}$$

is a subset of the identification region, i.e., that for all the points o' in the interval given by (25), there exists a legitimate joint distribution with its corresponding $\mathbb{P}(y \mid do(x)) \equiv o'$. To achieve this goal, we now consider a region

$$\mathcal{O}_\varepsilon := \Big\{\mathbb{P}(y \mid do(x)) : \forall u \in U, \mathbb{P}(u, x) \geq \varepsilon, \mathbb{P}(Y, U, X) \text{ is compatible with } \mathbb{P}(X, Y), \mathbb{P}(U)\Big\}. \tag{26}$$

Now if we treat $\mathbb{P}(x', y', u')$ $(x', y', u' \in \{0, 1\})$ as parameters and $\mathbb{P}(y \mid do(x))$ as a function of these parameters, one can easily verify that the parameter space restricted by $\mathcal{O}_\varepsilon$ is a convex and compact set; moreover, $\mathbb{P}(y \mid do(x))$ is a continuous and well-defined function w.r.t. all parameters in the restricted parameter space (since in the parameter space given by $\mathcal{O}_\varepsilon$, the denominators in (24) is always nonzero). In light of these, it is straightforward that $\mathcal{O}_\varepsilon$ is a closed interval on $\mathbb{R}$. Letting $o_{\min}, o_{\max}$ be the left and right side of the interval $\mathcal{O}_\varepsilon$; then one can easily verify that with ε_0 sufficiently small, for all $\varepsilon \in (0, \varepsilon_0]$,

$$o_{\min} \leq \min_{t\in\{0,1\}} \mathcal{LB}_\varepsilon(\mathbb{P}(U = t)) \leq \max_{t\in\{0,1\}} \mathcal{UB}_\varepsilon(\mathbb{P}(U = t)) \leq o_{\max},$$

which means that the region given by (25) is a subset of $\mathcal{O}_\varepsilon$. Since $\mathcal{O}_\varepsilon$ is a subset of the identification region, it's straightforward that the interval (25) is a subset of the identification region as well, which proves the desired result.

■

B.1. Further discussion: The identification of the vanilla bound

Recall our objective function is stated below:

$$\mathbb{P}(y \mid do(x)) = \mathbb{P}(x, y) + \mathbb{P}(y \mid u_0, x)\mathbb{P}(u_0, \neg x) + \mathbb{P}(y \mid u_1, x)\mathbb{P}(u_1, \neg x).$$

The lower vanilla bound. When $\mathbb{P}(y \mid do(x)) = \mathbb{P}(x, y)$, we have $\forall t = 0, 1, \mathbb{P}(y, u_t, x)\mathbb{P}(u_t, \neg x) = 0$. According to Assumption 3.1, notice the fact that $\mathbb{P}(y, u_0, x) + \mathbb{P}(y, u_1, x) = \mathbb{P}(y, x) > 0$ and $\mathbb{P}(u_0, \neg x) + \mathbb{P}(u_1, \neg x) = \mathbb{P}(\neg x) > 0$, it leads to $\exists t \in \{0, 1\}, \mathbb{P}(y, u_{\neg t}, x) = \mathbb{P}(u_t, \neg x) = 0$. Then

$$\mathbb{P}(u_t) = \mathbb{P}(u_t, x) \in [\mathbb{P}(y, u_t, x), \mathbb{P}(x)] = [\mathbb{P}(x, y), \mathbb{P}(x)].$$

Hence the necessity result is

$$\{\mathbb{P}(U = 0), \mathbb{P}(U = 1)\} \cap [\mathbb{P}(x, y), \mathbb{P}(x)] \neq \emptyset.$$

For the sufficiency part, we refer to **CASE II** for the tight lower-bound construction.

The upper vanilla bound. When $\mathbb{P}(y \mid do(x)) = \mathbb{P}(x, y) + \mathbb{P}(\neg x)$, it is equivalent to

$$\mathbb{P}(\neg y \mid u_0, x)\mathbb{P}(u_0, \neg x) + \mathbb{P}(\neg y \mid u_1, x)\mathbb{P}(u_1, \neg x) = 0.$$

Hence $\forall t = 0, 1, \mathbb{P}(\neg y, u_t, x)\mathbb{P}(u_t, \neg x) = 0$. According to Assumption 3.1, noticing the fact that $\mathbb{P}(\neg y, u_0, x) + \mathbb{P}(\neg y, u_1, x) = \mathbb{P}(\neg y, x) > 0$ and $\mathbb{P}(u_0, \neg x) + \mathbb{P}(u_1, \neg x) = \mathbb{P}(\neg x) > 0$,it leads to $\exists t \in \{0, 1\}, \mathbb{P}(\neg y, u_{\neg t}, x) = \mathbb{P}(u_t, \neg x) = 0$. Hence,

$$\mathbb{P}(u_t) = \mathbb{P}(u_t, x) \in [\mathbb{P}(\neg y, u_t, x), \mathbb{P}(x)] = [\mathbb{P}(x, \neg y), \mathbb{P}(x)].$$

Hence the necessity result is $\{\mathbb{P}(U = 0), \mathbb{P}(U = 1)\} \cap [\mathbb{P}(x, \neg y), \mathbb{P}(x)] \neq \emptyset$, namely

$$\{\mathbb{P}(U = 0), \mathbb{P}(U = 1)\} \cap [\mathbb{P}(\neg x), \mathbb{P}(\neg x) + \mathbb{P}(x, y)] \neq \emptyset.$$

For the sufficiency part, we refer to the constructions in **CASE II** (upper bound).

The above analysis produces the same result as in Theorem 3.3.

C. Proof of Corollary 3.4

Apparently, given the prior knowledge that $\mathbb{P}(U) \in \mathcal{P}_\varepsilon$, the tight identification region of the interventional probability should be given by $\cup_{\mathbb{P}(U) \in \mathcal{P}_\varepsilon} \mathscr{O}_{\mathbb{P}(U)}$, where

$$\mathscr{O}_{\mathbb{P}(U)} = \left[\min_{t \in \{0,1\}} \mathcal{LB}\left(\mathbb{P}(U = t)\right), \max_{t \in \{0,1\}} \mathcal{UB}\left(\mathbb{P}(U = t)\right) \right].$$

We now prove that the identification region in (6) is equivalent to the region stated above. To prove this, first, apparently,

$$\text{Identification region in (6)} \subseteq \cup_{\mathbb{P}(U) \in \mathcal{P}_\varepsilon} \mathscr{O}_{\mathbb{P}(U)}.$$

We now prove that the converse is also true. Without loss of generality, we force $\mathbb{P}(u_0) \in (0, \varepsilon]$ and $\mathbb{P}(u_1) \in [1 - \varepsilon, 1)$. Then using the monotonicity of $\mathcal{LB}(t)$ and $\mathcal{UB}(t)$ within the range $t \in [0, \min\{\mathbb{P}(x), \mathbb{P}(\neg x)\}]$, we have

$$\mathcal{LB}(\mathbb{P}(u_0)) \geq \mathcal{LB}(\varepsilon), \mathcal{LB}(\mathbb{P}(u_1)) \geq \mathcal{LB}(1 - \varepsilon); \quad \mathcal{UB}(\mathbb{P}(u_0)) \leq \mathcal{UB}(\varepsilon), \mathcal{UB}(\mathbb{P}(u_1)) \leq \mathcal{UB}(1 - \varepsilon),$$

which leads to

$$\cup_{\mathbb{P}(U) \in \mathcal{P}_\varepsilon} \mathscr{O}_{\mathbb{P}(U)} \subseteq \text{Identification region in (6)}.$$

We now turn to (7). Using again the monotonicity of $\mathcal{LB}(t)$ and $\mathcal{UB}(t)$ and the symmetry between u_0 and u_1, we just need to prove that the bound given by (7) is equivalent to $\mathscr{O}_{\mathbb{P}(U)}$ when $\mathbb{P}(u_0) = \varepsilon \leq \min\{\mathbb{P}(x, y), \mathbb{P}(x, \neg y), \mathbb{P}(\neg x)\}$. In this case, the LHS of $\mathscr{O}_{\mathbb{P}(U)}$ contains a simple form

$$\begin{aligned} &\mathcal{LB}\big(\mathbb{P}(u_0)\big) = \frac{\mathbb{P}(x, y) - \mathbb{P}(u_0)}{\mathbb{P}(x) - \mathbb{P}(u_0)}\big(1 - \mathbb{P}(u_0)\big) + \mathbb{P}(u_0), \mathcal{LB}\big(\mathbb{P}(u_1)\big) = \mathbb{P}(y \mid x)\big(1 - \mathbb{P}(u_0)\big). \text{ Then} \\ &\min\Big\{\mathcal{LB}\big(\mathbb{P}(u_0)\big), \mathcal{LB}\big(\mathbb{P}(u_1)\big)\Big\} = \mathbb{P}(y \mid x) - \varepsilon \max\Big\{\mathbb{P}(y \mid x), \frac{\mathbb{P}(x, \neg y)\mathbb{P}(\neg x)}{\mathbb{P}(x)\big(\mathbb{P}(x) - \varepsilon\big)}\Big\}. \end{aligned} \tag{27}$$

Secondly, we consider the tight upper bound. Due to $\varepsilon \in [0, \min\{\mathbb{P}(\neg x), \mathbb{P}(x, \neg y)\}]$, we have $\mathbb{P}(u_0) \in [0, \mathbb{P}(\neg x)], \mathbb{P}(u_1) \in [1-\mathbb{P}(x,\neg y), 1], t \in \{0,1\}$. Then RHS of $\mathscr{O}_{\mathbb{P}(U)}$ contains a simple form:

$$\begin{aligned}&\mathcal{UB}\big(\mathbb{P}(u_0)\big) = \mathbb{P}(y \mid x)\big(1-\mathbb{P}(u_0)\big) + \mathbb{P}(u_0), \mathcal{UB}\big(\mathbb{P}(u_1)\big) = \frac{\mathbb{P}(x,y)\big(1-\mathbb{P}(u_0)\big)}{\mathbb{P}(x)-\mathbb{P}(u_0)}. \text{ Then}\\ &\max\Big\{\mathcal{UB}\big(\mathbb{P}(u_0)\big), \mathcal{UB}\big(\mathbb{P}(u_1)\big)\Big\} = \mathbb{P}(y \mid x) + \varepsilon \max\Big\{\mathbb{P}(\neg y \mid x), \frac{\mathbb{P}(x,y)\mathbb{P}(\neg x)}{\mathbb{P}(x)\big(\mathbb{P}(x)-\varepsilon\big)}\Big\}.\end{aligned} \tag{28}$$

Hence the tight region $\mathscr{O}_{\mathbb{P}(U)}$ under $\{\mathbb{P}(u_0), \mathbb{P}(u_1)\}$ where $\mathbb{P}(u_0) \leq \varepsilon$ can be transformed to (notice that $\neg\neg y = y$):

$$\Big[\mathbb{P}(y \mid x) - \mathbb{P}(u_0)\mathscr{E}_y\big(\mathbb{P}(u_0)\big), \mathbb{P}(y \mid x) + \mathbb{P}(u_0)\mathscr{E}_{\neg y}\big(\mathbb{P}(u_0)\big)\Big], \text{ where } \mathscr{E}_{y'}(t) := \max\Big\{\mathbb{P}(y' \mid x), \frac{\mathbb{P}(x,\neg y')\mathbb{P}(\neg x)}{\mathbb{P}(x)(\mathbb{P}(x)-t)}\Big\}. \tag{29}$$

Putting together, we obtain the desired result. ■

Comparison to Li's result (Li et al., 2023). It would be convenient to introduce the original theorem of (Li et al., 2023) as follows.

Theorem C.1 ((Li et al., 2023)). *If* $\mathbb{P}(u_0) \leq \mathbb{P}(x) - c$, $0 < c \leqslant \mathbb{P}(x)$, *and we have* $\mathbb{P}(u_0) \leq \eta_c \varepsilon_0$, *then*

$$\Big| P\left(y \mid do(x)\right) - \Big(P(y \mid x) + \lambda_c \varepsilon_0\Big)\Big| \leq \varepsilon_0, \tag{30}$$

where $\varepsilon_0 > 0, \eta_c = 2c\mathbb{P}(x)/(2c\mathbb{P}(x) + \mathbb{P}(x) + c), \lambda_c = (\mathbb{P}(x) - c)/(2c\mathbb{P}(x) + \mathbb{P}(x) + c)$.

The equivalent form in our main text. We first prove that the form provided in our main text is equivalent to the above form. Considering $\mathcal{P}_\varepsilon$ where $\varepsilon > 0$, according to the duality of $\{\mathbb{P}(u_0), \mathbb{P}(u_1)\}$, it is equal to $\mathbb{P}(u_0) \leq \varepsilon$. To make use of the above theorem, we must force

$$\mathbb{P}(u_0) \in (0, \varepsilon], \text{ where } \min\Big\{\mathbb{P}(x) - c, \eta_c\varepsilon_0\Big\} = \varepsilon.$$

On this basis, the lower bound of Li's result (Li et al., 2023) is controlled by

$$\mathbb{P}(y \mid x) + \lambda_c\varepsilon_0 - \varepsilon_0 \leq \mathbb{P}(y \mid x) + (\lambda_c - 1)\frac{\varepsilon}{\eta_c} = \mathbb{P}(y \mid x) - \frac{\mathbb{P}(x)+1}{\mathbb{P}(x)}\varepsilon, \tag{31}$$

and the upper bound is controlled by

$$\mathbb{P}(y \mid x) + \lambda_c\varepsilon_0 + \varepsilon_0 \geq \mathbb{P}(y \mid x) + (\lambda_c + 1)\frac{\varepsilon}{\eta_c} = \mathbb{P}(y \mid x) + \frac{c+1}{c}\varepsilon \geq \mathbb{P}(y \mid x) + \frac{\mathbb{P}(x) - \varepsilon + 1}{\mathbb{P}(x) - \varepsilon}\varepsilon. \tag{32}$$

Hence, the best identification region Li et al. (2023) could achieve is

$$\mathbb{P}(y \mid do(x)) \in [\mathcal{LB}_{li}, \mathcal{UB}_{li}] := \Big[\mathbb{P}(y \mid x) - \frac{\mathbb{P}(x)+1}{\mathbb{P}(x)}\varepsilon, \mathbb{P}(y \mid x) + \frac{\mathbb{P}(x) - \varepsilon + 1}{\mathbb{P}(x) - \varepsilon}\varepsilon\Big] \text{ when } \mathbb{P}(U) \in \mathcal{P}_\varepsilon, \tag{33}$$

where $\varepsilon \in (0, \mathbb{P}(x)]$. Actually, this bound can be achieved via choosing $\mathbb{P}(x) - c = \eta_c\varepsilon_0 = \varepsilon$. This assignment process is legitimate. In sum, (33), which is presented in our main text, is the equivalent result of Li et al. (2023).

The justification of ε *region.* We argue that the region (Li et al., 2023) only works for $\varepsilon \in [0, \min\{\mathbb{P}(x), \mathbb{P}(\neg x)]$, instead of their claim $\varepsilon \in [0, \mathbb{P}(x)]$. In other words, when $\mathbb{P}(\neg x) \leq \mathbb{P}(x)$, their result will significantly fail when $\varepsilon \in [\mathbb{P}(\neg x), \mathbb{P}(x)]$, both for the lower and upper bounds. In this case, the identification region of (33) is transformed to

$$\begin{aligned}&\mathcal{LB}_{li} \leq \mathbb{P}(y \mid x) - \frac{\mathbb{P}(x)+1}{\mathbb{P}(x)}\mathbb{P}(\neg x) < \frac{\mathbb{P}(x,y) - \mathbb{P}(x,y)\mathbb{P}(\neg x)}{\mathbb{P}(x)} = \mathbb{P}(x,y). \text{(vanilla lower bound)}\\ &\mathcal{UB}_{li} \geq \mathbb{P}(x,y) + \frac{\mathbb{P}(x)+1}{\mathbb{P}(x)}\mathbb{P}(\neg x) > \mathbb{P}(x,y) + \mathbb{P}(\neg x). \text{(vanilla upper bound)}\end{aligned} \tag{34}$$

Seriously, their identification region $[\mathcal{LB}_{li}, \mathcal{UB}_{li}]$ is non-informative.

The dominance of our bound over Li et al. (2023). We have already argued that Li et al. (2023) does not work when $\varepsilon > \mathbb{P}(\neg x)$, both on the lower and upper bounds. Therefore, it only requires comparison within $\varepsilon \in (0, \min\{\mathbb{P}(x), \mathbb{P}(\neg x)\}]$. Notice that

$$\mathcal{LB}(\varepsilon) = \begin{cases} \frac{\mathbb{P}(x,y)-\varepsilon}{\mathbb{P}(x)-\varepsilon}(1-\varepsilon)+\varepsilon & \text{if } \varepsilon \le \mathbb{P}(x,y) \\ \mathbb{P}(x,y) & \text{if } \varepsilon > \mathbb{P}(x,y) \end{cases}, \mathcal{LB}(1-\varepsilon) = \mathbb{P}(y \mid x)(1-\varepsilon).$$

Hence

$$\min\{\mathcal{LB}(\varepsilon), \mathcal{LB}(1-\varepsilon)\} = \begin{cases} \mathbb{P}(y \mid x) - \varepsilon \mathscr{E}_y(\varepsilon) & \text{If } \varepsilon \le \mathbb{P}(x,y) \\ \mathbb{P}(x,y) & \text{If } \varepsilon > \mathbb{P}(x,y) \end{cases}. \tag{35}$$

Here $\mathscr{E}_y(\varepsilon)$ is identified in (29). Thus the enhancement from Li et al. (2023) to our optimal result is

$$\min\{\mathcal{LB}(\varepsilon), \mathcal{LB}(1-\varepsilon)\} - \mathcal{LB}_{li} = \begin{cases} \frac{\mathbb{P}(x)+1}{\mathbb{P}(x)}\varepsilon - \varepsilon \mathscr{E}_y(\varepsilon) & \text{If } \varepsilon \le \mathbb{P}(x,y) \\ \mathbb{P}(x,y) - \mathbb{P}(y \mid x) + \frac{\mathbb{P}(x)+1}{\mathbb{P}(x)}\varepsilon & \text{If } \varepsilon > \mathbb{P}(x,y) \end{cases} \tag{36}$$

which can be measured by

$$\min\{\mathcal{LB}(\varepsilon), \mathcal{LB}(1-\varepsilon)\} - \mathcal{LB}_{li} \ge \varepsilon + \Delta_y, \text{ where } \Delta_{y'} = \min\left\{\varepsilon, \frac{1-\mathbb{P}(x,y')}{\mathbb{P}(x)}\varepsilon, \mathbb{P}(x,y')\right\}. \tag{37}$$

Analogously, we shift our attention to the upper bound comparison. Notice that ($\varepsilon \in (0, \min\{\mathbb{P}(x), \mathbb{P}(\neg x)\}$)

$$\mathcal{UB}(\varepsilon) = \mathbb{P}(y \mid x)(1-\varepsilon)+\varepsilon, \mathcal{UB}(1-\varepsilon) = \begin{cases} \mathbb{P}(x,y)+\mathbb{P}(\neg x) & \text{if } \varepsilon \ge \mathbb{P}(x,\neg y) \\ \frac{\mathbb{P}(x,y)(1-\varepsilon)}{\mathbb{P}(x)-\varepsilon} & \text{if } \varepsilon < \mathbb{P}(x,\neg y) \end{cases}.$$

Hence

$$\max\{\mathcal{UB}(\varepsilon), \mathcal{UB}(1-\varepsilon)\} = \begin{cases} \mathbb{P}(y \mid x) + \varepsilon \mathscr{E}_{\neg y}(\varepsilon) & \text{If } \varepsilon \le \mathbb{P}(x,\neg y) \\ \mathbb{P}(x,y)+\mathbb{P}(\neg x) & \text{If } \varepsilon > \mathbb{P}(x,\neg y) \end{cases}. \tag{38}$$

Thus, the enhancement from Li et al. (2023) to our optimal result is

$$\mathcal{UB}_{li} - \max\{\mathcal{UB}(\varepsilon), \mathcal{UB}(1-\varepsilon)\} = \begin{cases} \frac{\mathbb{P}(x)-\varepsilon+1}{\mathbb{P}(x)-\varepsilon}\varepsilon - \varepsilon \mathscr{E}_{\neg y}(\varepsilon) & \text{If } \varepsilon \le \mathbb{P}(x,\neg y) \\ \mathbb{P}(y \mid x) + \frac{\mathbb{P}(x)-\varepsilon+1}{\mathbb{P}(x)-\varepsilon}\varepsilon - \mathbb{P}(x,y) - \mathbb{P}(\neg x) & \text{If } \varepsilon > \mathbb{P}(x,\neg y) \end{cases}, \tag{39}$$

which can be measured by

$$\mathcal{UB}_{li} - \max\{\mathcal{UB}(\varepsilon), \mathcal{UB}(1-\varepsilon)\} \ge \varepsilon + \Delta_{\neg y}. \tag{40}$$

In sum, we have proved our bound is strictly stronger than Li et al. (2023) within $\varepsilon \in [0, \min\{\mathbb{P}(x), \mathbb{P}(\neg x)\}]$, with at least $\varepsilon + \Delta_y$ and $\varepsilon + \Delta_{\neg y}$ improvements, for the lower and upper bound respectively.

D. Proof of Theorem 3.5

Supplementary notations. For the simplicity of presentation, given $i, j, t \in \{0, 1\}$, we write $\mathbb{P}(X = i, Y = j, U = t)$ as $\mathbb{P}(x_i, y_j, u_t)$. Analogously, the expression of interventional probability can be simplified as

$$\mathbb{P}(y_j \mid do(x_i)) := \mathbb{P}(Y = j \mid do(X = i)), \text{ where } i, j \in \{0, 1\}.$$

Hence, the value of average treatment effect (ATE) can be written as

$$\text{ATE} := \mathbb{P}(Y = 1 \mid do(X = 1)) - \mathbb{P}(Y = 1 \mid do(X = 0)) = \mathbb{P}(y_1 \mid do(x_1)) - \mathbb{P}(y_1 \mid do(x_0)).$$

After preparation, we begin our proof with the following lemma about the valid identification region. We will further relax this region to be expressed solely in terms of observed data and demonstrate the tightness of the final bound through direct construction.

Lemma D.1 (Valid identification region of ATE)**.** *A valid identification region of average treatment effect (ATE) is given by*

$$\Big[\min_{t=\{0,1\}} \{\max(\mathcal{S}_{t,1}) - \min(\mathcal{S}_{t,0})\}, \max_{t=\{0,1\}} \{\min(\mathcal{S}_{t,1}) - \max(\mathcal{S}_{t,0})\}\Big].$$

Here

$$\mathcal{S}_{t,i} = \left\{ \frac{\mathbb{P}(x_i, y_1)}{\mathbb{P}(u_t, x_i)}\mathbb{P}(u_t), \frac{\mathbb{P}(x_i, y_1) - \mathbb{P}(u_t, x_i)}{\mathbb{P}(x_i) - \mathbb{P}(u_t, x_i)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) \right\}, \textit{ where } t, i \in \{0, 1\}. \tag{41}$$

Proof of Lemma D.1 We first consider the lower bound. If we apply Lemma B.1 on both $\mathbb{P}(y_1 \mid do(x_1))$ and $\mathbb{P}(y_1 \mid do(x_0))$, respectively, then it directly leads

$$\begin{aligned} \mathbb{P}(y_1 \mid do(x_i)) &\in \Big[\min_{t=\{0,1\}} \big\{\max(\mathcal{S}_{t,i})\big\}, \max_{t=\{0,1\}} \big\{\min(\mathcal{S}_{t,i})\big\}\Big], i \in [0, 1]. \\ \text{ATE} &\in \Big[\min_{t=\{0,1\}} \big\{\max(\mathcal{S}_{t,1})\big\} - \max_{t=\{0,1\}} \big\{\min(\mathcal{S}_{t,0})\big\}, \max_{t=\{0,1\}} \big\{\min(\mathcal{S}_{t,1})\big\} - \min_{t=\{0,1\}} \big\{\max(\mathcal{S}_{t,0})\big\}\Big]. \end{aligned} \tag{42}$$

We now prove the following property:

$$\max\{\mathcal{S}_{0,i}\} \leq \max\{\mathcal{S}_{1,i}\} \text{ IFF } \min\{\mathcal{S}_{0,\neg i}\} \geq \min\{\mathcal{S}_{1,\neg i}\}, i \in \{0, 1\}. \tag{43}$$

To prove this, we first apply the following transformation of $\mathcal{S}_{t,i}$ (see Lemma M.1 for its justification):

$$\mathcal{S}_{t,i} = \Big\{\mathbb{P}(y_1 \mid do(x_i)) + \Big[\frac{1}{\mathbb{P}(x_i \mid u_t)} - \frac{1}{\mathbb{P}(x_i \mid u_{\neg t})}\Big]p : p \in \{\mathbb{P}(x_i, y_1, u_{\neg t}), \mathbb{P}(x_i, y_0, u_t)\}\Big\}. \tag{44}$$

Since the elements p within the set are all non-negative, we have that

$$\max\{\mathcal{S}_{0,i}\} \leq \max\{\mathcal{S}_{1,i}\} \text{ IFF } \frac{1}{\mathbb{P}(x_i \mid u_1)} - \frac{1}{\mathbb{P}(x_i \mid u_0)} \geq 0 \text{ IFF } \frac{1}{\mathbb{P}(x_{\neg i} \mid u_1)} - \frac{1}{\mathbb{P}(x_{\neg i} \mid u_0)} \leq 0 \text{ IFF } \min\{\mathcal{S}_{0,\neg i}\} \geq \min\{\mathcal{S}_{1,\neg i}\},$$

which proves (43).

In light of (43) and (42), we can control the lower and upper bound via

$$\text{ATE} \geq \min\Big\{\max\{\mathcal{S}_{0,1}\}, \max\{\mathcal{S}_{1,1}\}\Big\} - \max\Big\{\min\{\mathcal{S}_{0,0}\}, \min\{\mathcal{S}_{1,0}\}\Big\} \overset{*}{=} \min_{t=0,1}\Big\{\max(\mathcal{S}_{t,1}) - \min(\mathcal{S}_{t,0})\Big\}. \tag{45}$$

$$\text{ATE} \leq \max\Big\{\min\{\mathcal{S}_{0,1}\}, \min\{\mathcal{S}_{1,1}\}\Big\} - \min\Big\{\max\{\mathcal{S}_{0,0}\}, \max\{\mathcal{S}_{1,0}\}\Big\} \overset{*}{=} \max_{t=\{0,1\}} \{\min(\mathcal{S}_{t,1}) - \max(\mathcal{S}_{t,0})\}. \tag{46}$$

Here $*$ is according to (43). From above, we finish the proof of Lemma D.1.

■

The proof of Theorem 3.5 **(VALIDITY)** We denote $\mathcal{S}_{t,i} = \{s_t(x_i), s'_t(x_i)\}, i, t \in \{0, 1\}$. We start our proof with the validity part. Firstly, we consider the lower bound. We take advantage of Lemma D.1, whose result is re-stated as follows:

$$\begin{aligned} \text{ATE} &\geq \min_{t=0,1}\Big\{\max(\mathcal{S}_{t,1}) - \min(\mathcal{S}_{t,0})\Big\} \\ &= \min_{t=0,1}\Big\{\max\big\{s_t(x_1), s'_t(x_1)\big\} - \min\big\{s_t(x_0), s'_t(x_0)\big\}\Big\} \\ &\geq \min_{t=0,1}\Big\{\max\big\{\underbrace{s_t(x_1) - s_t(x_0)}_{\Omega_1(t)}, \underbrace{s_t(x_1) - s'_t(x_0)}_{\Omega_2(t)}, \underbrace{s'_t(x_1) - s'_t(x_0)}_{\Omega_3(t)}\big\}\Big\}. \end{aligned} \tag{47}$$

From above, in order to prove the validity of the lower bound, we just need to prove that for each fixed $t \in \{0,1\}$,

$$\max_{i\in\{1,2,3\}} \Omega_i(t) \geq -\mathcal{B}(\mathbb{P}(u_{\neg t}); 0, 1) \tag{48}$$

for any choice of $\mathbb{P}(u_{\neg t}) \in (0,1)$. Again we prove that the above inequality hold when $\mathbb{P}(u_{\neg t})$ belongs to the intervals $\mathcal{I}_1 := \big(0, \mathbb{P}(x_0, y_0)\big]$, $\mathcal{I}_2 := \big(\mathbb{P}(x_0, y_0), \mathbb{P}(x_0, y_1) + \mathbb{P}(y_0)\big]$, and $\mathcal{I}_3 := (\mathbb{P}(x_0, y_1) + \mathbb{P}(y_0), 1)$ respectively.

CASE I: $\mathbb{P}(u_{\neg t}) \in \mathcal{I}_1$. We just need to prove that $\Omega_1(t) \geq -\mathcal{B}(\mathbb{P}(u_{\neg t}); 0, 1)$. notice that

$$\mathbb{P}(u_{\neg t}, x_1) \leq \mathbb{P}(u_{\neg t}) \leq \mathbb{P}(x_0, y_0),$$

then

$$\begin{aligned} s_t(x_1) &= \frac{\mathbb{P}(x_1, y_1)}{\mathbb{P}(u_t, x_1)}\mathbb{P}(u_t) \geq \max\{\mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_t), \mathbb{P}(x_1, y_1)\} \geq \mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_t), \\ s_t(x_0) &= \frac{\mathbb{P}(x_0, y_1)}{\mathbb{P}(u_t, x_0)}\mathbb{P}(u_t) = \frac{\mathbb{P}(x_0, y_1)}{\mathbb{P}(x_0) - \mathbb{P}(u_{\neg t}, x_0)}\mathbb{P}(u_t) \leq \frac{\mathbb{P}(x_0, y_1)}{\mathbb{P}(x_0) - \mathbb{P}(u_{\neg t})}\mathbb{P}(u_t), \end{aligned} \tag{49}$$

which proves the desired result.

CASE II: $\mathbb{P}(u_{\neg t}) \in \mathcal{I}_2$. We just need to prove that $\Omega_2(t) \geq -\mathcal{B}(\mathbb{P}(u_{\neg t}); 0, 1)$. Notice that

$$s'_t(x_0) = \frac{\mathbb{P}(x_0, y_1) - \mathbb{P}(u_t, x_0)}{\mathbb{P}(x_0) - \mathbb{P}(u_t, x_0)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) = \frac{-\mathbb{P}(x_0, y_0)}{\mathbb{P}(u_{\neg t}, x_0)}\mathbb{P}(u_{\neg t}) + 1 \leq 1 - \mathbb{P}(x_0, y_0), \tag{50}$$

Moreover, due to $[\mathbb{P}(x_0, y_1) - \mathbb{P}(u_t, x_0)]\mathbb{P}(x_0) \leq \mathbb{P}(x_0, y_1)[\mathbb{P}(x_0) - \mathbb{P}(u_t, x_0)]$, we have

$$s'_t(x_0) = \frac{\mathbb{P}(x_0, y_1) - \mathbb{P}(u_t, x_0)}{\mathbb{P}(x_0) - \mathbb{P}(u_t, x_0)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) \leq \frac{\mathbb{P}(x_0, y_1)}{\mathbb{P}(x_0)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t). \tag{51}$$

Combining with (49), (50) and (51) yields

$$\begin{aligned} s_t(x_1) - s'_t(x_0) &\geq \max\Big\{\mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_t), \mathbb{P}(x_1, y_1)\Big\} - \min\Big\{1 - \mathbb{P}(x_0, y_0), \mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t)\Big\}. \\ &\overset{*}{=} \begin{cases} \mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_t) + \mathbb{P}(x_0, y_0) - 1 & \mathbb{P}(u_{\neg t}) \in [\mathbb{P}(x_0, y_0), \mathbb{P}(x_0)] \\ \mathbb{P}(x_1, y_1) - \mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_{\neg t}) - \mathbb{P}(u_t) & \mathbb{P}(u_{\neg t}) \in [\mathbb{P}(x_0), \mathbb{P}(x_0, y_1) + \mathbb{P}(y_0)]. \end{cases} \end{aligned} \tag{52}$$

Here $*$ is due to

$$\mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_t) \geq \mathbb{P}(x_1, y_1) \text{ and } 1 - \mathbb{P}(x_0, y_0) \leq \mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) \text{ when } \mathbb{P}(u_t) \geq \mathbb{P}(x_1).$$

CASE III: $\mathbb{P}(u_{\neg t}) \in \mathcal{I}_3$, we prove (48) by showing that $\Omega_3(t) \geq -\mathcal{B}(\mathbb{P}(u_{\neg t}); 0, 1)$. Notice that

$$\begin{aligned} s'_t(x_1) &= \frac{\mathbb{P}(x_1, y_1) - \mathbb{P}(u_t, x_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_t, x_1)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) \geq \frac{\mathbb{P}(x_1, y_1) - \mathbb{P}(u_t)}{\mathbb{P}(x_1) - \mathbb{P}(u_t)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t). \\ s'_t(x_0) &= \frac{\mathbb{P}(x_0, y_1) - \mathbb{P}(u_t, x_0)}{\mathbb{P}(x_0) - \mathbb{P}(u_t, x_0)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) \leq \mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t). \end{aligned} \tag{53}$$

Combining with **CASEs I-III**, the lower bound (LHS) of the validity part has been proved.

Secondly, we consider the upper bound of ATE. According to Lemma D.1, we already have

$$\begin{aligned} \text{ATE} &\leq \max_{t=\{0,1\}} \{\min(\mathcal{S}_{t,1}) - \max(\mathcal{S}_{t,0})\} \\ &= \max_{t=0,1} \Big\{ \min\big\{s_t(x_1), s'_t(x_1)\big\} - \max\big\{s_t(x_0), s'_t(x_0)\big\}\Big\} \\ &\leq \max_{t=0,1} \Big\{ \min\big\{ \underbrace{s_t(x_1) - s_t(x_0)}_{\Phi_1(t)}, \underbrace{s'_t(x_1) - s_t(x_0)}_{\Phi_2(t)}, \underbrace{s'_t(x_1) - s'_t(x_0)}_{\Phi_3(t)} \big\}\Big\}. \end{aligned} \tag{54}$$

In order to prove (54), it is sufficient to prove that for any $t \in \{0,1\}$,

$$\min_{i\in\{1,2,3\}} \Phi_i(t) \leq \mathcal{B}(\mathbb{P}(u_{\neg t}); 1, 1) \tag{55}$$

for any choice of $\mathbb{P}(u_{\neg t}) \in (0,1)$. Again we consider the scenarios when $\mathbb{P}(u_{\neg t})$ belongs to $\mathcal{I}_1' = (0, \mathbb{P}(x_1, y_0)]$, $\mathcal{I}_2' = (\mathbb{P}(x_1, y_0), 1 - \mathbb{P}(x_0, y_1)]$ and $\mathcal{I}_3' = (1 - \mathbb{P}(x_0, y_1), 1)$.

CASE I: $\mathbb{P}(u_{\neg t}) \in \mathcal{I}_1'$. We prove (55) via showing that $\Phi_1(t) \leq \mathcal{B}(\mathbb{P}(u_{\neg t}); 1, 1)$. We have

$$\begin{aligned} s_t(x_1) &= \frac{\mathbb{P}(x_1, y_1)}{\mathbb{P}(u_1, x_1)}\mathbb{P}(u_t) = \frac{\mathbb{P}(x_1, y_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_{\neg t}, x_1)}\mathbb{P}(u_t) \leq \frac{\mathbb{P}(x_1, y_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_{\neg t})}\mathbb{P}(u_t). \\ s_t(x_0) &= \frac{\mathbb{P}(x_0, y_1)}{\mathbb{P}(u_t, x_0)}\mathbb{P}(u_t) \geq \max\{\mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_t), \mathbb{P}(x_0, y_1)\}. \end{aligned} \tag{56}$$

CASE II: $\mathbb{P}(u_{\neg t}) \in \mathcal{I}_2'$. We prove that $\Phi_2(t) \leq \mathcal{B}(\mathbb{P}(u_{\neg t}); 1, 1)$. Notice that

$$s_t'(x_1) = \frac{\mathbb{P}(x_1, y_1) - \mathbb{P}(u_t, x_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_t, x_1)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) = \frac{-\mathbb{P}(x_1, y_0)}{\mathbb{P}(u_{\neg t}, x_1)}\mathbb{P}(u_{\neg t}) + 1 \leq 1 - \mathbb{P}(x_1, y_0). \tag{57}$$

Moreover, due to $[\mathbb{P}(x_1, y_1) - \mathbb{P}(u_t, x_1)]\mathbb{P}(x_1) \leq \mathbb{P}(x_1, y_1)[\mathbb{P}(x_1) - \mathbb{P}(u_i, x_1)]$, we have

$$s_t'(x_1) = \frac{\mathbb{P}(x_1, y_1) - \mathbb{P}(u_t, x_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_t, x_1)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) \leq \frac{\mathbb{P}(x_1, y_1)}{\mathbb{P}(x_1)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t). \tag{58}$$

Combined with (56), (57) and (58), we have that

$$\begin{aligned} s_t'(x_1) - s_t(x_0) &\leq \min\Big\{1 - \mathbb{P}(x_1, y_0), \mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t)\Big\} - \max\Big\{\mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_t), \mathbb{P}(x_0, y_1)\Big\}. \\ &\overset{*}{=} \begin{cases} -\mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_t) - \mathbb{P}(x_1, y_0) + 1 & \mathbb{P}(u_{\neg t}) \in [\mathbb{P}(x_1, y_0), \mathbb{P}(x_1)] \\ -\mathbb{P}(x_0, y_1) + \mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) & \mathbb{P}(u_{\neg t}) \in [\mathbb{P}(x_1), 1 - \mathbb{P}(x_0, y_1)]. \end{cases} \end{aligned} \tag{59}$$

Here $*$ is due to

$$1 - \mathbb{P}(x_1, y_0) \leq \mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) \text{ and } \mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_t) \geq \mathbb{P}(x_0, y_1) \text{ iff } \mathbb{P}(u_{\neg t}) \leq \mathbb{P}(x_1).$$

CASE III: $\mathbb{P}(u_{\neg t}) \in \mathcal{I}_3'$, we prove that $\Phi_3(t) \leq \mathcal{B}(\mathbb{P}(u_{\neg t}); 1, 1)$. This is due to

$$\begin{aligned} s_t'(x_1) =& \frac{\mathbb{P}(x_1, y_1) - \mathbb{P}(u_t, x_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_t, x_1)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) \leq \frac{\mathbb{P}(x, y)}{\mathbb{P}(x)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t). \\ s_t'(x_0) =& \frac{\mathbb{P}(x_0, y_1) - \mathbb{P}(u_t, x_0)}{\mathbb{P}(x_0) - \mathbb{P}(u_t, x_0)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t) \geq \frac{\mathbb{P}(x_0, y_1) - \mathbb{P}(u_t)}{\mathbb{P}(x_0) - \mathbb{P}(u_t)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t). \end{aligned} \tag{60}$$

CASE I-III simultaneously lead to (55). Hence the upper bound (RHS) of the validity part has been proved.

Combining both our control of lower and upper bounds, we obtain the validity of the bound described in Theorem 3.5.

(TIGHTNESS) Our tightness proof contains two steps: First, we prove that given any $\mathbb{P}(X, Y), \mathbb{P}(U)$, there exist two joint distributions $\mathbb{P}(Y, X, U)$ such that their corresponding ATE's equal to the lower bound $\min_{t\in\{0,1\}}\{-\mathcal{B}(\mathbb{P}(U=t); 0, 1)\}$ and the upper bound $\max_{t\in\{0,1\}} \mathcal{B}(\mathbb{P}(U=t); 1, 1)$. Secondly, we further demonstrate that for all o' between these two bounds, there exists at least one compatible $\mathbb{P}(X, Y, U)$ with corresponding ATE equal to o'.

To prove the first step, we start by proving the tightness of the lower bound $\min_{t\in\{0,1\}}\{-\mathcal{B}(\mathbb{P}(U=t); 0, 1)\}$. Due to the symmetry between $\mathbb{P}(u_0)$ and $\mathbb{P}(u_1)$, we only need consider the case $\mathbb{P}(u_0) \in \mathcal{I}_i, i = 1, 2, 3$.

CASE I : $\mathbb{P}(u_0) \in \mathcal{I}_1 = (0, \mathbb{P}(x_0, y_0)]$, the following construction is compatible:

$$\mathbb{P}(u_0, x_1) = 0, \mathbb{P}(u_1, x_1) = \mathbb{P}(x_1), \mathbb{P}(u_0, x_0) = \mathbb{P}(u_0), \mathbb{P}(u_1, x_0) = \mathbb{P}(x_0) - \mathbb{P}(u_0).$$

On this basis, the conditional probabilities can be constructed as

$$\mathbb{P}(y_1 \mid u_0, x_1) = 0, \mathbb{P}(y_1 \mid u_1, x_1) = \mathbb{P}(y_1 \mid x_1), \mathbb{P}(y_1 \mid u_0, x_0) = 0, \mathbb{P}(y_1 \mid u_1, x_0) = \frac{\mathbb{P}(x_0, y_1)}{\mathbb{P}(x_0) - \mathbb{P}(u_0)}.$$

Then ATE can be computed as

$$\text{ATE} = 0 * \mathbb{P}(u_0) + \mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_1) - 0 * \mathbb{P}(u_0) - \frac{\mathbb{P}(x_0, y_1)}{\mathbb{P}(x_0) - \mathbb{P}(u_0)}\mathbb{P}(u_1) = \left[\mathbb{P}(y_1 \mid x_1) - \frac{\mathbb{P}(x_0, y_1)}{\mathbb{P}(x_0) - \mathbb{P}(u_0)}\right]\mathbb{P}(u_1).$$

CASE II: $\mathbb{P}(u_0) \in \mathcal{I}_2 = (\mathbb{P}(x_0, y_0), \mathbb{P}(x_0, y_1) + \mathbb{P}(y_0)]$, we separate the construction on $\mathcal{I}_2$ into two parts according to (52). $\forall \mathbb{P}(u_0) \in (\mathbb{P}(x_0, y_0), \mathbb{P}(x_0)]$, the following construction is compatible:

$$\mathbb{P}(u_0, x_1) = 0, \mathbb{P}(u_1, x_1) = \mathbb{P}(x_1), \mathbb{P}(u_0, x_0) = \mathbb{P}(u_0), \mathbb{P}(u_1, x_0) = \mathbb{P}(x_0) - \mathbb{P}(u_0).$$

On this basis, the conditional probability is constructed as

$$\mathbb{P}(y_1 \mid u_0, x_1) = 0, \mathbb{P}(y_1 \mid u_1, x_1) = \mathbb{P}(y_1 \mid x_1), \mathbb{P}(y_1 \mid u_0, x_0) = \frac{\mathbb{P}(u_0) - \mathbb{P}(x_0, y_0)}{\mathbb{P}(u_0)}, \mathbb{P}(y_1 \mid u_1, x_0) = 1.$$

Then ATE can be computed as

$$\text{ATE} = 0 * \mathbb{P}(u_0) + \mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_1) - \frac{\mathbb{P}(u_0) - \mathbb{P}(x_0, y_0)}{\mathbb{P}(u_0)} * \mathbb{P}(u_0) - 1 * \mathbb{P}(u_1) = \mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_1) + \mathbb{P}(x_0, y_0) - 1.$$

Moreover, $\forall \mathbb{P}(u_0) \in (\mathbb{P}(x_0), \mathbb{P}(x_0, y_1) + \mathbb{P}(y_0)]$, the following construction is compatible:

$$\mathbb{P}(u_0, x_1) = \mathbb{P}(x_1) - \mathbb{P}(u_1), \mathbb{P}(u_1, x_1) = \mathbb{P}(u_1), \mathbb{P}(u_0, x_0) = \mathbb{P}(x_0), \mathbb{P}(u_1, x_0) = 0.$$

On this basis, the conditional probability is constructed as

$$\mathbb{P}(y_1 \mid u_0, x_1) = 0, \mathbb{P}(y_1 \mid u_1, x_1) = \frac{\mathbb{P}(x_1, y_1)}{\mathbb{P}(u_1)}, \mathbb{P}(y_1 \mid u_0, x_0) = \mathbb{P}(y_1 \mid x_0), \mathbb{P}(y_1 \mid u_1, x_0) = 1.$$

Then ATE can be computed as

$$\text{ATE} = 0 * \mathbb{P}(u_0) + \frac{\mathbb{P}(x_1, y_1)}{\mathbb{P}(u_1)}\mathbb{P}(u_1) - \mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_0) - 1 * \mathbb{P}(u_1) = \mathbb{P}(x_1, y_1) - \mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_0) - \mathbb{P}(u_1).$$

CASE III: $\mathbb{P}(u_0) \in \mathcal{I}_3 = (\mathbb{P}(x_0, y_1) + \mathbb{P}(y_0), 1)$, we have $\mathbb{P}(u_1) < \mathbb{P}(x_1, y_1)$. The following construction is compatible:

$$\mathbb{P}(u_0, x_1) = \mathbb{P}(x_1) - \mathbb{P}(u_1), \mathbb{P}(u_1, x_1) = \mathbb{P}(u_1), \mathbb{P}(u_0, x_0) = \mathbb{P}(x_0), \mathbb{P}(u_1, x_0) = 0.$$

On this basis, the conditional probability is constructed as

$$\mathbb{P}(y_1 \mid u_0, x_1) = \frac{\mathbb{P}(x_1, y_1) - \mathbb{P}(u_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_1)}, \mathbb{P}(y_1 \mid u_1, x_1) = 1, \mathbb{P}(y_1 \mid u_0, x_0) = \mathbb{P}(y_1 \mid x_0), \mathbb{P}(y_1 \mid u_1, x_0) = 1.$$

Then ATE can be computed as

$$\text{ATE} = \frac{\mathbb{P}(x_1, y_1) - \mathbb{P}(u_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_1)} * \mathbb{P}(u_0) + 1 * \mathbb{P}(u_1) - \mathbb{P}(y_1 \mid x_0) * \mathbb{P}(u_0) - 1 * \mathbb{P}(u_1) = \left[\frac{\mathbb{P}(x_1, y_1) - \mathbb{P}(u_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_1)} - \mathbb{P}(y_1 \mid x_0)\right]\mathbb{P}(u_0).$$

In sum, via direct construction, we have proved $\min_{t \in \{0,1\}}\{-\mathcal{B}(\mathbb{P}(U = t); 0, 1)\}$ can be achieved given any $\mathbb{P}(X, Y), \mathbb{P}(U)$. We now consider how to achieve $\max_{t \in \{0,1\}} \mathcal{B}(\mathbb{P}(U = t); 1, 1)$. Again, we only need to consider $\mathbb{P}(u_0) \in \mathcal{I}'_i, i = 1, 2, 3$.

CASE I: $\mathbb{P}(u_0) \in \mathcal{I}_1' = (0, \mathbb{P}(x_1, y_0)]$, the following construction is compatible:

$$\mathbb{P}(u_0, x_1) = \mathbb{P}(u_0), \mathbb{P}(u_1, x_1) = \mathbb{P}(x_1) - \mathbb{P}(u_0), \mathbb{P}(u_0, x_0) = 0, \mathbb{P}(u_1, x_0) = \mathbb{P}(x_0).$$

On this basis, the conditional probability is constructed as

$$\mathbb{P}(y_1 \mid u_0, x_1) = 0, \mathbb{P}(y_1 \mid u_1, x_1) = \frac{\mathbb{P}(x_1, y_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_0)}, \mathbb{P}(y_1 \mid u_0, x_0) = 0, \mathbb{P}(y_1 \mid u_1, x_0) = \mathbb{P}(y_1 \mid x_0).$$

Then ATE can be computed as

$$\text{ATE} = 0 * \mathbb{P}(u_0) + \frac{\mathbb{P}(x_1, y_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_0)} * \mathbb{P}(u_1) - 0 * \mathbb{P}(u_0) - \mathbb{P}(y_1 \mid x_0) * \mathbb{P}(u_1) = \left[\frac{\mathbb{P}(x_1, y_1)}{\mathbb{P}(x_1) - \mathbb{P}(u_0)} - \mathbb{P}(y_1 \mid x_0)\right] \mathbb{P}(u_1).$$

CASE II: $\mathbb{P}(u_0) \in \mathcal{I}_2' = (\mathbb{P}(x_1, y_0), 1 - \mathbb{P}(x_0, y_1)]$. We partition $\mathcal{I}_2'$ into two parts according to (59). For the first part, $\forall \mathbb{P}(u_0) \in (\mathbb{P}(x_1, y_0), \mathbb{P}(x_1)]$, the following construction is compatible:

$$\mathbb{P}(u_0, x_1) = \mathbb{P}(u_0), \mathbb{P}(u_1, x_1) = \mathbb{P}(x_1) - \mathbb{P}(u_0), \mathbb{P}(u_0, x_0) = 0, \mathbb{P}(u_1, x_0) = \mathbb{P}(x_0).$$

On this basis, the conditional probability is constructed as

$$\mathbb{P}(y_1 \mid u_0, x_1) = \frac{\mathbb{P}(u_0) - \mathbb{P}(x_1, y_0)}{\mathbb{P}(u_0)}, \mathbb{P}(y_1 \mid u_1, x_1) = 1, \mathbb{P}(y_1 \mid u_0, x_0) = 0, \mathbb{P}(y_1 \mid u_1, x_0) = \mathbb{P}(y_1 \mid x_0).$$

Then ATE can be computed as

$$\text{ATE} = \frac{\mathbb{P}(u_0) - \mathbb{P}(x_1, y_0)}{\mathbb{P}(u_0)} * \mathbb{P}(u_0) + 1 * \mathbb{P}(u_1) - 0 * \mathbb{P}(u_0) - \mathbb{P}(y_1 \mid x_0) * \mathbb{P}(u_1) = 1 - \mathbb{P}(x_1, y_0) - \mathbb{P}(y_1 \mid x_0)\mathbb{P}(u_1).$$

Moreover, for the second part, $\forall \mathbb{P}(u_0) \in (\mathbb{P}(x_1), 1 - \mathbb{P}(x_0, y_1)]$, the following construction is compatible:

$$\mathbb{P}(u_0, x_1) = \mathbb{P}(x_1), \mathbb{P}(u_1, x_1) = 0, \mathbb{P}(u_0, x_0) = \mathbb{P}(x_0) - \mathbb{P}(u_1), \mathbb{P}(u_1, x_0) = \mathbb{P}(u_1).$$

On this basis, the conditional probability is constructed as

$$\mathbb{P}(y_1 \mid u_0, x_1) = \mathbb{P}(y_1 \mid x_1), \mathbb{P}(y_1 \mid u_1, x_1) = 1, \mathbb{P}(y_1 \mid u_0, x_0) = 0, \mathbb{P}(y_1 \mid u_1, x_0) = \frac{\mathbb{P}(x_0, y_1)}{\mathbb{P}(u_1)}.$$

Then ATE can be computed as

$$\text{ATE} = \mathbb{P}(y_1 \mid x_1) * \mathbb{P}(u_0) + 1 * \mathbb{P}(u_1) - 0 * \mathbb{P}(u_0) - \frac{\mathbb{P}(x_0, y_1)}{\mathbb{P}(u_1)} * \mathbb{P}(u_1) = \mathbb{P}(y_1 \mid x_1)\mathbb{P}(u_0) + \mathbb{P}(u_1) - \mathbb{P}(x_0, y_1).$$

CASE III: $\mathbb{P}(u_0) \in \mathcal{I}_3' = (1 - \mathbb{P}(x_0, y_1), 1)$, we have $\mathbb{P}(u_1) < \mathbb{P}(x_0, y_1)$. The following construction is compatible:

$$\mathbb{P}(u_0, x_1) = \mathbb{P}(x_1), \mathbb{P}(u_1, x_1) = 0, \mathbb{P}(u_0, x_0) = \mathbb{P}(x_0) - \mathbb{P}(u_1), \mathbb{P}(u_1, x_0) = \mathbb{P}(u_1).$$

On this basis, the conditional probability is constructed as

$$\mathbb{P}(y_1 \mid u_0, x_1) = \mathbb{P}(y_1 \mid x_1), \mathbb{P}(y_1 \mid u_1, x_1) = 1, \mathbb{P}(y_1 \mid u_0, x_0) = \frac{\mathbb{P}(x_0, y_1) - \mathbb{P}(u_1)}{\mathbb{P}(x_0) - \mathbb{P}(u_1)}, \mathbb{P}(y_1 \mid u_1, x_0) = 1.$$

Then ATE can be computed as

$$\text{ATE} = \mathbb{P}(y_1 \mid x_1) * \mathbb{P}(u_0) + 1 * \mathbb{P}(u_1) - \frac{\mathbb{P}(x_0, y_1) - \mathbb{P}(u_1)}{\mathbb{P}(x_0) - \mathbb{P}(u_1)} * \mathbb{P}(u_0) - 1 * \mathbb{P}(u_1) = \left[\mathbb{P}(y_1 \mid x_1) - \frac{\mathbb{P}(x_0, y_1) - \mathbb{P}(u_1)}{\mathbb{P}(x_0) - \mathbb{P}(u_1)}\right] \mathbb{P}(u_0).$$

In sum, we have proved $\max_{t \in \{0,1\}} \mathcal{B}(\mathbb{P}(U = t); 1, 1)$ can be achieved given any $\mathbb{P}(X, Y)$ and $\mathbb{P}(U)$.

Now we have demonstrated that for every given specification of $\mathbb{P}(X, Y)$ and $\mathbb{P}(U)$, there exists a compatible joint distribution $\mathbb{P}(X, Y, U)$, whose induced ATE could be equivalent to the lower bound $-\min_{t\in\{0,1\}} \mathcal{B}(\mathbb{P}(U = t); 0, 1)$ and upper bound $\max_{t\in\{0,1\}} \mathcal{B}(\mathbb{P}(U = t); 1, 1)$. We are now left with illustrating that for each o' between these two bounds, there exists a compatible $\mathbb{P}(X, Y, U)$ whose corresponding ATE is equal to o'.

We first consider the case $\mathbb{P}(u_0)$ or $\mathbb{P}(u_1)$ is equal to $\mathbb{P}(x_1)$. Without loss of generality, we just consider the case $\mathbb{P}(u_0) = \mathbb{P}(x_1)$. In this case, our proposed identification region is $[-\mathbb{P}(x_1, y_0) - \mathbb{P}(x_0, y_1), \mathbb{P}(x_0, y_0) + \mathbb{P}(x_1, y_1)]$.

Then we construct

$$\mathbb{P}(u_0, x_1) = \mathbb{P}(x_1), \mathbb{P}(u_1, x_1) = \mathbb{P}(u_0, x_0) = 0, \mathbb{P}(u_1, x_0) = \mathbb{P}(x_0).$$

Moreover, we set the conditional probability $\mathbb{P}(y_1 \mid u_0, x_1) = \mathbb{P}(y_1 \mid x_1)$, $\mathbb{P}(y_1 \mid u_1, x_1) = \varepsilon_1$, $\mathbb{P}(y_1 \mid u_0, x_0) = \varepsilon_2$, $\mathbb{P}(y_1 \mid u_1, x_0) = \mathbb{P}(y_1 \mid x_0)$ and $\mathbb{P}(y_0 \mid u', x') = 1 - \mathbb{P}(y_1 \mid u', x'), u', x' \in \{0, 1\}$. Here all $\varepsilon_1, \varepsilon_2 \in [0, 1]$. Apparently, this construction is non-negative and compatible with the observed marginal distributions $\mathbb{P}(X, Y)$ and $\mathbb{P}(U)$. Under this construction, we get

$$\text{ATE} = \mathbb{P}(y_1 \mid x_1)\mathbb{P}(x_1) + \varepsilon_1\mathbb{P}(x_0) - \varepsilon_2\mathbb{P}(x_1) - \mathbb{P}(y_1 \mid x_0)\mathbb{P}(x_0). \tag{61}$$

One can arbitrarily select points $(\varepsilon_1, \varepsilon_2)$ on the plane $\mathbb{R}^2 : [0, 1] \times [0, 1]$. By varying $(\varepsilon_1, \varepsilon_2)$ along $\varepsilon_1 + \varepsilon_2 = 1$ from $(1, 0)$ to $(0, 1)$, all values within our proposed identification region $[-\mathbb{P}(x_1, y_0) - \mathbb{P}(x_0, y_1), \mathbb{P}(x_0, y_0) + \mathbb{P}(x_1, y_1)]$ is achievable, which proves our desired result.

Below we consider the more general case where $\mathbb{P}(u_0), \mathbb{P}(u_1) \neq \mathbb{P}(x_1)$. Given any fixed $\varepsilon > 0$, let

$$\mathcal{B}_\varepsilon(t; x, y) := \begin{cases} \left(-\frac{\mathbb{P}(\neg x, y)}{\mathbb{P}(\neg x)-\varepsilon} + \frac{\mathbb{P}(x,y)}{\mathbb{P}(x)-t+\varepsilon}\right)(1-t) & t \in (0, \mathbb{P}(x, \neg y)] \\ \frac{t-\mathbb{P}(x,\neg y)}{t-\varepsilon} t + \left(\frac{\mathbb{P}(x)-t}{\mathbb{P}(x)-t+\varepsilon} - \frac{\mathbb{P}(\neg x,y)}{\mathbb{P}(\neg x)-\varepsilon}\right)(1-t) & t \in (\mathbb{P}(x, \neg y), \mathbb{P}(x)] \\ \left(\frac{\mathbb{P}(x,y)-\varepsilon}{\mathbb{P}(x)-\varepsilon} - \frac{\varepsilon}{\varepsilon-\mathbb{P}(x)+t}\right)t + \frac{1-t-\mathbb{P}(\neg x,y)}{1-t-\varepsilon}(1-t) & t \in (\mathbb{P}(x), 1-\mathbb{P}(\neg x, y)] \\ \left(\frac{\mathbb{P}(x,y)-\varepsilon}{\mathbb{P}(x)-\varepsilon} - \frac{\mathbb{P}(\neg x,y)-(1-t)+\varepsilon}{\mathbb{P}(\neg x)-(1-t)+\varepsilon}\right)t & t \in (1-\mathbb{P}(\neg x, y), 1) \end{cases} \tag{62}$$

It is easy to verify that $\forall x', y' \in \{0, 1\}, t \in \{\mathbb{P}(u_0), \mathbb{P}(u_1)\}, \mathcal{B}_\varepsilon(t; x', y')$ converges to $\mathcal{B}(t; x', y')$, as $\varepsilon \to 0$. Notice that since $t \notin \{\mathbb{P}(x_0), \mathbb{P}(x_1)\}$, the denominators in $\mathcal{B}_\varepsilon(t; x', y')$ would not approach to zero as $\varepsilon \to 0$.

From this, in order to make sure that for each $o' \in \left(\min_{t\in\{0,1\}} -\mathcal{B}(\mathbb{P}(U = t); 0, 1), \max_{t\in\{0,1\}} \mathcal{B}(\mathbb{P}(U = t); 1, 1)\right)$, there is a legitimate $\mathbb{P}(X, Y, U)$ whose induced value of ATE is equal to o', it is sufficient to demonstrate there exists a sufficiently small $\varepsilon_0 > 0$ such that $\forall\varepsilon \in (0, \varepsilon_0]$,

$$\left[\min_{t\in\{0,1\}} -\mathcal{B}_\varepsilon(\mathbb{P}(U = t); 0, 1), \max_{t\in\{0,1\}} \mathcal{B}_\varepsilon(\mathbb{P}(U = t); 1, 1)\right] \tag{63}$$

is a subset of the identification region. This is because once this is proved, we can further conclude that for any $o' \in \left(\min_{t\in\{0,1\}} -\mathcal{B}(\mathbb{P}(U = t); 0, 1), \max_{t\in\{0,1\}} \mathcal{B}(\mathbb{P}(U = t); 1, 1)\right)$, there exists a ε so that o' lies in the region defined by (63).

Now to prove (63) is a subset of the identification region, we now consider an auxiliary region

$$\mathcal{O}'_\varepsilon = \{\mathbb{P}(y_1 \mid do(x_1)) - \mathbb{P}(y_1 \mid do(x_0)) : \mathbb{P}(u', x') \geq \varepsilon, u', x' \in \{0, 1\} \;\&\; \mathbb{P}(Y, U, X) \text{ is compatible with } \mathbb{P}(X, Y), \mathbb{P}(U)\},$$

which is apparently a subset of the identification region.

Analogous to the above analysis in the proof of Theorem 3.3, if we treat $\mathbb{P}(x', y', u'), (x', y', u' \in \{0, 1\})$ as parameters and ATE as a function of these parameters, it can be verified that the parameter space restricted by $\mathcal{O}'_\varepsilon$ is a convex and compact set; moreover, since the denominator $\mathbb{P}(u', x'), u', x' \in \{0, 1\}$ is larger than ε in the restricted parameter space $\mathcal{O}'_\varepsilon$, ATE is a well-defined and bounded continuous function w.r.t all parameters. In light of these, $\mathcal{O}'_\varepsilon$ is a closed interval on $\mathbb{R}$. Letting $o'_{\text{min}}, o'_{\text{max}}$ be the left and right side of interval $\mathcal{O}'_\varepsilon$; then one can easily verify that with ε_0 sufficiently small, for all $\varepsilon \in (0, \varepsilon_0]$,

$$o'_{\text{min}} \leq \min_{t\in\{0,1\}} -\mathcal{B}_\varepsilon(\mathbb{P}(U = t); 0, 1) \leq \max_{t\in\{0,1\}} \mathcal{B}_\varepsilon(\mathbb{P}(U = t); 1, 1) \leq o'_{\text{max}},$$

which means the region given by (63) serves as a sub-region of $\mathcal{O}'_\varepsilon$. Since $\mathcal{O}'_\varepsilon$ is a subset of the identification region; it concludes that the interval (63) is a subset of the identification region as well. It completes the proof. ■

D.1. Further discussion: The identification of the vanilla bound of ATE

We first consider the vanilla lower bound. For the necessity part, under Assumption 3.1, we have

$$\begin{aligned}\text{ATE} =&\mathbb{P}(y_1 \mid do(x_1)) - \mathbb{P}(y_1 \mid do(x_0))\\ =&\mathbb{P}(y_1 \mid u_0, x_1)\mathbb{P}(u_0) + \mathbb{P}(y_1 \mid u_1, x_1)\mathbb{P}(u_1) - \mathbb{P}(y_1 \mid u_0, x_0)\mathbb{P}(u_0) - \mathbb{P}(y_1 \mid u_1, x_0)\mathbb{P}(u_1)\\ =&\mathbb{P}(x_1, y_1) + \sum_{i=0,1}\mathbb{P}(y_1 \mid u_i, x_1)\mathbb{P}(u_i, x_0) - \mathbb{P}(x_0, y_1) - \sum_{i=0,1}\mathbb{P}(y_1 \mid u_i, x_0)\mathbb{P}(u_i, x_1).\end{aligned} \tag{64}$$

When (64) $= -\mathbb{P}(\neg x, y) - \mathbb{P}(x, \neg y)$, it is equivalent to

$$\mathbb{P}(x_1) + \sum_{i=0,1}\mathbb{P}(y_1 \mid u_i, x_1)\mathbb{P}(u_i, x_0) - \sum_{i=0,1}\mathbb{P}(y_1 \mid u_i, x_0)\mathbb{P}(u_i, x_1) = 0. \tag{65}$$

(65) is equal to

$$\sum_{i=0,1}\mathbb{P}(y_1 \mid u_i, x_1)\mathbb{P}(u_i, x_0) + \sum_{i=0,1}\mathbb{P}(y_0 \mid u_i, x_0)\mathbb{P}(u_i, x_1) = 0. \tag{66}$$

Notice that

$$\begin{aligned}\text{LHS of (66)} \geq& \max\Big\{\sum_{i=0,1}\mathbb{P}(y_1, u_i, x_1)\mathbb{P}(u_i, x_0), \sum_{i=0,1}\mathbb{P}(y_0, u_i, x_0)\mathbb{P}(u_i, x_1)\Big\}\\ \geq& \max\Big\{\min\Big\{\mathbb{P}(u_0, x_0), \mathbb{P}(u_1, x_0)\Big\}\mathbb{P}(x_1, y_1), \min\Big\{\mathbb{P}(u_0, x_1), \mathbb{P}(u_1, x_1)\Big\}\mathbb{P}(x_0, y_0)\Big\} \geq 0.\end{aligned} \tag{67}$$

Under the Assumption 3.1, combined with (66) and (67), when LHS of (66) achieves 0, then it must be $\mathbb{P}(u_t, x_0) = \mathbb{P}(u_{\neg t}, x_1) = 0, \exists t \in \{0, 1\}$. Therefore, the necessary condition of the vanilla lower bound of ATE can be derived:

$$\mathbb{P}(u_t) = \mathbb{P}(u_t, x_1) + \mathbb{P}(u_t, x_0) = \mathbb{P}(u_t, x_1) + \mathbb{P}(u_{\neg t}, x_1) = \mathbb{P}(x_1), \exists t \in \{0, 1\}. \tag{68}$$

On the other hand, we consider when ATE achieves the vanilla upper bound, namely (64) $= \mathbb{P}(x_1, y_1) + \mathbb{P}(x_0, y_0)$. It is equivalent to

$$\sum_{i=0,1}\mathbb{P}(y_0 \mid u_i, x_1)\mathbb{P}(u_i, x_0) + \sum_{i=0,1}\mathbb{P}(y_1 \mid u_i, x_0)\mathbb{P}(u_i, x_1) = 0. \tag{69}$$

Analogously, it leads to

$$\begin{aligned}\text{LHS of (69)} \geq& \max\Big\{\sum_{i=0,1}\mathbb{P}(y_0, u_i, x_1)\mathbb{P}(u_i, x_0), \sum_{i=0,1}\mathbb{P}(y_1, u_i, x_0)\mathbb{P}(u_i, x_1)\Big\}\\ \geq& \max\Big\{\min\left\{\mathbb{P}(u_0, x_0), \mathbb{P}(u_1, x_0)\right\}\mathbb{P}(x_1, y_0), \min\left\{\mathbb{P}(u_0, x_1), \mathbb{P}(u_1, x_1)\right\}\mathbb{P}(x_0, y_1)\Big\} \geq 0.\end{aligned} \tag{70}$$

Under Assumption 3.1, when LHS of (70) achieves 0, then it also must be $\mathbb{P}(u_t, x_0) = \mathbb{P}(u_{\neg t}, x_1) = 0, \exists t \in \{0, 1\}$. Hence we repeat (68) and also have $\{\mathbb{P}(u_0), \mathbb{P}(u_1)\} \cap \{\mathbb{P}(x)\} \neq 0$. In sum, the necessity part has been proved.

For the sufficiency part, we resort to the construction in (61). Hence the IFF condition has been demonstrated.

■

E. The proof of Theorem 4.1

Notice that

$$\mathbb{P}(y \mid do(x)) - \mathbb{P}(x, y) = \sum_{u=0}^{d_u - 1}\mathbb{P}(y \mid u, x)\mathbb{P}(u, \neg x) \tag{71}$$

and

$$\mathbb{P}(x,y)+\mathbb{P}(\neg x)-\mathbb{P}(y\mid do(x))=\sum_{u=0}^{d_u-1}\mathbb{P}(\neg y\mid u,x)\mathbb{P}(u,\neg x). \tag{72}$$

(NECESSITY) We first consider the vanilla lower bound. If $\mathbb{P}(y\mid do(x))=\mathbb{P}(x,y)$, it induces that (71) $=0$. Hence, $\exists\mathcal{U}\subseteq\mathbb{R}$ such that $\forall u\in\mathcal{U}$, $\mathbb{P}(u,\neg x)=0$, and $\forall u\in\mathcal{U}^c$, $\mathbb{P}(y,u,x)=0$. According to this partition, the subset sum $\mathbb{P}(U\in\mathcal{U})$ could be bounded:

$$\begin{aligned}&\mathbb{P}(U\in\mathcal{U})=\mathbb{P}(U\in\mathcal{U},x)+\mathbb{P}(U\in\mathcal{U},\neg x)=\mathbb{P}(U\in\mathcal{U},x)+0\leq\mathbb{P}(x).\\&\mathbb{P}(U\in\mathcal{U})\geq\mathbb{P}(U\in\mathcal{U},x,y)+0=\mathbb{P}(U\in\mathcal{U},x,y)+\mathbb{P}(U\in\mathcal{U}^c,x,y)=\mathbb{P}(x,y).\end{aligned} \tag{73}$$

Analogously, for the upper bound, if $\mathbb{P}(y\mid do(x))=\mathbb{P}(x,y)+\mathbb{P}(\neg x)$, then $\exists\mathcal{U}\in\mathbb{R}$ such that $\forall U\in\mathcal{U}$, $\mathbb{P}(u,\neg x)=0$, and $\forall U\in\mathcal{U}^c$, $\mathbb{P}(\neg y,u,x)=0$. Hence

$$\begin{aligned}&\mathbb{P}(U\in\mathcal{U})=\mathbb{P}(U\in\mathcal{U},x)+\mathbb{P}(U\in\mathcal{U},\neg x)=\mathbb{P}(U\in\mathcal{U},x)+0\leq\mathbb{P}(x).\\&\mathbb{P}(U\in\mathcal{U})\geq\mathbb{P}(U\in\mathcal{U},x,\neg y)+0=\mathbb{P}(U\in\mathcal{U},x,\neg y)+\mathbb{P}(U\in\mathcal{U}^c,x,\neg y)=\mathbb{P}(x,\neg y).\end{aligned} \tag{74}$$

This proves the necessity part.

(SUFFICIENCY) For the vanilla lower bound, we take the following construction of the joint distribution $\mathbb{P}(U,X)$:

$$\begin{bmatrix}\mathbb{P}(U\in\mathcal{U},x) & \mathbb{P}(U\in\mathcal{U}^c,x)\\ \mathbb{P}(U\in\mathcal{U},\neg x) & \mathbb{P}(U\in\mathcal{U}^c,\neg x)\end{bmatrix}=\begin{bmatrix}\mathbb{P}(U\in\mathcal{U}) & \mathbb{P}(x)-\mathbb{P}(U\in\mathcal{U}^c)\\ 0 & \mathbb{P}(\neg x)\end{bmatrix}. \tag{75}$$

Notice that RHS of (75) is constructed by observed data. Moreover, the conditional probability $\mathbb{P}(Y\mid U,X)$ is constructed as

$$\begin{aligned}&\forall u\in\mathcal{U},\mathbb{P}(y\mid u,x)=\mathbb{P}(x,y)/\mathbb{P}(U\in\mathcal{U}),\mathbb{P}(y\mid u,\neg x)=0;\\&\forall u\in\mathcal{U}^c,\mathbb{P}(y\mid u,x)=0,\mathbb{P}(y\mid u,\neg x)=\mathbb{P}(y\mid\neg x).\end{aligned} \tag{76}$$

We choose $\mathbb{P}(\neg y\mid u,x')=1-\mathbb{P}(y\mid u,x'),\forall u\in\{0,1,...d_u-1\},x'\in\{0,1\}$. For a complete visualization, the total construction is summarized as the following Table 2. Noteworthy, each term among each summation in Table 2 could be chosen as arbitrary non-negative numbers. The non-negativity and compatibility of the construction (75) and (76) are easily verified.

$\mathcal{A}$	$\mathbb{P}(y,u\in\mathcal{A},x)$	$\mathbb{P}(\neg y,u\in\mathcal{A},x)$	$\mathbb{P}(y,u\in\mathcal{A},\neg x)$	$\mathbb{P}(\neg y,u\in\mathcal{A},\neg x)$
$\mathcal{U}$	$\mathbb{P}(x,y)$	$\mathbb{P}(U\in\mathcal{U})-\mathbb{P}(x,y)$	0	0
$\mathcal{U}^c$	0	$\mathbb{P}(x)-\mathbb{P}(U\in\mathcal{U})$	$\mathbb{P}(\neg x,y)$	$\mathbb{P}(\neg x,\neg y)$

TABLE 2: *The construction of the vanilla lower bound of* $\mathbb{P}(y\mid do(x))$.

According to the fact that $\forall u\in\mathcal{U},\mathbb{P}(u,\neg x)=0$. $\forall u\in\mathcal{U}^c,\mathbb{P}(y\mid u,x)=0$, (71) can be transformed as

$$\mathbb{P}(y\mid do(x))=\mathbb{P}(x,y)+\sum_{U\in\mathcal{U}}\mathbb{P}(y\mid u,x)\mathbb{P}(u,\neg x)+\sum_{U\in\mathcal{U}^c}\mathbb{P}(y\mid u,x)\mathbb{P}(u,\neg x)=\mathbb{P}(x,y).$$

For the vanilla upper bound, we inherit the construction of $\mathbb{P}(U,X)$ in (75), and then establish the new conditional probability:

$$\begin{aligned}&\forall u\in\mathcal{U},\mathbb{P}(y\mid u,x)=1-\mathbb{P}(x,\neg y)/\mathbb{P}(u\in\mathcal{U}),\mathbb{P}(y\mid u,\neg x)=0;\\&\forall u\in\mathcal{U}^c,\mathbb{P}(y\mid u,x)=1,\mathbb{P}(y\mid u,\neg x)=\mathbb{P}(y\mid\neg x).\end{aligned} \tag{77}$$

Analogously, the total construction is summarized as the following Table (3).

According to the fact that $\forall u\in\mathcal{U},\mathbb{P}(u,\neg x)=0$. $\forall U\in\mathcal{U}^c,\mathbb{P}(\neg y\mid u,x)=0$. According to (72), we have

$$\mathbb{P}(y\mid do(x)=\mathbb{P}(x,y)+\mathbb{P}(\neg x)-\sum_{U\in\mathcal{U}}\mathbb{P}(\neg y\mid u,x)\mathbb{P}(u,\neg x)-\sum_{U\in\mathcal{U}^c}\mathbb{P}(\neg y\mid u,x)\mathbb{P}(u,\neg x)=\mathbb{P}(x,y)+\mathbb{P}(\neg x).$$

$\mathcal{A}$	$\mathbb{P}(y, u \in \mathcal{A}, x)$	$\mathbb{P}(\neg y, u \in \mathcal{A}, x)$	$\mathbb{P}(y, u \in \mathcal{A}, \neg x)$	$\mathbb{P}(\neg y, u \in \mathcal{A}, \neg x)$
$\mathcal{U}$	$\mathbb{P}(U \in \mathcal{U}) - \mathbb{P}(x, \neg y)$	$\mathbb{P}(x, \neg y)$	0	0
$\mathcal{U}^c$	$\mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U})$	0	$\mathbb{P}(\neg x, y)$	$\mathbb{P}(\neg x, \neg y)$

TABLE 3: *The construction of the vanilla upper bound of* $\mathbb{P}(y \mid do(x))$.

Until here the sufficiency part has also been proved. Combining with the necessity part and the sufficiency part, the desired result follows.

■

F. The proof of Theorem 4.2

For brevity, we still follow the supplementary notations in Appendix D and adopt $\mathrm{ATE}^{\mathrm{L}}_{\mathrm{vanilla}}$ and $\mathrm{ATE}^{\mathrm{U}}_{\mathrm{vanilla}}$ to denote the vanilla lower and upper bound of ATE, i.e., $\mathrm{ATE}^{\mathrm{L}}_{\mathrm{vanilla}} = -\mathbb{P}(x_1, y_0) - \mathbb{P}(x_0, y_1)$, $\mathrm{ATE}^{\mathrm{U}}_{\mathrm{vanilla}} = \mathbb{P}(x_1, y_1) + \mathbb{P}(x_0, y_0)$. According to (71) and (72), we have

$$\mathrm{ATE} - \mathrm{ATE}^{\mathrm{L}}_{\mathrm{vanilla}} = \sum_{u=0}^{d_u-1} \big[\mathbb{P}(y_1 \mid u, x_1)\mathbb{P}(u, x_0) + \mathbb{P}(y_0 \mid u, x_0)\mathbb{P}(u, x_1)\big]. \tag{78}$$

$$\mathrm{ATE}^{\mathrm{U}}_{\mathrm{vanilla}} - \mathrm{ATE} = \sum_{u=0}^{d_u-1} \big[\mathbb{P}(y_1 \mid u, x_0)\mathbb{P}(u, x_1) + \mathbb{P}(y_0 \mid u, x_1)\mathbb{P}(u, x_0)\big]. \tag{79}$$

(NECESSITY) We first consider the vanilla lower bound. If we have (78) $= 0$, then $\exists \mathcal{R}_0 \subseteq \mathbb{R}$, such that $\forall U \in \mathcal{R}_0$, we have $\mathbb{P}(u, x_1) = 0, \forall u \in \mathcal{R}_0^c$, we have $\mathbb{P}(y_0, u, x_0) = 0$. For the same reason, $\exists \mathcal{R}_1 \subseteq \mathbb{R}$, such that $\forall U \in \mathcal{R}_1$, we have $\mathbb{P}(u, x_0) = 0, \forall u \in \mathcal{R}_1^c$, we have $\mathbb{P}(y_1, u, x_1) = 0$. These properties are summarized as

$$\mathbb{P}(U \in \mathcal{R}_1, x_0) = \mathbb{P}(U \in \mathcal{R}_0, x_1) = \mathbb{P}(y_1, U \in \mathcal{R}_1^c, x_1) = \mathbb{P}(y_0, U \in \mathcal{R}_0^c, x_0) = 0. \tag{80}$$

Apparently, we have $\mathcal{R}_0 \cap \mathcal{R}_1 \subseteq \{u : \mathbb{P}(U = u) = 0\}$. On this basis, we construct the desired pair $\{\mathcal{U}_0, \mathcal{U}_1\}$ via truncating joint parts of $\{\mathcal{R}_0, \mathcal{R}_1\}$:

$$\mathcal{U}_0 := \mathcal{R}_0/(\mathcal{R}_0 \cap \mathcal{R}_1), \mathcal{U}_1 := \mathcal{R}_1/(\mathcal{R}_0 \cap \mathcal{R}_1), \mathcal{U}_0 \cap \mathcal{U}_1 = \emptyset. \tag{81}$$

Recalling the strategy in (73) and (74), we take advantage of (80) and achieve the following bounds:

$$\begin{aligned}
&\mathbb{P}(U \in \mathcal{U}_1) = \mathbb{P}(U \in \mathcal{R}_1) = \mathbb{P}(U \in \mathcal{R}_1, x_1) \leq \mathbb{P}(x_1),\\
&\mathbb{P}(U \in \mathcal{U}_0) = \mathbb{P}(U \in \mathcal{R}_0) = \mathbb{P}(U \in \mathcal{R}_0, x_0) \leq \mathbb{P}(x_0).\\
&\mathbb{P}(U \in \mathcal{U}_1) = \mathbb{P}(U \in \mathcal{R}_1) \geq \mathbb{P}(U \in \mathcal{R}_1, x_1, y_1) = \mathbb{P}(U \in \mathcal{R}_1, x_1, y_1) + \mathbb{P}(U \in \mathcal{R}_1^c, x_1, y_1) = \mathbb{P}(x_1, y_1).\\
&\mathbb{P}(U \in \mathcal{U}_0) = \mathbb{P}(U \in \mathcal{R}_0) \geq \mathbb{P}(U \in \mathcal{R}_0, x_0, y_0) = \mathbb{P}(U \in \mathcal{R}_0, x_0, y_0) + \mathbb{P}(U \in \mathcal{R}_0^c, x_0, y_0) = \mathbb{P}(x_0, y_0).
\end{aligned} \tag{82}$$

Hence it holds that $\mathbb{P}(U \in \mathcal{U}_1) \in [\mathbb{P}(x_1, y_1), \mathbb{P}(x_1)] = \mathcal{I}_{1,1}$ and $\mathbb{P}(U \in \mathcal{U}_0) \in [\mathbb{P}(x_0, y_0), \mathbb{P}(x_0)] = \mathcal{I}_{0,0}$. The necessity part of the vanilla lower bound has been implied.

On the other hand, we consider the vanilla upper bound. Compared (79) with (78), it just need to exchange the symbols $\{x_0, x_1\}$ with each other. On this basis, it implies that $\exists \mathcal{Q}_0, \mathcal{Q}_1 \subseteq \mathbb{R}$ such that

$$\mathbb{P}(U \in \mathcal{Q}_1, x_1) = \mathbb{P}(U \in \mathcal{Q}_0, x_0) = \mathbb{P}(y_1, U \in \mathcal{Q}_1^c, x_0) = \mathbb{P}(y_0, U \in \mathcal{Q}_0^c, x_1) = 0. \tag{83}$$

Then we choose

$$\mathcal{U}_0 := \mathcal{Q}_0/(\mathcal{Q}_0 \cap \mathcal{Q}_1), \mathcal{U}_1 := \mathcal{Q}_1/(\mathcal{Q}_0 \cap \mathcal{Q}_1), \mathcal{U}_0 \cap \mathcal{U}_1 \neq \emptyset. \tag{84}$$

With the same strategy, we aim to bound the summation $\mathbb{P}(U \in \mathcal{U}_i), i = 0, 1$. We get

$$\begin{aligned}
&\mathbb{P}(U \in \mathcal{U}_1) = \mathbb{P}(U \in \mathcal{Q}_1) = \mathbb{P}(U \in \mathcal{Q}_1, x_0) \leq \mathbb{P}(x_0),\\
&\mathbb{P}(U \in \mathcal{U}_0) = \mathbb{P}(U \in \mathcal{Q}_0) = \mathbb{P}(U \in \mathcal{Q}_0, x_1) \leq \mathbb{P}(x_1).\\
&\mathbb{P}(U \in \mathcal{U}_1) = \mathbb{P}(U \in \mathcal{Q}_1) \geq \mathbb{P}(U \in \mathcal{Q}_1, x_0, y_1) = \mathbb{P}(U \in \mathcal{Q}_1, x_0, y_1) + \mathbb{P}(U \in \mathcal{Q}_1^c, x_0, y_1) = \mathbb{P}(x_0, y_1).\\
&\mathbb{P}(U \in \mathcal{U}_0) = \mathbb{P}(U \in \mathcal{Q}_0) \geq \mathbb{P}(U \in \mathcal{Q}_0, x_1, y_0) = \mathbb{P}(U \in \mathcal{Q}_0, x_1, y_0) + \mathbb{P}(U \in \mathcal{Q}_0^c, x_1, y_0) = \mathbb{P}(x_1, y_0).
\end{aligned} \tag{85}$$

Hence we get $\mathbb{P}(U \in \mathcal{U}_1) \in [\mathbb{P}(x_0, y_1), \mathbb{P}(x_0)] = \mathcal{I}_{0,1}$ and $\mathbb{P}(U \in \mathcal{U}_0) \in [\mathbb{P}(x_1, y_0), \mathbb{P}(x_1)] = \mathcal{I}_{1,0}$.

In conclusion, the necessity part has been demonstrated.

(SUFFICIENCY) We first consider the vanilla lower bound with the following construction:

$$\begin{bmatrix} \mathbb{P}(U \in \mathcal{U}_0, x_1) & \mathbb{P}(U \in \mathcal{U}_1, x_1) & \mathbb{P}(U \in (\mathcal{U}_0 \cup \mathcal{U}_1)^c, x_1) \\ \mathbb{P}(U \in \mathcal{U}_0, x_0) & \mathbb{P}(U \in \mathcal{U}_1, x_0) & \mathbb{P}(U \in (\mathcal{U}_0 \cup \mathcal{U}_1)^c, x_0) \end{bmatrix} = \begin{bmatrix} 0 & \mathbb{P}(U \in \mathcal{U}_1) & \mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U}_1) \\ \mathbb{P}(U \in \mathcal{U}_0) & 0 & \mathbb{P}(\neg x) - \mathbb{P}(U \in \mathcal{U}_0) \end{bmatrix}. \tag{86}$$

Moreover, the conditional probability $\mathbb{P}(Y \mid U, X)$ is constructed by

$$\begin{aligned} &\forall u \in \mathcal{U}_0, \mathbb{P}(y_1 \mid u, x_1) = 0, \mathbb{P}(y_1 \mid u, x_0) = 1 - \mathbb{P}(x_0, y_0)/\mathbb{P}(U \in \mathcal{U}_0); \\ &\forall u \in \mathcal{U}_1, \mathbb{P}(y_1 \mid u, x_1) = \mathbb{P}(x_1, y_1)/\mathbb{P}(U \in \mathcal{U}_1), \mathbb{P}(y_1 \mid u, x_0) = 1; \\ &\forall u \in (\mathcal{U}_0 \cup \mathcal{U}_1)^c, \mathbb{P}(y_1 \mid u, x_1) = 0, \mathbb{P}(y_1 \mid u, x_0) = 1. \end{aligned} \tag{87}$$

we also choose $\mathbb{P}(y_0 \mid u_i, x') = 1 - \mathbb{P}(y_1 \mid u_i, x')$. Here $u \in \{0, 1, ...d_u - 1\}$, $x' \in \{0, 1\}$. For better visualization, the whole construction can be expanded in the following Table (4) (with $\mathbb{P}(Y, U, X)$ as the parameter).

$\mathcal{A}$	$\mathbb{P}(y_1, u \in \mathcal{A}, x_1)$	$\mathbb{P}(y_0, u \in \mathcal{A}, x_1)$	$\mathbb{P}(y_1, u \in \mathcal{A}, x_0)$	$\mathbb{P}(y_0, u \in \mathcal{A}.x_0)$
$\mathcal{U}_0$	0	0	$\mathbb{P}(U \in \mathcal{U}_0) - \mathbb{P}(x_0, y_0)$	$\mathbb{P}(x_0, y_0)$
$\mathcal{U}_1$	$\mathbb{P}(x_1, y_1)$	$\mathbb{P}(U \in \mathcal{U}_1) - \mathbb{P}(x_1, y_1)$	0	0
$(\mathcal{U}_0 \cup \mathcal{U}_1)^c$	0	$\mathbb{P}(x_1) - \mathbb{P}(U \in \mathcal{U}_1)$	$\mathbb{P}(x_0) - \mathbb{P}(U \in \mathcal{U}_0)$	0

TABLE 4: *The construction of the vanilla lower bound of ATE.*

According to the fact $\mathbb{P}(U \in \mathcal{U}_1) \in [\mathbb{P}(x_1, y_1), \mathbb{P}(x_1)] = \mathcal{I}_{1,1}$ and $\mathbb{P}(U \in \mathcal{U}_0) \in [\mathbb{P}(x_0, y_0), \mathbb{P}(x_0)] = \mathcal{I}_{0,0}$, all the elements in Table 4 is non-negative. We compute $\mathbb{P}(y_1 \mid do(x_1))$ via dividing the summation into three groups $\mathcal{U}_0, \mathcal{U}_1, (\mathcal{U}_0 \cup \mathcal{U}_1)^c$ as follows:

$$\mathbb{P}(y_1 \mid do(x_1)) = \mathbb{P}(x_1, y_1) + \sum_{\mathcal{A}=\mathcal{U}_1, \mathcal{U}_1^c} \sum_{u \in \mathcal{A}} \mathbb{P}(y_1 \mid u, x_1)\mathbb{P}(u, x_0) \overset{(a)}{=} \mathbb{P}(x_1, y_1). \tag{88}$$

The last operation (a) is due to $\forall U \in \mathcal{U}_1$, $\mathbb{P}(u, x_0) = 0$, $\forall u \in (\mathcal{U}_1)^c$, $\mathbb{P}(y_1 \mid u, x_1) = 0$. On the other hand,

$$\mathbb{P}(y_1 \mid do(x_0)) = \mathbb{P}(x_0, y_1) + \mathbb{P}(x_1) - \sum_{\mathcal{A}=\mathcal{U}_0, \mathcal{U}_0^c} \sum_{u \in \mathcal{A}} \mathbb{P}(y_0 \mid u, x_0)\mathbb{P}(u, x_1) \overset{(b)}{=} \mathbb{P}(x_0, y_1) + \mathbb{P}(x_1). \tag{89}$$

Analogously, the last operation (b) is due to $\forall u \in \mathcal{U}_0$, $\mathbb{P}(u, x_1) = 0$, $\forall u \in (\mathcal{U}_0)^c$, $\mathbb{P}(y_0 \mid u, x_0) = 0$. Hence,

$$\text{ATE} = \mathbb{P}(y_1 \mid do(x_1)) - \mathbb{P}(y_1 \mid do(x_0)) = \mathbb{P}(x_1, y_1) - \mathbb{P}(x_0, y_1) - \mathbb{P}(x_1) = -\mathbb{P}(x_1, y_0) - \mathbb{P}(x_0, y_1). \tag{90}$$

On the other hand, we consider the vanilla upper bound. Analogously, considering the structure of (78)-(79), it only requires that $\{x_0, x_1\}$ exchanges with each other. Inspired by this, we set

$$\begin{bmatrix} \mathbb{P}(U \in \mathcal{U}_0, x_1) & \mathbb{P}(U \in \mathcal{U}_1, x_1) & \mathbb{P}(U \in (\mathcal{U}_0 \cup \mathcal{U}_1)^c, x_1) \\ \mathbb{P}(U \in \mathcal{U}_0, x_0) & \mathbb{P}(U \in \mathcal{U}_1, x_0) & \mathbb{P}(U \in (\mathcal{U}_0 \cup \mathcal{U}_1)^c, x_0) \end{bmatrix} = \begin{bmatrix} \mathbb{P}(U \in \mathcal{U}_0) & 0 & \mathbb{P}(\neg x) - \mathbb{P}(U \in \mathcal{U}_0) \\ 0 & \mathbb{P}(U \in \mathcal{U}_1) & \mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U}_1) \end{bmatrix}. \tag{91}$$

Moreover, we construct the conditional probability:

$$\begin{aligned} &\forall u \in \mathcal{U}_0, \mathbb{P}(y_1 \mid u, x_0) = 0, \mathbb{P}(y_1 \mid u, x_1) = 1 - \mathbb{P}(x_1, y_0)/\mathbb{P}(U \in \mathcal{U}_0); \\ &\forall u \in \mathcal{U}_1, \mathbb{P}(y_1 \mid u, x_0) = \mathbb{P}(x_0, y_1)/\mathbb{P}(u \in \mathcal{U}_1), \mathbb{P}(y_1 \mid u, x_1) = 1; \\ &\forall u \in (\mathcal{U}_0 \cup \mathcal{U}_1)^c, \mathbb{P}(y_1 \mid u, x_0) = 0, \mathbb{P}(y_1 \mid u, x_1) = 1. \end{aligned} \tag{92}$$

Here $\mathbb{P}(y_0 \mid u, x') = 1 - \mathbb{P}(y_1 \mid u, x')$. Here $u \in \{0, 1, ...d_u - 1\}$, $x' \in \{0, 1\}$. The whole construction can be expanded as the following Table (5) to justify the non-negativity and compatibility:

Under the construction in Table 5, we re-compute the $\mathbb{P}(y_1 \mid do(x_1))$ and $\mathbb{P}(y_1 \mid do(x_0))$:

$$\begin{aligned} \mathbb{P}(y_1 \mid do(x_1)) &= \mathbb{P}(x_1, y_1) + \mathbb{P}(x_0) - \sum_{\mathcal{A}=\mathcal{U}_0, (\mathcal{U}_0)^c} \sum_{u \in \mathcal{A}} \mathbb{P}(y_0 \mid u, x_1)\mathbb{P}(u, x_0) \overset{(c)}{=} \mathbb{P}(x_1, y_1) + \mathbb{P}(x_0). \\ \mathbb{P}(y_1 \mid do(x_0)) &= \mathbb{P}(x_0, y_1) + \sum_{\mathcal{A}=\mathcal{U}_1, (\mathcal{U}_1)^c} \sum_{u \in \mathcal{A}} \mathbb{P}(y_1 \mid u, x_0)\mathbb{P}(u, x_1) \overset{(d)}{=} \mathbb{P}(x_0, y_1). \end{aligned} \tag{93}$$

$\mathcal{A}$	$\mathbb{P}(y_1, u \in \mathcal{A}, x_1)$	$\mathbb{P}(y_0, u \in \mathcal{A}, x_1)$	$\mathbb{P}(y_1, u \in \mathcal{A}, x_0)$	$\mathbb{P}(y_0, u \in \mathcal{A}, x_0)$
$\mathcal{U}_0$	$\mathbb{P}(U \in \mathcal{U}_0) - \mathbb{P}(x_1, y_0)$	$\mathbb{P}(x_1, y_0)$	0	0
$\mathcal{U}_1$	0	0	$\mathbb{P}(x_0, y_1)$	$\mathbb{P}(U \in \mathcal{U}_1) - \mathbb{P}(x_0, y_1)$
$(\mathcal{U}_0 \cup \mathcal{U}_1)^c$	$\mathbb{P}(x_1) - \mathbb{P}(U \in \mathcal{U}_0)$	0	0	$\mathbb{P}(x_0) - \mathbb{P}(U \in \mathcal{U}_1)$

TABLE 5: *The construction of the vanilla upper bound of ATE.*

The operation $(c), (d)$ is according to the following fact, respectively:

$$\begin{aligned} &\forall u \in \mathcal{U}_0, \mathbb{P}(u, x_0) = 0, \forall u \in (\mathcal{U}_0)^c, \mathbb{P}(y_0 \mid u, x_1) = 0. \\ &\forall u \in \mathcal{U}_1, \mathbb{P}(u, x_1) = 0, \forall u \in (\mathcal{U}_0)^c, \mathbb{P}(y_1 \mid u, x_0) = 0. \end{aligned} \tag{94}$$

Hence we achieve

$$\text{ATE} = \mathbb{P}(y_1 \mid do(x_1)) - \mathbb{P}(y_1 \mid do(x_0)) = \mathbb{P}(x_1, y_1) + \mathbb{P}(x_0) - \mathbb{P}(x_0, y_1) = \mathbb{P}(x_1, y_1) + \mathbb{P}(x_0, y_0). \tag{95}$$

■

G. The proof of Corollary 4.3

Proof. Without loss of generalization, we let $x = y = 1$ in this proof.

Property (i) Considering $\mathcal{P}$, it is easy to verify

$$\mathcal{P} := \left\{\mathbb{P}(U) : \exists\, \mathcal{U} \subseteq \mathbb{R} \text{ s.t. } \mathbb{P}(U \in \mathcal{U}) \in [\mathbb{P}(x, y_0) \vee \mathbb{P}(x, y_1), \mathbb{P}(x)]\right\}.$$

then for $\mathcal{P}_{\text{ATE}}$, it could be verified that

$$\mathcal{P}'_{\text{ATE}} := \{\mathbb{P}(U) : \exists\, \mathcal{U}_0, \mathcal{U}_1 \subseteq \mathbb{R} \text{ with } \mathcal{U}_0 \cap \mathcal{U}_1 = \emptyset, \text{ s.t. } \forall z \in \{0,1\}, \mathbb{P}(U \in \mathcal{U}_z) \in \mathcal{I}_{z,\neg z} \cap \mathcal{I}_{z,z}\} \subseteq \mathcal{P}_{\text{ATE}}.$$

For both these two cases, for any given $\mathbb{P}(X, Y)$ under Assumption 3.1 and Assumption 3.2, we could choose

$$\mathbb{P}^\star(U) = \begin{cases} \mathbb{P}(x) & U = t_0 \\ \mathbb{P}(\neg x) & U = t_1 \\ 0 & U \neq t_0, t_1 \end{cases} \quad \text{where } t_0, t_1 \in \{0, 1, ...d_u - 1\}, t_0 \neq t_1.$$

It is easy to verify $\mathbb{P}^\star(U) \in \mathcal{P} \cap \mathcal{P}_{\text{ATE}}$. Hence for any given $\mathbb{P}(X, Y)$, legitimate $\mathbb{P}(U)$ exists such that $\mathcal{P} \neq \emptyset$ and $\mathcal{P}_{\text{ATE}} \neq \emptyset$ hold.

Property (ii) We consider the specific construction which is modified from the above:

$$\mathbb{P}^{\star\star}(U) = \begin{cases} \mathbb{P}(x) - \varepsilon & U = t_0 \\ \mathbb{P}(\neg x) + \varepsilon & U = t_1 \\ 0 & U \neq t_0, t_1 \end{cases} \quad \text{where } t_0, t_1 \in \{0, 1, ...d_u - 1\}, t_0 \neq t_1, 0 < \varepsilon < \min\left\{\mathbb{P}(x, y_0), \mathbb{P}(x, y_1), |\mathbb{P}(x) - \mathbb{P}(\neg x)|\right\}.$$

We take the lower bound for instance. Since $\mathbb{P}^{\star\star}(U = t_0) \in [\mathbb{P}(x, y_0) \vee \mathbb{P}(x, y_1), \mathbb{P}(x)]$, we get $\mathbb{P}^{\star\star}(U) \in \mathcal{P} \subseteq \mathcal{P}^L$. Furthermore, it is sufficient to prove $\mathbb{P}^{\star\star}(U) \notin \mathcal{P}^L_{\text{ATE}}$. We make it via contradiction: Recalling the definition, if $\exists \mathcal{U}_0, \mathcal{U}_1$ such that $\exists \mathcal{U}_0 \cap \mathcal{U}_1 = \emptyset$ and $\mathbb{P}(U \in \mathcal{U}_0) \in \mathcal{I}_{0,0}$, $\mathbb{P}(U \in \mathcal{U}_1) \in \mathcal{I}_{1,1}$. According to the fact that $\mathbb{P}(U \in \mathcal{U}_z) > 0, z = 0, 1$, we get $\{t_0, t_1\} \subseteq \mathcal{U}_0 \cup \mathcal{U}_1$, and hence

$$1 = \mathbb{P}(U = t_0) + \mathbb{P}(U = t_1) \leq \mathbb{P}(U \in \mathcal{U}_0) + \mathbb{P}(U \in \mathcal{U}_1) \leq 1.$$

Definitely, it leads to $\mathbb{P}(U \in \mathcal{U}_0) = \mathbb{P}(\neg x)$ and $\mathbb{P}(U \in \mathcal{U}_1) = \mathbb{P}(x)$. Thus we have $\mathcal{U}_z = \{t_z\}, z = 0, 1$. Namely, we have $\mathbb{P}(x) - \varepsilon = \mathbb{P}(\neg x)$, which is equal to $\mathbb{P}(\neg x) + \epsilon = \mathbb{P}(x)$. According to the constraint $\varepsilon < |\mathbb{P}(x) - \mathbb{P}(\neg x)|$ as above, we get the contradiction.

In conclusion, due to $\mathbb{P}^{\star\star}(U) \in \mathcal{P}^L \cap (\mathcal{P}^L_{\text{ATE}})^c$, we have $\mathcal{P}^L_{\text{ATE}} \subsetneqq \mathcal{P}^L$. Totally with the same strategy, we achieve $\mathcal{P}^U_{\text{ATE}} \subsetneqq \mathcal{P}^U$. It leads to $\mathcal{P}_{\text{ATE}} = \mathcal{P}^L_{\text{ATE}} \cap \mathcal{P}^U_{\text{ATE}} \subsetneqq \mathcal{P}^L \cap \mathcal{P}^U = \mathcal{P}$. The desired result follows.

■

H. The proof of Proposition 4.4

Lemma H.1. *Suppose Assumption 3.1-3.2 hold. Given prior knowledge of $\mathbb{P}(U)$, for the interventional probability and ATE, the sufficient conditions for the tight identification regions degenerate to be vanilla are $\mathbb{P}(U)$ belongs to*

$$\mathcal{P}_\forall(\min_{y'\in\{0,1\}} \mathbb{P}(x,y')) \quad \& \quad \mathcal{P}_\forall(\min_{x',y'\in\{0,1\}} \mathbb{P}(x',y')),$$

respectively. Here $\mathcal{P}_\forall(t) := \{\mathbb{P}(U) : \forall i \in \{0,1,...d_u-1\}, \mathbb{P}(U=i) \le t\}$.

The proof of Lemma H.1 is presented as follows.

(INTERVENTIONAL PROBABILITY) We first consider the interventional probability. It is sufficient to prove $\mathcal{P}_\forall(\min_{y'\in\{0,1\}} \mathbb{P}(x,y')) \subseteq \mathcal{P}$. $\forall \mathbb{P}(U) \in \mathcal{P}_\forall(\min_{y'\in\{0,1\}} \mathbb{P}(x,y'))$, we consider the following item:

$$\mathcal{U}^* := \arg\max_{\mathcal{U}} \Big\{\mathbb{P}(U\in\mathcal{U}) : \mathbb{P}(U\in\mathcal{U}) \le \max_{y'\in\{0,1\}} \mathbb{P}(x,y')\Big\}. \tag{96}$$

Apparently, $\mathcal{U}^* \subsetneqq \{0,1,...d_u-1\}$. We consider $\mathbb{P}(\mathcal{U}^* \cup u^*_c)$ with $u^*_c \in (\mathcal{U}^*)^c$.

On the one hand, by definition of $\mathcal{U}^*$, we have $\mathbb{P}(\mathcal{U}^* \cup u^*_c) > \max_{y'\in\{0,1\}} \mathbb{P}(x,y')$; on the other hand, since $\mathbb{P}(U=u^*_c) \le \min_{y'\in\{0,1\}} \mathbb{P}(x,y')$ due to $\mathbb{P}(U) \in \mathcal{P}_\forall(\min_{y'\in\{0,1\}} \mathbb{P}(x,y'))$, we also get

$$\mathbb{P}(\mathcal{U}^* \cup u^*_c) = \mathbb{P}(\mathcal{U}^*) + \mathbb{P}(U=u^*_c) \le \max_{y'\in\{0,1\}} \mathbb{P}(x,y') + \min_{y'\in\{0,1\}} \mathbb{P}(x,y') = \mathbb{P}(x).$$

Hence $\mathbb{P}(\mathcal{U}^* \cup u^*_c) \in [\max_{y'\in\{0,1\}} \mathbb{P}(x,y'), \mathbb{P}(x)]$. Due to the arbitrary of the selection of $\mathbb{P}(U)$ within $\mathcal{P}_\forall(\min_{y'\in\{0,1\}} \mathbb{P}(x,y'))$, it is proved that $\mathcal{P}_\forall(\min_{y'\in\{0,1\}} \mathbb{P}(x,y')) \subseteq \mathcal{P}$.

(ATE) Stepping forwards, we consider the case of ATE. We aim to prove $\mathcal{P}_\forall(\min_{x',y'\in\{0,1\}} \mathbb{P}(x',y')) \subseteq \mathcal{P}_{\text{ATE}}$. Recall that

$$\mathcal{P}_{\text{ATE}} \supseteq \{\mathbb{P}(U) : \exists\, \mathcal{U}_0, \mathcal{U}_1 \subseteq \mathbb{R} \textit{ with } \mathcal{U}_0 \cap \mathcal{U}_1 = \emptyset, \text{ s.t. } \forall z \in \{0,1\}, \mathbb{P}(U\in\mathcal{U}_z) \in \mathcal{I}_{z,0} \cap \mathcal{I}_{z,1}\}. \tag{97}$$

is a subset of $\mathcal{P}_{\text{ATE}}$. Hence it is sufficient to prove $\mathcal{P}_\forall(\min_{x',y'\in\{0,1\}} \mathbb{P}(x',y')) \subseteq$ RHS of (97).

Inspired by the proof of the interventional probability as above, for each legitimate $\mathbb{P}(U)$ in $\mathcal{P}_\forall(\min_{x',y'\in\{0,1\}} \mathbb{P}(x',y'))$, we consider

$$\mathcal{U}^*_0 := \arg\max_{\mathcal{U}} \Big\{\mathbb{P}(U\in\mathcal{U}) : \mathbb{P}(U\in\mathcal{U}) \le \min(\mathcal{I}_{0,0} \cap \mathcal{I}_{0,1}) = \max_{y'\in\{0,1\}} \mathbb{P}(x_0,y')\Big\}. \tag{98}$$

With the same strategy as above, we could bound $\mathbb{P}(U \in \mathcal{U}^*_0 \cup u^*_{0,c})$ with $u^*_{0,c} \in (\mathcal{U}^*_0)^c$:

$$\mathbb{P}(U \in \mathcal{U}^*_0 \cup u^*_{0,c}) \in [\max_{y'\in\{0,1\}} \mathbb{P}(x_0,y'), \max_{y'\in\{0,1\}} \mathbb{P}(x_0,y') + \min_{x',y'\in\{0,1\}} \mathbb{P}(x',y')] \subseteq \mathcal{I}_{0,0} \cap \mathcal{I}_{0,1}. \tag{99}$$

Naturally, we can choose $\mathcal{U}_0 := \mathcal{U}^*_0 \cup u^*_{0,c}$ in Theorem 4.2. Apparently, $(\mathcal{U}_0)^c \neq \emptyset$. Hence we could consider

$$\mathcal{U}^*_1 := \arg\max_{\mathcal{U}} \Big\{\mathbb{P}(U\in\mathcal{U}) : \mathbb{P}(U\in\mathcal{U}) \le \min(\mathcal{I}_{1,0} \cap \mathcal{I}_{1,1}) = \max_{y'\in\{0,1\}} \mathbb{P}(x_1,y'), \mathcal{U} \subseteq (\mathcal{U}_0)^c)\Big\}. \tag{100}$$

Noteworthy, here $\mathcal{U}^*_1 \subsetneqq (\mathcal{U}_0)^c$ because

$$\mathbb{P}(U \in (\mathcal{U}_0)^c) \ge 1 - \max(\mathcal{I}_{0,0} \cap \mathcal{I}_{0,1}) = \mathbb{P}(x_1) > \max_{y'\in\{0,1\}} \mathbb{P}(x_1,y') \ge \mathbb{P}(\mathcal{U}^*_1).$$

Then we could bound $\mathbb{P}(U \in \mathcal{U}_1^* \cup u_{1,c}^*)$ with $u_{1,c}^* \in (\mathcal{U}_1^*)^c \cap (\mathcal{U}_0)^c$:

$$\mathbb{P}(U \in \mathcal{U}_1^* \cup u_{1,c}^*) \in [\max_{y' \in \{0,1\}} \mathbb{P}(x_1, y'), \max_{y' \in \{0,1\}} \mathbb{P}(x_1, y') + \min_{x', y' \in \{0,1\}} \mathbb{P}(x', y')] \subseteq \mathcal{I}_{1,0} \cap \mathcal{I}_{1,1}. \tag{101}$$

Naturally we choose $\mathcal{U}_1 := \mathcal{U}_1^* \cup u_{1,c}^*$ in Theorem 4.2. Notice that

$$\mathcal{U}_1^* \subsetneqq (\mathcal{U}_0)^c \text{ in (100) and } u_{1,c}^* \in (\mathcal{U}_0)^c \text{ by definition,}$$

it leads to $\mathcal{U}_0 \cap \mathcal{U}_1 = \emptyset$. In sum, for each $\mathbb{P}(U)$ arbitrarily selected from $\mathcal{P}_\forall(\min_{x', y' \in \{0,1\}} \mathbb{P}(x', y'))$, there exists disjoint subsets $\mathcal{U}_0, \mathcal{U}_1$ which locate in $\mathcal{I}_{0,0} \cap \mathcal{I}_{0,1}$ and $\mathcal{I}_{1,0} \cap \mathcal{I}_{1,1}$, respectively. Hence such $\mathbb{P}(U)$ belongs to $\mathcal{P}_{\text{ATE}}$. The desired result follows.

■

Equipped with Lemma H.1, we start the proof of Proposition 4.4.

Proof. (CONVERGENCE RATE) We first consider the interventional probability. It is sufficient to prove that the probability of $\mathbb{P}(U)$ falling into $\mathbb{P}_\forall(t)$ is bounded by $d_u(1-t)^{d_u - 1}$. Namely,

$$\mathbb{P}\Big(\mathbb{P}(U) \notin \mathcal{P}\Big) \overset{(1)}{\leq} \mathbb{P}\Big(\mathbb{P}(U) \notin \mathbb{P}_\forall(t)\Big) \overset{(2)}{\leq} d_u(1-t)^{d_u-1} < 1, \text{ where } t = \min_{y' \in \{0,1\}} \mathbb{P}(x, y').$$

Here $\mathbb{P}(U)$ is induced by the parameters $\{\mathbb{P}(U=0), \mathbb{P}(U=1), ...\mathbb{P}(U=d_u-1)\}$ under a uniform prior. The first inequality (1) has already been proved via Lemma H.1. For the second inequality (2), we take advantage of the union bound. We get

$$\mathbb{P}\Big(\mathbb{P}(U) \notin \mathbb{P}_\forall(t)\Big) = \mathbb{P}\Big(\exists i, \mathbb{P}(U=i) > t\Big) \leq \sum_{i=0}^{d_u-1} \Big(\mathbb{P}(U=i) > t\Big) = d_u(1-t)^{d_u-1}. \tag{102}$$

The last inequality is according to under a uniform prior, for all $u = 0, 1, ...d_u - 1$, the marginal cumulative distribution function of $\mathbb{P}(U=u)$ is $F_{\mathbb{P}(U=u)}(x) = (1-x)^{d_u-1}, x \in [0,1)$.

For the ATE case, we only need take $t = \min_{x', y' \in \{0,1\}} \mathbb{P}(x', y')$ and then the whole process holds totally the same. Hence we get

$$\mathbb{P}\Big(\mathbb{P}(U) \notin \mathcal{P}_{\text{ATE}}\Big) \leq \mathbb{P}\Big(\mathbb{P}(U) \notin \mathbb{P}_\forall(t)\Big) \leq d_u(1-t)^{d_u-1} < 1, \text{ where } t = \min_{x', y' \in \{0,1\}} \mathbb{P}(x', y').$$

(MONOTONICITY) Finally, we consider the monotonicity. We first consider the interventional probability case. According to the auxiliary Lemma M.4 in Appendix M, under a uniform prior, the probability of falling into the "vanilla" $\mathbb{P}(\mathbb{P}(U) \in \mathcal{P})$ is equal to

$$\mathbb{P}(\mathcal{S}_{d_u}), \text{ where event } \mathcal{S}_n := \exists \mathcal{A} \in \{0, 1, ...n-1\}, s.t. \sum_{j \in \mathcal{A}} \Big(p_{i(j+1)} - p_{i(j)}\Big) \in \Big[\max_{y' \in \{0,1\}} \{\mathbb{P}(x, y'), \mathbb{P}(x)\}\Big]. \tag{103}$$

Here $\{p_{i(j)}\}_{j=0}^{d_u}$ are re-ordered $\{p_i\}_{i=0}^{d_u}$ satisfying $p_{i(d_u)} \geq p_{i(d_u-1)} \geq ... \geq \mathbb{P}_{i(0)}$, and each original p_i is independently uniformly sampled within the interval $[0,1]$. In order to prove the probability of falling into the "non-vanilla" region $\mathcal{P}^c$ is non-increasing, it is sufficient to demonstrate

$$\mathbb{P}(\mathcal{S}_n) \leq \mathbb{P}(\mathcal{S}_{n+1}), \ \forall n \in \mathbb{N}^+.$$

Notice that

$$\mathbb{P}(\mathcal{S}_{n+1}) = \int_{\alpha \in [0,1]} \mathbb{P}(\mathcal{S}_{n+1} \mid p_{n+1} = \alpha) f_{p_{n+1}}(\alpha) d\alpha = \int_{\alpha \in [0,1]} \mathbb{P}(\mathcal{S}_{n+1} \mid p_{n+1} = \alpha) d\alpha. \tag{104}$$

Here $f_{p_{n+1}}(\alpha)=1, \alpha\in[0,1]$ denotes the uniform distribution of p_{n+1}. Consider each set $\{p_i\}_{i=0}^n \in [0,1]^{n+1}$ as above. If $\mathcal{S}_n$ happens, then apparently, $\mathcal{S}_{n+1}$ must happen with fixed $p_{n+1}=\alpha$. Hence,

$$\mathbb{P}(\mathcal{S}_{n+1} \mid p_{n+1}=\alpha) \geq \mathbb{P}(\mathcal{S}_n), \text{ thus } \mathbb{P}(\mathcal{S}_{n+1}) \geq \int_{\alpha\in[0,1]} \mathbb{P}(\mathcal{S}_n) d\alpha = \mathbb{P}(\mathcal{S}_n), \forall n \in \mathbb{N}^+.$$

It completes the proof on the interventional probability. Furthermore, for the ATE case, it only needs to change the event $\mathcal{S}_n$ to $\mathcal{S}_{n,\mathrm{ATE}}$:

$$\mathcal{S}_{n,\mathrm{ATE}}^L := \exists \mathcal{A}_0, \mathcal{A}_1 \in \{0,1,...n-1\}, \mathcal{A}_0\cap\mathcal{A}_1=\emptyset, s.t. \sum_{j\in\mathcal{A}_0}\Big(p_{i(j+1)}-p_{i(j)}\Big)\in\mathcal{I}_{0,0}, \sum_{j\in\mathcal{A}_1}\Big(p_{i(j+1)}-p_{i(j)}\Big)\in\mathcal{I}_{1,1},$$

$$\mathcal{S}_{n,\mathrm{ATE}}^U := \exists \mathcal{A}_0, \mathcal{A}_1 \in \{0,1,...n-1\}, \mathcal{A}_0\cap\mathcal{A}_1=\emptyset, s.t. \sum_{j\in\mathcal{A}_0}\Big(p_{i(j+1)}-p_{i(j)}\Big)\in\mathcal{I}_{0,1}, \sum_{j\in\mathcal{A}_1}\Big(p_{i(j+1)}-p_{i(j)}\Big)\in\mathcal{I}_{1,0},$$

$$\mathcal{S}_{n,\mathrm{ATE}}^U := \mathcal{S}_{n,\mathrm{ATE}}^L \cap \mathcal{S}_{n,\mathrm{ATE}}^U.$$

The rest analysis holds the same. Namely, with the same strategy, we get

$$\mathbb{P}(\mathcal{S}_{n+1,\mathrm{ATE}}) = \int_{\alpha\in[0,1]} \mathbb{P}(\mathcal{S}_{n+1,\mathrm{ATE}} \mid p_{n+1}=\alpha)d\alpha \geq \int_{\alpha\in[0,1]} \mathbb{P}(\mathcal{S}_{n+1,\mathrm{ATE}})d\alpha = \mathbb{P}(\mathcal{S}_{n,\mathrm{ATE}}).$$

The monotonicity has been proved. ■

I. The proof of Theorem 4.5

Supplementary notation We follow the supplementary notations in Appendix D. Moreover, we use $\mathcal{P}_{XYU}$ to denote the set of all possible $\mathbb{P}(X,Y,U)$ which is compatible with observed data $\mathbb{P}(X,Y), \mathbb{P}(U)$. Naturally, our original optimization problem (1) could be transformed to explore the minimum and maximum of

$$\{\mathbb{P}(y \mid do(x)) : \mathbb{P}(X,Y,U) \in \mathcal{P}_{XYU}\}. \tag{105}$$

Furthermore, we consider the set

$$\mathcal{P}_{XYU}^{(k)} := \Big\{\mathbb{P}(X,Y,U) : \exists \Omega \in \mathbb{R}, |\Omega|=k, s.t. \forall u\in\Omega, \mathbb{P}(y,u,x)\wedge\mathbb{P}(\neg y,u,x)=0\Big\}\bigcap \mathcal{P}_{XYU}.$$

Naturally, it holds that $\mathcal{P}_{XYU}^{(d_u)} \subseteq \mathcal{P}_{XYU}^{(d_u-1)} \subseteq ... \subseteq \mathcal{P}_{XYU}^{(1)} \subseteq \mathcal{P}_{XYU} := \mathcal{P}_{XYU}^{(0)}$. Furthermore, for brevity, we denote the sub identification region of interventional probability:

$$\mathcal{P}_{y|do(x)}^{(k)} := \{\mathbb{P}(y \mid do(x)) : \mathbb{P}(X,Y,U) \in \mathcal{P}_{XYU}^{(k)}\}, k=0,1,...d_u.$$

We now prove our identification bound in Theorem 4.5, namely $[\min \mathcal{P}_{y|do(x)}^{(0)}, \max \mathcal{P}_{y|do(x)}^{(0)}]$ is valid and tight.

(VALIDITY) In order to prove the validity of bounds given by Theorem 4.5, it is sufficient to prove

$$\min \mathcal{P}_{y|do(x)}^{(0)} \geq \mathcal{LB}_{x,y}^{\mathrm{mul}}(\mathbb{P}(U)) \text{ and } \max \mathcal{P}_{y|do(x)}^{(0)} \leq \mathcal{UB}_{x,y}^{\mathrm{mul}}(\mathbb{P}(U)). \tag{106}$$

To achieve this goal, the following two claims should be brought forward:

Claim I: $\mathcal{P}_{XYU}^{d_u-1} \neq \emptyset$.

We prove it via direct construction. For any given $\mathbb{P}(U)$, a legitimate joint distribution $\mathbb{P}(X,Y,U)$ could be constructed which belongs to $\mathcal{P}_{XYU}^{(d_u-1)}$. Details are deferred into Lemma M.2 in Appendix M. Consequently, we get $\mathcal{P}_{XYU}^{k} \neq \emptyset$, where $k=0,1,...d_u-1$.

Claim II: $\min \mathcal{P}_{y|do(x)}^{(d_u-1)} = \min \mathcal{P}_{y|do(x)}^{(0)}, \max \mathcal{P}_{y|do(x)}^{(d_u-1)} = \max \mathcal{P}_{y|do(x)}^{(0)}$.

It means the lower and upper tight identification bounds are equal to the minimum and maximum of $\{\mathbb{P}(y \mid do(x)) : \mathbb{P}(X,Y,U) \in \mathcal{P}_{XYU}^{(d_u-1)}\}$, respectively.

To prove **Claim II**, on the one hand, due to $\mathcal{P}_{XYU}^{(d_u-1)} \subseteq \mathcal{P}_{XYU}$ and **Claim I**, it definitely holds that

$$[\min \mathcal{P}_{y|do(x)}^{(d_u-1)}, \max \mathcal{P}_{y|do(x)}^{(d_u-1)}] \subseteq [\min \mathcal{P}_{y|do(x)}^{(0)}, \min \mathcal{P}_{y|do(x)}^{(0)}]. \tag{107}$$

On the other hand, we consider the series of sub-regions $\{\mathcal{P}_{XYU}^{(k)}\}_{k=1}^{d_u}$ iteratively. For two adjacent sets $\mathcal{P}_{XYU}^{(j)}, \mathcal{P}_{XYU}^{(j+1)}$, $j = 0, 1, ...d_u - 2$. If $\mathcal{P}_{XYU}^{(j)}/\mathcal{P}_{XYU}^{(j+1)} = \emptyset$, then $\mathcal{P}_{XYU}^{(j)} = \mathcal{P}_{XYU}^{(j+1)}$ naturally holds; otherwise, it can be inferred that $\forall \mathbb{P}^{(j)}(X,Y,U) \in \mathcal{P}_{XYU}^{(j)}/\mathcal{P}_{XYU}^{(j+1)}$:

$$\exists u_1^+, u_2^+ \in U, s.t.\ \mathbb{P}^{(j)}(y', u, x) > 0, \text{ where } y' \in \{0,1\}, u = u_1^+, u_2^+. \tag{108}$$

We construct two legitimate $\mathbb{P}_\omega^{(j+1)}(X,Y,U)$ within $\mathcal{P}_{XYU}^{(j+1)}, \omega \in \{1,-1\}$ by perturbing $\mathbb{P}^{(j)}(X,Y,U)$. Here $\mathbb{P}_\omega^{(j+1)}(X,Y,U)$ is established by

$$\mathbb{P}_\omega^{(j+1)}(y \mid u, x') = \begin{cases} (\mathbb{P}^{(j)}(y,u,x') + \omega\eta)/\mathbb{P}^{(j)}(u,x') & u = u_1^+, x' = x \\ (\mathbb{P}^{(j)}(y,u,x') - \omega\eta)/\mathbb{P}^{(j)}(u,x') & u = u_2^+, x' = x \\ \mathbb{P}^{(j)}(y \mid u, x') & \text{otherwise} \end{cases}, \text{ and } \mathbb{P}_\omega^{(j+1)}(u,x') = \mathbb{P}^{(j)}(u,x'), i = 1,2.$$

for all $u \in U$ and $x \in \{0,1\}$. Here $\eta = \min\left\{\mathbb{P}(y',u,x) : y' \in \{0,1\}, u \in \{u_1^+, u_2^+\}\right\} > 0$.

It is easy to verify $\{\mathbb{P}_\omega^{(j+1)}(X,Y,U)\}_{\omega=1,-1} \subseteq \mathcal{P}_{XYU}^{(j+1)}$. Noteworthy, if we abbreviate the interventional probability induced by $\mathbb{P}_\omega^{(j+1)}(X,Y,U), \mathbb{P}^{(j)}(X,Y,U)$ as $\mathbb{P}_\omega^{(j+1)}(y \mid do(x)), \mathbb{P}^{(j)}(y \mid do(x))$, respectively. It holds that

$$\mathbb{P}_\omega^{(j+1)}(y \mid do(x)) = \mathbb{P}^{(j)}(y \mid do(x)) + \left[1/\mathbb{P}_\omega^{(j)}(u_1^+, x) - 1/\mathbb{P}_\omega^{(j)}(u_2^+, x)\right]\omega\eta. \tag{109}$$

Hence

$$\begin{aligned} \mathbb{P}^{(j)}(y \mid do(x)) &\in [\min\{\mathbb{P}_\omega^{(j+1)}(y \mid do(x))\}_{\omega=1,-1}, \max\{\mathbb{P}_\omega^{(j+1)}(y \mid do(x))\}_{\omega=1,-1}] \\ &\in [\min \mathcal{P}_{y|do(x)}^{(j+1)}, \max \mathcal{P}_{y|do(x)}^{(j+1)}]. \end{aligned} \tag{110}$$

Due to the arbitrary selection of $\mathbb{P}^{(j)}(X,Y,U)$, it concludes that

$$[\min \mathcal{P}_{y|do(x)}^{(j+1)}, \max \mathcal{P}_{y|do(x)}^{(j+1)}] \supseteq [\min \mathcal{P}_{y|do(x)}^{(j)}, \min \mathcal{P}_{y|do(x)}^{(j)}], j = 0, 1, ...d_u - 2. \tag{111}$$

The combination of (107) and (111) indicates **Claim II**.

According to **Claim I-II**, in order to prove the validity of bounds given by Theorem 4.5, it is sufficient to prove

$$\min \mathcal{P}_{y|do(x)}^{(d_u-1)} \geq \mathcal{LB}_{x,y}^{\text{mul}}(\mathbb{P}(U)), \max \mathcal{P}_{y|do(x)}^{(d_u-1)} \leq \mathcal{UB}_{x,y}^{\text{mul}}(\mathbb{P}(U)).$$

We first consider the lower bound. Apparently, when $\mathbb{P}(U) \in \mathcal{P}^L$, it leads to $\min \mathcal{P}_{y|do(x)}^{(d_u-1)} \geq \mathbb{P}(x,y) = \mathcal{LB}_{x,y}^{\text{mul}}(\mathbb{P}(U))$. Hence, in the following part, we focus on the non-vanilla case $\mathbb{P}(U) \in (\mathcal{P}^L)^c$. Notice that $\mathcal{P}_{y|do(x)}^{d_u-1}$ could be transformed to the following structure:

$$\mathcal{P}_{y|do(x)}^{d_u-1} = \left\{\mathbb{P}(U \in \mathcal{U}) + \mathbb{P}(y \mid u_t, x)\mathbb{P}(u_t) : \mathcal{U} \subseteq \mathbb{R}/\{t\}, \forall u \in (\mathcal{U} \cup \{t\})^c, \mathbb{P}(y,u,x) = 0, \forall u \in \mathcal{U}, \mathbb{P}(\neg y, u, x) = 0\right\}. \tag{112}$$

It could be verified that for each compatible $\mathcal{U}$ within $\mathcal{P}_{y|do(x)}^{d_u-1}$ under $\mathbb{P}(U) \in (\mathcal{P}^L)^c$, we get

$$\mathbb{P}(U \in \mathcal{U}) \in \left[\max\left\{0, \mathbb{P}(x) - \mathbb{P}(u_t)\right\}, \mathbb{P}(x,y) + \mathbb{P}(\neg x)\right]. \tag{113}$$

We refer readers to Lemma M.5 in Appendix M for the constraints in (113). To prove the validity, it is sufficient to prove each value among (112) locates in $[\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)), \mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U))]$. First, for the valid low bound, we consider the minimum of RHS via separating (112) into **Cases I-II**:

CASE I: under $\mathbb{P}(U) \in (\mathcal{P}^L)^c$, when $\mathbb{P}(U \in \mathcal{U}) \in \Big[\max\big\{0, \mathbb{P}(x) - \mathbb{P}(u_t)\big\}, \mathbb{P}(x,y)\Big] \neq \emptyset$, it holds

$$\mathbb{P}(y, u_t, x) \overset{*}{=} \mathbb{P}(x,y) - \mathbb{P}(y, u \in \mathcal{U}, x) \geq \mathbb{P}(x,y) - \mathbb{P}(u \in \mathcal{U}).$$

Here $*$ is due to the fact $\forall (\mathcal{U} \cup \{t\})^c, \mathbb{P}(y,u,x) = 0$. Then

$$\min \mathcal{P}^{d_u-1}_{y|do(x)} \geq \mathbb{P}(U \in \mathcal{U}) + \frac{\mathbb{P}(x,y) - \mathbb{P}(U \in \mathcal{U})}{\mathbb{P}(x,y) - \mathbb{P}(U \in \mathcal{U}) + \mathbb{P}(x, \neg y, u_t)} \mathbb{P}(u_t) \geq \mathbb{P}(U \in \mathcal{U}) + \frac{\mathbb{P}(x,y) - \mathbb{P}(U \in \mathcal{U})}{\mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U})} \mathbb{P}(u_t). \tag{114}$$

CASE II: under $\mathbb{P}(U) \in (\mathcal{P}^L)^c$, when $\mathbb{P}(U \in \mathcal{U}) \in \Big[\mathbb{P}(x,y), \mathbb{P}(x,y) + \mathbb{P}(\neg x)\Big]$, it holds

$$\min \mathcal{P}^{d_u-1}_{y|do(x)} \geq \mathbb{P}(U \in \mathcal{U}) \geq \min\Big\{\mathbb{P}(U \in \mathcal{U}) : \mathbb{P}(U \in \mathcal{U}) > \mathbb{P}(x)\Big\}. \tag{115}$$

The last inequality is due to $\mathbb{P}(U \in \mathcal{U})$ would not fall into $[\mathbb{P}(x,y), \mathbb{P}(x)]$.

Noteworthy, $\mathbb{P}(U \in \mathcal{U})$ in **CASE I** is non-empty, since we can choose $t \in \mathcal{U}^* := \text{argmin}_{\mathcal{U}'}\{\mathcal{U}' : \mathbb{P}(U \in \mathcal{U}') > \mathbb{P}(x)\}$ and set $\mathcal{U} := \mathcal{U}^*/\{t\}$. Then according to $\mathbb{P}(U) \in (\mathcal{P}^L)^c$, $\mathbb{P}(U \in \mathcal{U})$ falls into the above interval of **CASE I**. Moreover, due to

$$\mathbb{P}(U \in \mathcal{U}) + \frac{\mathbb{P}(x,y) - \mathbb{P}(U \in \mathcal{U})}{\mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U})} \mathbb{P}(u_t) < \mathbb{P}(U \in \mathcal{U} \cup \{t\}) = \mathbb{P}(U^*) = \min\Big\{\mathbb{P}(U \in \mathcal{U}) : \mathbb{P}(U \in \mathcal{U}) > \mathbb{P}(x)\Big\}, \tag{116}$$

combining with (113), (114), (115) and (116), finally, we get

$$\begin{aligned} \min \mathcal{P}^{d_u-1}_{y|do(x)} &\geq \min\Big\{s + \frac{\mathbb{P}(x,y) - s}{\mathbb{P}(x) - s}\mathbb{P}(u_t) : \mathcal{U} \subseteq \mathbb{R}/\{t\}, s = \mathbb{P}(U \in \mathcal{U}) \in \Big[\max\big\{0, \mathbb{P}(x) - \mathbb{P}(u_t)\big\}, \mathbb{P}(x,y)\Big]\Big\} \\ &= \Big\{s + \frac{\mathbb{P}(x,y) - s}{\mathbb{P}(x) - s}\mathbb{P}(u_t) : \mathcal{U} \subseteq \mathbb{R}/\{t\}, s \in \{p_{\min}(\mathcal{I}_t, \mathcal{I}'_t), p_{\max}(\mathcal{I}_t, \mathcal{I}'_t)\} \neq \emptyset\Big\}. \end{aligned} \tag{117}$$

Here $\mathcal{I}_t = \mathbb{R}/\{t\}$ and $\mathcal{I}'_t = \Big[\max\big\{0, \mathbb{P}(x) - \mathbb{P}(u_t)\big\}, \mathbb{P}(x,y)\Big]$. $p_{\min}$ and $p_{\max}$ are identified in our main text. The last inequality is blessed with the monotonicity. In sum, we get

$$\min \mathbb{P}(y \mid do(x)) = \min \mathcal{P}^{d_u-1}_{y|do(x)} \geq \begin{cases} \mathcal{B}'(\mathbb{P}(U); x, y) & \mathbb{P}(U) \in (\mathcal{P}^L)^c \\ \mathbb{P}(x,y) & \mathbb{P}(U) \in \mathcal{P}^L \end{cases} =: \mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)). \tag{118}$$

Here $\mathcal{B}'(\mathbb{P}(U); x, y)$ is identified in the main text. We conclude that (118) is the valid lower identification bound of $\mathbb{P}(y \mid do(x))$ in the multi-valued confounder case.

Furthermore, we consider the valid upper identification bound. Following the same strategy as above, the valid lower identification bound of $\mathbb{P}(\neg y \mid do(x))$ can be constructed as

$$\begin{cases} \mathcal{B}'(\mathbb{P}(U); x, \neg y) & \mathbb{P}(U) \in (\mathcal{P}^U)^c \\ \mathbb{P}(x, \neg y) & \mathbb{P}(U) \in \mathcal{P}^U \end{cases}. \tag{119}$$

Due to the fact $\mathbb{P}(y \mid do(x)) = 1 - \mathbb{P}(\neg y \mid x)$, the valid upper identification bound of $\mathbb{P}(y \mid do(x))$ is formalized as

$$\begin{cases} 1 - \mathcal{B}'(\mathbb{P}(U); x, \neg y) & \mathbb{P}(U) \in (\mathcal{P}^U)^c \\ \mathbb{P}(x,y) + \mathbb{P}(\neg x) & \mathbb{P}(U) \in \mathcal{P}^U \end{cases} =: \mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U)). \tag{120}$$

In sum, the validity part is completed.

(TIGHTNESS) We first take the lower bound for instance. Notice that the legitimate $\mathbb{P}(X, Y, U)$ under $\mathbb{P}(U) \in \mathcal{P}^L$ has already been established in Theorem 4.1, we only need to consider the non-vanilla case $\mathbb{P}(U) \in (\mathcal{P}^L)^c$. According to (117), it is sufficient to prove that for each legitimate pair $\{t, \mathcal{U}\}$ that satisfies $\mathbb{P}(U \in \mathcal{U}) \in \left[0 \vee \left(\mathbb{P}(x) - \mathbb{P}(u_t)\right), \mathbb{P}(x, y)\right]$ and $t \in (\mathcal{U})^c$, we could construct legitimate $\mathbb{P}(X, Y, U)$ which induces $\mathbb{P}(y \mid do(x)) = \mathbb{P}(U \in \mathcal{U}) + \frac{\mathbb{P}(x,y) - \mathbb{P}(U \in \mathcal{U})}{\mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U})} \mathbb{P}(u_t)$.

Notice that it must holds $\mathbb{P}(u_t) \geq \mathbb{P}(x, \neg y)$. The construction is as follows:

$$\mathbb{P}(U \in \mathcal{U}, x) = \mathbb{P}(U \in \mathcal{U}), \mathbb{P}(U \in (\mathcal{U} \cup \{u_t\})^c, x) = 0, \text{ and } \mathbb{P}(u_t, x) = \mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U}). \tag{121}$$

Moreover, the conditional probability $\mathbb{P}(Y \mid U, X)$ is set as

$$\forall u \in \mathcal{U}, \mathbb{P}(y \mid u, x) = 1, \forall u \in (\mathcal{U} \cup \{t\})^c, \mathbb{P}(y \mid u, x) = 0, \text{ and } \mathbb{P}(y \mid u_t, x) = \frac{\mathbb{P}(x, y) - \mathbb{P}(U \in \mathcal{U})}{\mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U})}.$$

$\forall u \in U, \mathbb{P}(u, \neg x)$ is supplemented by $\mathbb{P}(u) - \mathbb{P}(u, x)$ based on (121). Additionally, $\forall u \in U$, we set $\mathbb{P}(y \mid u, \neg x) = \mathbb{P}(y \mid \neg x)$ and $\mathbb{P}(\neg y \mid u, x') = 1 - \mathbb{P}(y \mid u, x'), x' \in \{0, 1\}$. It is easy to verify construction (121) is non-negative and compatible with the observed $\mathbb{P}(X, Y), \mathbb{P}(U)$.

We further consider the tightness of upper bound with the same strategy. Compared with (121), we re-construct the conditional probability $\mathbb{P}(Y \mid U, X)$ as

$$\forall u \in \mathcal{U}, \mathbb{P}(y \mid u, x) = 0, \forall u \in (\mathcal{U} \cup \{t\})^c, \mathbb{P}(y \mid u, x) = 1, \text{ and } \mathbb{P}(y \mid u_t, x) = \frac{\mathbb{P}(x, y)}{\mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U})}.$$

The other part holds the same as that of the lower bound. In sum, the tightness of the identification bound has been demonstrated.

As illustrated as above, we have proved for each specification of $\mathbb{P}(X, Y)$ and $\mathbb{P}(U)$, there exists a compatible joint distribution so that its induced $\mathbb{P}(y \mid do(x))$ is equal to the lower bound $\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U))$ and upper bound $\mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U))$. Now we are left with illustrating that for each o between these two bounds, there exists a legitimate corresponding $\mathbb{P}(X, Y, U)$ whose induced interventional probability is o.

To achieve this goal, the strategy is inherited from the proof of Theorem 3.3. The difference is that when dealing with the multi-valued confounders, our new construction would be more general and it is legitimate for any given $\mathbb{P}(U)$. Given any $\varepsilon > 0$ and a legitimate joint distribution $\mathbb{P}(X, Y, U)$, we construct a new legitimate joint distribution $\mathbb{P}^*(X, Y, U)$ satisfying

$$\begin{aligned}
&\mathbb{P}^*(x, y', u) = \begin{cases} \varepsilon \mathbb{P}(y' \mid u, x) & u \in \mathcal{U}_\varepsilon \\ \mathbb{P}(y', x, u) - \sum_{u' \in \mathcal{U}_\varepsilon} \mathbb{P}(y' \mid u', x)(\varepsilon - \mathbb{P}(u', x)) & u = \arg\max_{u'} \{\mathbb{P}(u', x, y') : u' \in \mathbb{R}\} \\ \mathbb{P}(y', x, u) & \text{Otherwise.} \end{cases}, \\
&\mathbb{P}^*(\neg x, y', u) = \mathbb{P}(y' \mid \neg x)(\mathbb{P}(u) - \mathbb{P}^*(u, x)), \text{ where } y' \in \{0, 1\}, \ \mathcal{U}_\varepsilon := \{u : \mathbb{P}(u, x) \leq \varepsilon\}.
\end{aligned} \tag{122}$$

It is easy to verify that (122) is a legitimate joint distribution generated from arbitrary given legitimate joint distribution $\mathbb{P}(X, Y, U)$. Then in the following part, we consider the sub region

$$\mathcal{P}^\varepsilon_{y|do(x)} = \Big\{\mathbb{P}(y \mid do(x)) : \{X, Y, U\} \text{ obeys } \mathbb{P}^*(X, Y, U) \text{ in (122) with some } \mathbb{P}(X, Y, U) \in \mathcal{P}_{XYU}\Big\}. \tag{123}$$

As $\varepsilon \to 0$, we show that $\min \mathcal{P}^\varepsilon_{y|do(x)}$ approaches the lower bound $\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U))$ and $\max \mathcal{P}^\varepsilon_{y|do(x)}$ approaches the upper bound $\mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U))$. We defer the detailed proof of legitimacy and convergence to Lemma M.3 in Appendix M.

In this sense, in order to prove $\forall o \in [\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)), \mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U))]$, there exists a legitimate joint probability so that $\mathbb{P}(y \mid do(x)) = o$, it is sufficient to prove that $\exists \varepsilon_0 > 0$ sufficiently small, such that $\forall \varepsilon \in (0, \varepsilon_0]$,

$$[\min \mathcal{P}^\varepsilon_{y|do(x)}, \max \mathcal{P}^\varepsilon_{y|do(x)}] \tag{124}$$

derived by (123) is a subset of the true identification region. Namely, for each point o' in the interval given by (123), there exists a legitimate joint distribution with its corresponding $\mathbb{P}(y \mid do(x)) \equiv o'$. To achieve this goal, we now recall the region given by (26):

$$\mathcal{O}_\varepsilon := \Big\{\mathbb{P}(y \mid do(x)) : \forall u \in U, \mathbb{P}(u, x) \geq \varepsilon, \mathbb{P}(Y, U, X) \text{ is compatible with } \mathbb{P}(X, Y), \mathbb{P}(U)\Big\}. \tag{125}$$

As we have already demonstrated in the proof of Theorem 3.3, $\mathcal{O}_\varepsilon$ in (125) is a closed interval on $\mathbb{R}$. Following the previous notations, we take $o_{\min}$ and $o_{\max}$ as the left and right side of the interval $\mathcal{O}_\varepsilon$, it is easily verified that $\exists\varepsilon_0$ sufficiently small, such that $\forall\varepsilon \in (0, \varepsilon_0]$,

$$o_{\min} \leq \min \mathcal{P}^{\varepsilon}_{y|do(x)} \leq \max \mathcal{P}^{\varepsilon}_{y|do(x)} \leq o_{\max}.$$

It indicates that the interval given by (124) is a subset of $\mathcal{O}_\varepsilon$. Furthermore, since $\mathcal{O}_\varepsilon$ is also a subset of the identification region by definition, it is straightforward that the interval (124) is the subset of the identification region. It completes the proof.

■

J. The proof of Proposition 4.6

Proof. (Lower bound) We first consider the lower bound and choose $\mathcal{I} = [\mathbb{P}(x,y), \mathbb{P}(x)]$. When $D\big(\mathbb{P}(U), \mathcal{I}\big) = 0$, it immediately leads to $\exists\mathcal{U} \in \mathbb{R}$, such that $\mathbb{P}(U \in \mathcal{U}) \in [\mathbb{P}(x,y), \mathbb{P}(x)]$ holds. Hence the if and only if condition in Theorem 3.3 holds and $\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U))$ equals to the vanilla lower bound $\mathbb{P}(x,y)$. In this sense, we only need to consider the case $D\big(\mathbb{P}(U), \mathcal{I}\big) > 0$:

$$\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) = \min_{t,s}\Big\{s + \frac{\mathbb{P}(x,y) - s}{\mathbb{P}(x) - s}\mathbb{P}(u_t)\Big\} = \min_{t,s}\Big\{\mathbb{P}(x,y) + (\mathbb{P}(x,y) - s)\frac{\mathbb{P}(u_t) - \mathbb{P}(x) + s}{\mathbb{P}(x) - s}\Big\}. \tag{126}$$

Here t spans $\{0, 1, ...d_u - 1\}$ and then s spans every legitimate $\mathbb{P}(U \in \mathcal{U})$ for each t. Namely,

$$s = \mathbb{P}(U \in \mathcal{U}) \in \Big[\max\big\{0, \mathbb{P}(x) - \mathbb{P}(u_t)\big\}, \mathbb{P}(x,y)\Big], t \notin \mathcal{U}. \tag{127}$$

Hence

$$\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) \geq \mathbb{P}(x,y) + \frac{D\big(\mathbb{P}(U), \mathcal{I}\big)^2}{\mathbb{P}(x) - s} \geq \mathbb{P}(x,y) + \frac{D\big(\mathbb{P}(U), \mathcal{I}\big)^2}{\mathbb{P}(x)}, \mathcal{I} = [\mathbb{P}(x,y), \mathbb{P}(x)], \tag{128}$$

On the other hand, we denote $\{\mathcal{U}^{\text{opt}}, t^{\text{opt}}\} = \arg\min_{\mathcal{U}\subseteq\mathbb{R}, t\in\mathcal{I}} |\mathbb{P}(U \in \mathcal{U}) - t|$. There are two possibilities:

(i) $\mathbb{P}(U \in \mathcal{U}^{\text{opt}}) \in [0, \mathbb{P}(x,y)], t^{\text{opt}} = \mathbb{P}(x,y)$, then in (127) we choose $s = \mathbb{P}(U \in \mathcal{U}^{\text{opt}}), t \in (\mathcal{U}^{\text{opt}})^c$. We have that

$$(126) \leq \mathbb{P}(x,y) + \Big(\mathbb{P}(x,y) - \mathbb{P}(U \in \mathcal{U}^{\text{opt}})\Big)\frac{\mathbb{P}(\neg x)}{\mathbb{P}(x) - s} \leq \mathbb{P}(x,y) + D\big(\mathbb{P}(U), \mathcal{I}\big)\frac{\mathbb{P}(\neg x)}{\mathbb{P}(x, \neg y)}. \tag{129}$$

(ii) $\mathbb{P}(U \in \mathcal{U}^{\text{opt}}) \in [\mathbb{P}(x), 1], t^{\text{opt}} = \mathbb{P}(x)$, then in (127) we choose $s = \mathbb{P}(U \in \mathcal{U}^{\text{opt}}/u_t) < \mathbb{P}(x,y)$, where $u_t \in \mathcal{U}^{\text{opt}}$. We have that

$$(126) \leq \mathbb{P}(x,y) + \mathbb{P}(x,y)\frac{\mathbb{P}(U \in \mathcal{U}^{\text{opt}}) - \mathbb{P}(x)}{\mathbb{P}(x) - s} \leq \mathbb{P}(x,y) + \mathbb{P}(x,y)\frac{D\big(\mathbb{P}(U), \mathcal{I}\big)}{\mathbb{P}(x, \neg y)}. \tag{130}$$

The combination of (129) and (130) leads to

$$\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) \leq \mathbb{P}(x,y) + \frac{(\mathbb{P}(\neg x) \vee \mathbb{P}(x,y))}{\mathbb{P}(x, \neg y)}D\big(\mathbb{P}(U), \mathcal{I}\big), \mathcal{I} = [\mathbb{P}(x,y), \mathbb{P}(x)]. \tag{131}$$

The combination of (128) and (131) leads to the first result.

(Upper bound) Second, we consider the upper bound and choose $\mathcal{I} = [\mathbb{P}(\neg x), 1 - \mathbb{P}(x, \neg y)]$. Following the same strategy, when $D\big(\mathbb{P}(U), \mathcal{I}\big) = 0$, then due to the if and only if condition in Theorem 3.3, we have that $\mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U))$ equals to the vanilla upper bound $\mathbb{P}(x,y) + \mathbb{P}(\neg x)$. Hence we also only need to consider $D\big(\mathbb{P}(U), \mathcal{I}\big) > 0$:

$$\mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) = 1 - \min_{t,s}\Big\{s + \frac{\mathbb{P}(x, \neg y) - s}{\mathbb{P}(x) - s}\mathbb{P}(u_t)\Big\} = 1 - \mathbb{P}(x, \neg y) - \min_{t,s}\Big\{(\mathbb{P}(x, \neg y) - s)\frac{\mathbb{P}(u_t) - \mathbb{P}(x) + s}{\mathbb{P}(x) - s}\Big\}. \tag{132}$$

Here $\{s, t\}$ follow the same setting as in the lower bound case. Then (132) leads to

$$\mathcal{UB}_{x,y}^{\text{mul}}(\mathbb{P}(U)) \leq 1 - \mathbb{P}(x, \neg y) - \frac{D\big(\mathbb{P}(U), \mathcal{I}\big)^2}{\mathbb{P}(x)} = \mathbb{P}(x, y) + \mathbb{P}(\neg x) - \alpha_x \Delta_{x,\neg y}^2, \tag{133}$$

On the other hand, we follow the notations $\{\mathcal{U}^{\text{opt}}, t^{\text{opt}}\}$ as above with $\mathcal{I} = [\mathbb{P}(\neg x), 1 - \mathbb{P}(x, \neg y)]$. There are also two possibilities:

(i) $\mathbb{P}(U \in \mathcal{U}^{\text{opt}}) \in [0, \mathbb{P}(\neg x)], t^{\text{opt}} = \mathbb{P}(\neg x)$, then in (127) we choose $s = \mathbb{P}(U \in (\mathcal{U}^{\text{opt}})^c / u_t) < \mathbb{P}(x, \neg y)$, where $u_t \in (\mathcal{U}^{\text{opt}})^c$. We have that

$$(132) \geq 1 - \mathbb{P}(x, \neg y) - \mathbb{P}(x, \neg y) \frac{\mathbb{P}(\neg x) - \mathbb{P}(U \in \mathcal{U}^{\text{opt}})}{\mathbb{P}(x) - s} \geq \mathbb{P}(x, y) + \mathbb{P}(\neg x) - \mathbb{P}(x, \neg y) \frac{D\big(\mathbb{P}(U), \mathcal{I}\big)}{\mathbb{P}(x, y)}. \tag{134}$$

(ii) $\mathbb{P}(U \in \mathcal{U}^{\text{opt}}) \in [1 - \mathbb{P}(x, \neg y), 1], t^{\text{opt}} = 1 - \mathbb{P}(x, \neg y)$, then in (127) we choose $s = \mathbb{P}(U \in (\mathcal{U}^{\text{opt}})^c)$ and $t \in \mathcal{U}^{\text{opt}}$. We have that

$$(132) \geq 1 - \mathbb{P}(x, \neg y) - \Big(\mathbb{P}(U \in (\mathcal{U}^{\text{opt}})) - \big(1 - \mathbb{P}(x, \neg y)\big)\Big) \frac{\mathbb{P}(\neg x)}{\mathbb{P}(x, y)} \geq \mathbb{P}(x, y) + \mathbb{P}(\neg x) - D\big(\mathbb{P}(U), \mathcal{I}\big) \frac{\mathbb{P}(\neg x)}{\mathbb{P}(x, y)}. \tag{135}$$

The combination of (134) and (135) leads to

$$\mathcal{UB}_{x,y}^{\text{mul}}(\mathbb{P}(U)) \leq \mathbb{P}(x, y) + \mathbb{P}(\neg x) - \frac{(\mathbb{P}(\neg x) \vee \mathbb{P}(x, \neg y))}{\mathbb{P}(x, y)} D\big(\mathbb{P}(U), \mathcal{I}\big) = \mathbb{P}(x, y) + \mathbb{P}(\neg x) - \beta_{x,\neg y} D\big(\mathbb{P}(U), \mathcal{I}\big).$$

Here $\mathcal{I} = [\mathbb{P}(\neg x), 1 - \mathbb{P}(x, \neg y)]$, and it completes the proof.

■

K. The proof of Proposition 4.7

Proof. According to the natural composition of Proposition 4.6, it directly leads to $\underline{\text{ATE}} - \text{ATE}_{\text{vanilla}}^{\text{L}} \geq \alpha_1 \Delta_{1,1}^2 + \alpha_0 \Delta_{0,0}^2$ and $\text{ATE}_{\text{vanilla}}^{\text{U}} - \overline{\text{ATE}} \geq \alpha_0 \Delta_{0,1}^2 + \alpha_1 \Delta_{1,0}^2$. Hence we only need consider the rest part.

(Lower bound) We only need analyze the non-vanilla case. Under $\mathbb{P}(U) \notin \mathcal{P}_{\text{ATE}}^{L}$, We aim to prove $\text{ATE}_{\text{vanilla}}^{\text{L}} + D_{\text{ATE}}\big(\mathbb{P}(U), \{\mathcal{I}_{0,0}, \mathcal{I}_{1,1}\}\big)^2 / d_u$ serves as the valid lower bound. Recalling that in (78), we have that

$$\begin{aligned}
\text{ATE} - \text{ATE}_{\text{vanilla}}^{\text{L}} =& \sum_u \left[\mathbb{P}(y_1 \mid u, x_1)\mathbb{P}(u, x_0) + \mathbb{P}(y_0 \mid u, x_0)\mathbb{P}(u, x_1)\right] \\
\geq& \frac{1}{\mathbb{P}(x_1)} \sum_u \Big(\mathbb{P}(y_1, u, x_1) \wedge \mathbb{P}(u, x_0)\Big)^2 + \frac{1}{\mathbb{P}(x_0)} \sum_u \Big(\mathbb{P}(y_0, u, x_0) \wedge \mathbb{P}(u, x_1)\Big)^2.
\end{aligned} \tag{136}$$

We denote that

$$\mathcal{U}_t^* := \{u : \mathbb{P}(y_t, u, x_t) > \mathbb{P}(u, x_{\neg t})\}, t \in \{0, 1\}. \tag{137}$$

According to the Cauchy–Schwartz inequality, we have that

$$\sum_u \big(\mathbb{P}(y_t, u, x_t) \wedge \mathbb{P}(u, x_{\neg t})\big)^2 \geq \frac{1}{d_u} \Big(\mathbb{P}(y_t, x_t, U \in (\mathcal{U}_t^*)^c) + \mathbb{P}(U \in \mathcal{U}_t^*, x_{\neg t})\Big)^2 \tag{138}$$

Combined with (136) and (138), we have

$$\begin{aligned}
\text{ATE} - \text{ATE}_{\text{vanilla}}^{\text{L}} \geq& \sum_{t=0,1} \frac{1}{\mathbb{P}(x_t)} \sum_u \Big(\mathbb{P}(y_t, u, x_t) \wedge \mathbb{P}(u, x_{\neg t})\Big)^2 \\
\geq& \frac{1}{d_u} \sum_{t=0,1} \frac{1}{\mathbb{P}(x_t)} \Big(\mathbb{P}(y_t, x_t, U \in (\mathcal{U}_t^*)^c) + \mathbb{P}(U \in \mathcal{U}_t^*, x_{\neg t})\Big)^2 \sum_{t=0,1} \mathbb{P}(x_t) \\
\geq& \frac{1}{d_u} \Big(\sum_{t=0,1} \mathbb{P}(y_t, x_t, U \in (\mathcal{U}_t^*)^c) + \mathbb{P}(U \in \mathcal{U}_t^*, x_{\neg t})\Big)^2
\end{aligned} \tag{139}$$

The last inequality in (139) is also due to the Cauchy–Schwarz inequality. Moreover, $\mathbb{P}(U \in \mathcal{U}_t^*), t = 0, 1$ can be bounded as

$$\begin{aligned}\mathbb{P}(U \in \mathcal{U}_t^*) &= \mathbb{P}(U \in \mathcal{U}_t^*, x_t) + \mathbb{P}(U \in \mathcal{U}_t^*, x_{\neg t}) \le \mathbb{P}(x_t) + \mathbb{P}(U \in \mathcal{U}_t^*, x_{\neg t}),\\ \mathbb{P}(U \in \mathcal{U}_t^*) &\ge \mathbb{P}(U \in \mathcal{U}_t^*, x_t, y_t) = \mathbb{P}(x_t, y_t) - \mathbb{P}(U \in (\mathcal{U}_t^*)^c, x_t, y_t),\end{aligned} \tag{140}$$

Importantly, according to the definition of $\mathcal{U}_t^*, t = 0, 1$, we have $\mathcal{U}_0^* \cap \mathcal{U}_1^* \neq \emptyset$. Otherwise $\exists u \in \mathbb{R}$, such that $\sum_{t=0,1} \mathbb{P}(y_t, u, x_t) > \sum_{t=0,1} \mathbb{P}(u, x_{\neg t})$, which is a contradiction. It indicates that $\mathbb{P}(U \in \mathcal{U}_t^*)$ locates in

$$D_{\text{ATE}}\big(\mathbb{P}(U), \{\mathcal{I}_{0,0}, \mathcal{I}_{1,1}\}\big) \le \sum_{t=0,1} \mathbb{P}(U \in (\mathcal{U}_t^*)^c, x_t, y_t) \vee \mathbb{P}(U \in \mathcal{U}_t^*, x_{\neg t}). \tag{141}$$

Combined with (139) and (141), we have

$$\text{ATE} - \text{ATE}^{\text{L}}_{\text{vanilla}} \ge \frac{1}{d_u} D_{\text{ATE}}\big(\mathbb{P}(U), \{\mathcal{I}_{0,0}, \mathcal{I}_{1,1}\}\big)^2. \tag{142}$$

According to the above analysis, we get $\underline{\text{ATE}} - \text{ATE}^{\text{L}}_{\text{vanilla}} \ge \Delta^2_{\text{ATE}}/d_u,$ where $\Delta_{\text{ATE}} = D_{\text{ATE}}(\mathbb{P}(U), \{\mathcal{I}_{0,0}, \mathcal{I}_{1,1}\})$. On this basis, we are only left with the demonstration of proving $\underline{\text{ATE}} - \text{ATE}^{\text{L}}_{\text{vanilla}} \le (\beta_{1,1} + \beta_{0,1})\Delta_{\text{ATE}}$. We prove it via direct construction. We denote

$$\{\mathcal{U}_0^{\text{opt}}, \mathcal{U}_1^{\text{opt}}\} = \arg\min_{\mathcal{U}_0, \mathcal{U}_1} \Big(|\mathbb{P}(U \in \mathcal{U}_0) - t_0| + |\mathbb{P}(U \in \mathcal{U}_1) - t_1|\Big), \tag{143}$$

where $\mathcal{U}_0, \mathcal{U}_1 \subseteq \mathbb{R}, \mathcal{U}_0 \cap \mathcal{U}_1 = \emptyset, t_0 \in [\mathbb{P}(x_1, y_1), \mathbb{P}(x_1)], t_1 \in [\mathbb{P}(x_0, y_0), \mathbb{P}(x_0)]$. It could be separated into two cases:
(i) $\mathbb{P}(U \in \mathcal{U}_0^{\text{opt}}) \le \mathbb{P}(x_0)$ and $\mathbb{P}(U \in \mathcal{U}_1^{\text{opt}}) \le \mathbb{P}(x_1)$. We follow the construction of $\mathbb{P}(U, X)$ in (144):

$$\begin{bmatrix}\mathbb{P}(U \in \mathcal{U}_0^{\text{opt}}, x_1) & \mathbb{P}(U \in \mathcal{U}_1^{\text{opt}}, x_1) & \mathbb{P}(U \in (\mathcal{U}_0^{\text{opt}} \cup \mathcal{U}_1^{\text{opt}})^c, x_1)\\ \mathbb{P}(U \in \mathcal{U}_0^{\text{opt}}, x_0) & \mathbb{P}(U \in \mathcal{U}_1^{\text{opt}}, x_0) & \mathbb{P}(U \in (\mathcal{U}_0^{\text{opt}} \cup \mathcal{U}_1^{\text{opt}})^c, x_0)\end{bmatrix} = \begin{bmatrix}0 & \mathbb{P}(U \in \mathcal{U}_1^{\text{opt}}) & \mathbb{P}(x_1) - \mathbb{P}(U \in \mathcal{U}_1^{\text{opt}})\\ \mathbb{P}(U \in \mathcal{U}_0^{\text{opt}}) & 0 & \mathbb{P}(x_0) - \mathbb{P}(U \in \mathcal{U}_0^{\text{opt}})\end{bmatrix}. \tag{144}$$

Moreover, the conditional probability $\mathbb{P}(Y \mid U, X)$ is constructed by

$$\begin{aligned}&\forall u \in \mathcal{U}_0^{\text{opt}}, \mathbb{P}(y_1 \mid u, x_1) = 0, \mathbb{P}(y_1 \mid u, x_0) = \delta_0/\mathbb{P}(U \in \mathcal{U}_0^{\text{opt}});\\ &\forall u \in \mathcal{U}_1^{\text{opt}}, \mathbb{P}(y_1 \mid u, x_1) = 1 - \delta_1/\mathbb{P}(U \in \mathcal{U}_1^{\text{opt}}), \mathbb{P}(y_1 \mid u, x_0) = 1;\\ &\forall u \in (\mathcal{U}_0^{\text{opt}} \cup \mathcal{U}_1^{\text{opt}})^c, \mathbb{P}(y_1 \mid u, x_1) = \frac{\mathbb{P}(x_1, y_1) - \mathbb{P}(U \in \mathcal{U}_1^{\text{opt}}) + \delta_1}{\mathbb{P}(x_1) - \mathbb{P}(U \in \mathcal{U}_1^{\text{opt}})}, \mathbb{P}(y_1 \mid u, x_0) = \frac{\mathbb{P}(x_0, y_1) - \delta_0}{\mathbb{P}(x_0) - \mathbb{P}(U \in \mathcal{U}_0^{\text{opt}})}.\end{aligned} \tag{145}$$

Here

$$\delta_t = (\mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}) - \mathbb{P}(x_t, y_t))\mathbb{I}(\mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}) > \mathbb{P}(x_t, y_t)), t = 0, 1.$$

Moreover, we also choose $\mathbb{P}(y_0 \mid u_i, x') = 1 - \mathbb{P}(y_1 \mid u_i, x')$. Here $u \in \{0, 1, ...d_u - 1\}, x' \in \{0, 1\}$. It is easy to verify the construction (144)-(145) is non-negative and compatible with the observed $\mathbb{P}(X, Y), \mathbb{P}(U)$. We can verify that

$$\begin{aligned}\mathbb{P}(y_t \mid do(x_t)) =& \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}) - \delta_t + \frac{\mathbb{P}(x_t, y_t) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}) + \delta_t}{\mathbb{P}(x_t) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}})}\mathbb{P}(U \in (\mathcal{U}_t^{\text{opt}} \cup \mathcal{U}_{\neg t}^{\text{opt}})^c)\\ =& \mathbb{P}(x_t, y_t) + [\mathbb{P}(x_t, y_t) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}) + \delta_t]\frac{\mathbb{P}(U \in (\mathcal{U}_t^{\text{opt}} \cup \mathcal{U}_{\neg t}^{\text{opt}})^c) - \mathbb{P}(x_t) + \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}})}{\mathbb{P}(x_t) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}})}\\ \le& \mathbb{P}(x_t, y_t) + \Delta_{\text{ATE}}\frac{1 - \mathbb{P}(x_t)}{\mathbb{P}(x_t) - \mathbb{P}(x_t, y_t)} = \mathbb{P}(x_t, y_t) + \Delta_{\text{ATE}}\frac{\mathbb{P}(x_{\neg t})}{\mathbb{P}(x_t, y_{\neg t})}, t = 0, 1.\end{aligned} \tag{146}$$

Therefore, according to $\text{ATE} = \mathbb{P}(y_1 \mid do(x_1)) - \mathbb{P}(y_1 \mid do(x_0)) = -1 + \sum_{t=0,1} \mathbb{P}(y_t \mid do(x_t))$, ATE under the construction in (144)-(145) can be upper-bounded by

$$-\mathbb{P}(x_1, y_0) - \mathbb{P}(x_0, y_1) + (\frac{\mathbb{P}(x_0)}{\mathbb{P}(x_1, y_0)} + \frac{\mathbb{P}(x_1)}{\mathbb{P}(x_0, y_1)})\Delta_{\text{ATE}}.$$

Hence it can be concluded as $\underline{\text{ATE}} \le \text{ATE}^{\text{L}}_{\text{vanilla}} + (\beta_{1,1} + \beta_{0,0}))\Delta_{\text{ATE}}$.

(ii) $\exists t \in \{0,1\}, s.t. \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}) > \mathbb{P}(x_t)$. In this case, according to the definition (143), it immediately leads to $\mathbb{P}(U \in \mathcal{U}_{\neg t}^{\text{opt}}) = 1 - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}) \le \mathbb{P}(x_{\neg t})$. Moreover, We select $u_t^{\text{opt}} \in \mathcal{U}_t^{\text{opt}}$, it must hold $\mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/u_t^{\text{opt}}) \le \mathbb{P}(x_t, y_t)$. We take construction on different groups:

$$\begin{aligned}&\begin{bmatrix} \mathbb{P}(U \in \mathcal{U}_{\neg t}^{\text{opt}}, x_t) & \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\}, x_t) & \mathbb{P}(U = u_t^{\text{opt}}, x_t) \\ \mathbb{P}(U \in \mathcal{U}_{\neg t}^{\text{opt}}, x_{\neg t}) & \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\}, x_{\neg t}) & \mathbb{P}(U = u_t^{\text{opt}}, x_{\neg t}) \end{bmatrix} \\ =&\begin{bmatrix} 0 & \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\}) & \mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\}) \\ \mathbb{P}(U \in \mathcal{U}_{\neg t}^{\text{opt}}) & 0 & \mathbb{P}(x_0) - \mathbb{P}(U \in \mathcal{U}_{\neg t}^{\text{opt}}) \end{bmatrix}.\end{aligned} \tag{147}$$

Moreover, the conditional probability $\mathbb{P}(Y \mid U, X)$ is constructed by

$$\begin{aligned}&\forall u \in \mathcal{U}_{\neg t}^{\text{opt}}, \mathbb{P}(y_t \mid u, x_t) = 0, \mathbb{P}(y_t \mid u, x_{\neg t}) = \delta_{\neg t}/\mathbb{P}(U \in \mathcal{U}_{\neg t}^{\text{opt}}); \\ &\forall u \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\}, \mathbb{P}(y_t \mid u, x_t) = 1, \mathbb{P}(y_t \mid u, x_{\neg t}) = 1; \\ &\forall u = u_t^{\text{opt}}, \mathbb{P}(y_t \mid u, x_t) = \frac{\mathbb{P}(x_t, y_t) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\})}{\mathbb{P}(x_t) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\})}, \mathbb{P}(y_t \mid u, x_{\neg t}) = \frac{\mathbb{P}(x_{\neg t}, y_t) - \delta_{\neg t}}{\mathbb{P}(x_{\neg t}) - \mathbb{P}(U \in \mathcal{U}_0)}.\end{aligned} \tag{148}$$

Here $\delta_{t'}, t' = 0,1$ has been identified in the case (i). We can verify that

$$\begin{aligned}\mathbb{P}(y_t \mid do(x_t)) &= \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\}) + \frac{\mathbb{P}(x_t, y_t) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\})}{\mathbb{P}(x_t) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\})}\mathbb{P}(u_t^{\text{opt}}) \\ &= \mathbb{P}(x_t, y_t) + \Big(\mathbb{P}(x_t, y_t) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\})\Big)\frac{\mathbb{P}(u_t^{\text{opt}}) - \mathbb{P}(x_t) + \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\})}{\mathbb{P}(x_t) - \mathbb{P}(U \in \mathcal{U}_t^{\text{opt}}/\{u_t^{\text{opt}}\})} \\ &\le \mathbb{P}(x_t, y_t) + \mathbb{P}(x_t, y_t)\frac{\Delta_{\text{ATE}}}{\mathbb{P}(x_t) - \mathbb{P}(x_t, y_t)} = \mathbb{P}(x_t, y_t) + \frac{\mathbb{P}(x_t, y_t)}{\mathbb{P}(x_t, y_{\neg t})}\Delta_{\text{ATE}}.\end{aligned} \tag{149}$$

On the other hand,

$$\mathbb{P}(y_{\neg t} \mid do(x_{\neg t})) \le \mathbb{P}(x_{\neg t}, y_{\neg t}) + \Delta_{\text{ATE}}\frac{\mathbb{P}(x_t)}{\mathbb{P}(x_{\neg t}, y_t)}. \tag{150}$$

Hence under the construction (147) and (148), the ATE can be upper bounded by

$$\text{ATE} = -1 + \sum_{t=0,1}\mathbb{P}(y_t \mid do(x_t)) \le -\mathbb{P}(x_1, y_0) - \mathbb{P}(x_0, y_1) + \sum_{t=0,1}\frac{\mathbb{P}(x_t, y_t) \vee \mathbb{P}(x_{\neg t})}{\mathbb{P}(x_t, y_{\neg t})}\Delta_{\text{ATE}}. \tag{151}$$

it can also be concluded as $\underline{\text{ATE}} \le \text{ATE}^{\text{L}}_{\text{vanilla}} + (\beta_{1,1} + \beta_{0,0})\Delta_{\text{ATE}}$. Combining case (i)-(ii), the desired result follows.

(Upper bound) We adopt the same strategy. Considering the non-vanilla case $\mathbb{P}(U) \in (\mathcal{P}^U_{ATE})^c$, we only need to further demonstrate that $\text{ATE}^{\text{U}}_{\text{vanilla}} - D_{\text{ATE}}\big(\mathbb{P}(U), \{\mathcal{I}_{0,1}, \mathcal{I}_{1,0}\}\big)^2/d_u$ serves as a valid upper bound. Compared with (137), we re-denote

$$\mathcal{U}_t^* := \{u : \mathbb{P}(y_t, u, x_{\neg t}) \ge \mathbb{P}(u, x_t)\}, t \in \{0,1\}.$$

Recalling that in (79), it holds that

$$\begin{aligned}\text{ATE}^{\text{U}}_{\text{vanilla}} - \text{ATE} &= \sum_u \left[\mathbb{P}(y_1 \mid u, x_0)\mathbb{P}(u, x_1) + \mathbb{P}(y_0 \mid u, x_1)\mathbb{P}(u, x_0)\right] \\ &\ge \sum_{t=0,1}\frac{1}{\mathbb{P}(x_{\neg t})}\sum_u \Big(\mathbb{P}(y_t, u, x_{\neg t}) \wedge \mathbb{P}(u, x_t)\Big)^2 \\ &\ge \frac{1}{d_u}\sum_{t=0,1}\frac{1}{\mathbb{P}(x_{\neg t})}\Big(\mathbb{P}(y_t, x_{\neg t}, U \in (\mathcal{U}_t^*)^c) + \mathbb{P}(U \in \mathcal{U}_t^*, x_t)\Big)^2 \sum_{t=0,1}\mathbb{P}(x_{\neg t}) \\ &\ge \frac{1}{d_u}\Big(\sum_{t=0,1}\mathbb{P}(y_t, x_{\neg t}, U \in (\mathcal{U}_t^*)^c) + \mathbb{P}(U \in \mathcal{U}_t^*, x_t)\Big)^2.\end{aligned} \tag{152}$$

Here the last two inequalities are both due to the Cauchy–Schwartz inequality. We get

$$\begin{aligned}\mathbb{P}(U \in \mathcal{U}_t^*) &= \mathbb{P}(U \in \mathcal{U}_t^*, x_{\neg t}) + \mathbb{P}(U \in \mathcal{U}_t^*, x_t) \leq \mathbb{P}(x_{\neg t}) + \mathbb{P}(U \in \mathcal{U}_t^*, x_t), \\ \mathbb{P}(U \in \mathcal{U}_t^*) &\geq \mathbb{P}(U \in \mathcal{U}_t^*, x_{\neg t}, y_t) = \mathbb{P}(x_{\neg t}, y_t) - \mathbb{P}(U \in (\mathcal{U}_t^*)^c, x_{\neg t}, y_t).\end{aligned} \tag{153}$$

According to $\mathcal{U}_0^* \cap \mathcal{U}_1^* \neq \emptyset$ with the same reason as above, we claim that $\mathbb{P}(U \in \mathcal{U}_t^*)$ locates in

$$D_{\mathrm{ATE}}\big(\mathbb{P}(U), \{\mathcal{I}_{0,1}, \mathcal{I}_{1,0}\}\big) \leq \sum_{t=0,1} \mathbb{P}(U \in (\mathcal{U}_t^*)^c, x_{\neg t}, y_t) \vee \mathbb{P}(U \in \mathcal{U}_t^*, x_t). \tag{154}$$

Combined with (152) and (154), we have

$$\mathrm{ATE}^U_{\mathrm{vanilla}} - \mathrm{ATE} \geq \frac{1}{d_u} D_{\mathrm{ATE}}\big(\mathbb{P}(U), \{\mathcal{I}_{0,0}, \mathcal{I}_{1,1}\}\big)^2.. \tag{155}$$

The desired result follows.

Moreover, notice that the analysis on the upper bound $\mathbb{P}(y_1 \mid do(x_1)) - \mathbb{P}(y_1 \mid do(x_0))$ is equivalent to the analysis on the lower bound of $\mathbb{P}(y_1 \mid do(x_0)) - \mathbb{P}(y_1 \mid do(x_1))$. Based on the above analysis on the lower bound and exchange $\{x_0, x_1\}$ with each other, we directly get that there exists legitimate $\mathbb{P}(X, Y, U)$ such that

$$\mathbb{P}(y_1 \mid do(x_0)) - \mathbb{P}(y_1 \mid do(x_1)) \leq -\mathbb{P}(x_0, y_0) - \mathbb{P}(x_1, y_1) + \sum_{t=0,1} \frac{\mathbb{P}(x_{\neg t}, y_t) \vee \mathbb{P}(x_t)}{\mathbb{P}(x_{\neg t}, y_{\neg t})} \Delta_{\mathrm{ATE}}.$$

It is concluded that the tight upper bound of ATE is lower-bounded by

$$\mathbb{P}(x_0, y_0) + \mathbb{P}(x_1, y_1) - \sum_{t=0,1} \frac{\mathbb{P}(x_{\neg t}, y_t) \vee \mathbb{P}(x_t)}{\mathbb{P}(x_{\neg t}, y_{\neg t})} \Delta_{\mathrm{ATE}} = \mathrm{ATE}^{\mathrm{U}}_{\mathrm{vanilla}} - (\beta_{1,0} + \beta_{0,1}) \Delta_{\mathrm{ATE}}.$$

Here $\Delta_{\mathrm{ATE}} = D_{\mathrm{ATE}}(\mathbb{P}(U), \{\mathcal{I}_{0,1}, \mathcal{I}_{1,0}\})$. Hence we get $\mathrm{ATE}^{\mathrm{U}}_{\mathrm{vanilla}} - \overline{\mathrm{ATE}} \leq (\beta_{1,0} + \beta_{0,1}) \Delta_{\mathrm{ATE}}$. It completes the proof.

■

L. Auxiliary algorithms

Algorithm 1 Approximation TPI algorithm.

Require: Observed data $\mathbb{P}(x', y'), \mathbb{P}(u_i), i = 0, 1, ...d_u - 1$; Null set $\mathcal{S}_{y'} = \emptyset$, $x', y' \in \{0, 1\}$; approximation error η.

Ensure: The approximated tight identification region $[\widehat{\mathcal{LB}}^{\text{mul}}_{x,y}(\mathbb{P}(U)), \widehat{\mathcal{UB}}^{\text{mul}}_{x,y}(\mathbb{P}(U))] := [\min \mathcal{S}_y, 1 - \min \mathcal{S}_{\neg y}]$.

output The tight identification region of $\mathbb{P}(y \mid do(x))$ with approximation error $\beta_{x,y'}\eta$, where constant $\beta_{x,y'}$ has been identified in Proposition 4.6. Namely, we produce

$$|\widehat{\mathcal{LB}}^{\text{mul}}_{x,y}(\mathbb{P}(U)) - \mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U))| \leq \beta_{x,y}\eta, |\widehat{\mathcal{UB}}^{\text{mul}}_{x,y}(\mathbb{P}(U)) - \mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U))| \leq \beta_{x,\neg y}\eta.$$

for $y' = y, \neg y$ **do**
  if SSP-min$\Big(\{\mathbb{P}(u_t)\}_{t=0}^{d_u-1}, \mathbb{P}(x, y')\Big) \leq \mathbb{P}(x)$ **then**
    return $\mathcal{S}_{y'} = \mathbb{P}(x, y')$.
  else
    for each t satisfying $\mathbb{P}(u_t) \geq \mathbb{P}(x, \neg y')$ **do**

$$s_{\min} = \min \mathbf{SSP}\Big(\{\mathbb{P}(u_i)\}_{i=0}^{d_u-1}/\{\mathbb{P}(u_t)\}, \mathcal{I}^l, \eta\Big), s_{\max} = \max \mathbf{SSP}\Big(\{\mathbb{P}(u_t)\}_{t=0}^{d_u-1}/\{\mathbb{P}(u_t)\}, \mathcal{I}^u, \eta\Big),$$

      where $[\mathcal{I}^l, \mathcal{I}^u] := [0 \vee \big(\mathbb{P}(x') - \mathbb{P}(u_t)\big), \mathbb{P}(x', y')]$.
      Moreover, $\mathcal{S}_{y'} = \mathcal{S}_{y'} \cup \{s_{\min} + \frac{\mathbb{P}(x,y) - s_{\min}}{\mathbb{P}(x) - s_{\min}}\mathbb{P}(u_t)\}$ when $s_{\min} \leq \mathcal{I}^u$;, and $\mathcal{S}_{y'} = \mathcal{S}_{y'} \cup \{s_{\max} + \frac{\mathbb{P}(x,y) - s_{\max}}{\mathbb{P}(x) - s_{\max}}\mathbb{P}(u_t)\}$ when $s_{\max} \geq \mathcal{I}^l$; .
    end for
  end if
end for

The traditional subset-sum problem (SSP) problem is to explore the sub-optimal subset, such that its sum is larger (smaller) than a certain threshold. The algorithm for SSP is illustrated as follows. For brevity, we denote
$\min(\emptyset) = -\infty, \max(\emptyset) = +\infty$,

Algorithm 2 SSP$(\mathcal{I}, \mathcal{I}', \eta)$ algorithm.

Require: Region $\mathcal{I}, \mathcal{I}'$ and approximation error η.
Ensure: sub-optimal subsets $\widehat{\mathcal{U}}_{\min}, \widehat{\mathcal{U}}_{\max} \subseteq R$ such that $\mathbb{P}(U \in \widehat{\mathcal{U}}^{\text{T}}_{\min})/p_{\min}(\mathcal{I}, \mathcal{I}') \in [1 - \eta, 1 + \eta]$ and $\mathbb{P}(U \in \widehat{\mathcal{U}}^{\text{T}}_{\max})/p_{\max}(\mathcal{I}, \mathcal{I}') \in [1 - \eta, 1 + \eta]$.
Initialize $\mathcal{S}_{\min} = \{\mathbb{P}(U \in \mathcal{I})\}, \mathcal{S}_{\max} = \{0\}$.
for $u \in \mathcal{I}$ **do**
  $\mathcal{S}_{\min} = \mathcal{S}_{\min} \cup \{\mathcal{S}_{\min} - \mathbb{P}(u)\}, \mathcal{S}_{\max} = \mathcal{S}_{\max} \cup \{\mathcal{S}_{\max} + \mathbb{P}(u)\}$.
  Update $\mathcal{S}_{\min}$ by removing each element that is lower than $\min \mathcal{I}'$; Update $\mathcal{S}_{\max}$ by removing each element that is upper than $\max \mathcal{I}'$.
  For each element $q \in \mathcal{S}_{\mathcal{A}}$, if $\exists q' \in \mathcal{S}_{\mathcal{A}}$, such that $q'/q \in [1 - \eta/d_u, 1 + \eta/d_u]$, then remove q'. $\mathcal{A} \in \{\min, \max\}$.
end for
Set $\widehat{\mathcal{U}}_{\min} = \min \mathcal{S}_{\min}, \widehat{\mathcal{U}}_{\max} = \max \mathcal{S}_{\max}$.

M. Auxiliary lemmas

Lemma M.1 (Justification of (44))**.**

$$\mathcal{S}_{t,i} = \Big\{\mathbb{P}(y_1 \mid do(x_i)) + \Big[\frac{1}{\mathbb{P}(x_i \mid u_t)} - \frac{1}{\mathbb{P}(x_i \mid u_{\neg t})}\Big]s : s \in \{\mathbb{P}(x_i, y_1, u_{\neg t}), \mathbb{P}(x_i, y_0, u_t)\}\Big\}. \tag{156}$$

Proof. We have

$$\mathcal{S}_{t,i} = \mathbb{P}(y_1 \mid do(x_i)) + \Big[\frac{1}{\mathbb{P}(x_i \mid u_t)} - \frac{1}{\mathbb{P}(x_i \mid u_{\neg t})}\Big]p = \frac{\mathbb{P}(y_1, u_t, x_i) + p}{\mathbb{P}(u_t, x_i)}\mathbb{P}(u_t) + \frac{\mathbb{P}(y_1, u_{\neg t}, x_i) - p}{\mathbb{P}(u_{\neg t}, x_i)}\mathbb{P}(u_{\neg t}). \tag{157}$$

When we choose $p = \mathbb{P}(x_i, y_1, u_{\neg t})$, we have

$$S_{t,i} = \frac{\mathbb{P}(x_i, y_1)}{\mathbb{P}(u_t, x_i)}\mathbb{P}(u_t). \tag{158}$$

When we choose $p = \mathbb{P}(x_i, y_0, u_t)$, we have

$$\mathcal{S}_{t,i} = \frac{\mathbb{P}(u_t, x_i)}{\mathbb{P}(u_t, x_i)}\mathbb{P}(u_t) + \frac{\mathbb{P}(y_1, u_{\neg t}, x_i) - \mathbb{P}(x_i, y_0, u_t)}{\mathbb{P}(u_{\neg t}, x_i)}\mathbb{P}(u_{\neg t}) = \frac{\mathbb{P}(x_i, y_1) - \mathbb{P}(u_t, x_i)}{\mathbb{P}(x_i) - \mathbb{P}(u_t, x_i)}\mathbb{P}(u_{\neg t}) + \mathbb{P}(u_t). \tag{159}$$

(158) and (159) are consistent with the definition of $\mathcal{S}_{t,i}$ in (41). ■

Lemma M.2 (Justification of **Claim I** in Appendix I)**.** $\mathcal{P}_{XYU}^{d_u-1} \neq \emptyset$.

Proof. We aim to construct a legitimate $\mathbb{P}(X, Y, U)$ within $\mathcal{P}_{XYU}^{d_u-1}$. For given $\mathbb{P}(U)$, we choose

$$\mathcal{U}' := \underset{\mathcal{U}}{\operatorname{argmax}}\{\mathbb{P}(U \in \mathcal{U}) \mid \mathbb{P}(U \in \mathcal{U}) < \mathbb{P}(x, y)\}.$$

Apparently, $\mathbb{P}(U \in \mathcal{U}') \leq \mathbb{P}(x, y)$. We then choose $u_c \in (\mathcal{U}')^c$, and thus legitimate constructions could be constructed. There are at most two possibilities:

CASE I: $\mathbb{P}(U \in \mathcal{U}' \cup u_c) \geq \mathbb{P}(x)$:

We choose

$$\mathbb{P}(U \in \mathcal{U}', x) = \mathbb{P}(U \in \mathcal{U}'), \mathbb{P}(U \in (\mathcal{U}' \cup \{u_c\})^c, x) = 0, \text{ and } \mathbb{P}(u_c, x) = \mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U}'). \tag{160}$$

Moreover, the conditional probability $\mathbb{P}(Y \mid U, X)$ is set as

$$\forall U \in \mathcal{U}', \mathbb{P}(y \mid u, x) = 1, \forall U \in (\mathcal{U}' \cup \{u_c\})^c, \mathbb{P}(y \mid u, x) = 0, \text{ and } \mathbb{P}(y \mid u_c, x) = \frac{\mathbb{P}(x, y) - \mathbb{P}(U \in \mathcal{U}')}{\mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U}')}.$$

CASE II: $\mathbb{P}(U \in \mathcal{U}' \cup u_c) \in [\mathbb{P}(x, y), \mathbb{P}(x)]$:

We choose

$$\mathbb{P}(U \in \mathcal{U}', x) = \mathbb{P}(U \in \mathcal{U}'), \mathbb{P}(U \in (\mathcal{U}' \cup \{u_c\})^c, x) = \mathbb{P}(x) - \mathbb{P}(U \in \mathcal{U}' \cup \{u_c\}), \text{ and } \mathbb{P}(u_c, x) = \mathbb{P}(u_c). \tag{161}$$

The conditional probability $\mathbb{P}(Y \mid U, X)$ is constructed as $\mathbb{P}(Y \mid U, X)$ is set as

$$\forall U \in \mathcal{U}', \mathbb{P}(y \mid u, x) = 1, \forall U \in (\mathcal{U}' \cup \{u_c\})^c, \mathbb{P}(y \mid u, x) = 0, \text{ and } \mathbb{P}(y \mid u_c, x) = \frac{\mathbb{P}(x, y) - \mathbb{P}(U \in \mathcal{U}')}{\mathbb{P}(u_c)}.$$

In both two cases, $\forall u \in U, \mathbb{P}(u, \neg x)$ is supplemented by $\mathbb{P}(u) - \mathbb{P}(u, x)$ based on (160) and (161). Additionally, $\forall u \in U$, we set $\mathbb{P}(y \mid u, \neg x) = \mathbb{P}(y \mid \neg x)$ and $\mathbb{P}(\neg y \mid u, x') = 1 - \mathbb{P}(y \mid u, x'), x' \in \{0, 1\}$.

It is easy to verify these three cases of constructions are non-negative and compatible with $\mathbb{P}(X, Y)$ and $\mathbb{P}(U)$. Moreover, it always holds $\forall u \in \mathbb{R}/\{u_c\}, \mathbb{P}(y, u, x) \wedge \mathbb{P}(\neg y, u, x) = 0$. According to this direct construction, we say $\mathcal{P}_{XYU}^{d_u-1} \neq \emptyset$.

■

Lemma M.3 (Justification of (122)-(123))**.** *The construction given by* (122) *is legitimate and satisfies*

$$\min \mathcal{P}^{\varepsilon}_{y|do(x)} \in \Big[\mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)), \mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) + \frac{2}{c}\varepsilon\Big],\ \max \mathcal{P}^{\varepsilon}_{y|do(x)} \in \Big[\mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) - \frac{2}{c}\varepsilon, \mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U))\Big],$$

where $\varepsilon \in [0, c]$, *where* $c = (\mathbb{P}(x, y_0) \wedge \mathbb{P}(x, y_1))/2d_u^2$ *is a constant.*

Proof.

Apparently, construction given by (122) is consistent with confounder distribution $\mathbb{P}(U)$. To demonstrate the legitimacy, we are only left with proving that the constructed $\mathbb{P}^*(Y, X, U)$ is always non-negative and compatible with observed $\mathbb{P}(X, Y)$. In order to achieve this goal, it is sufficient to verify $\sum_u \mathbb{P}^*(y', x, u) = \mathbb{P}(y', x)$ and $\mathbb{P}^*(x, u) \in [0, \mathbb{P}(u)]$ in (122).

The first part is easy to verify by summation. For brevity, we denote $u^*_{y'} := \arg\max_{y'}\{\mathbb{P}^*(u', x, y')\}, y' \in \{0, 1\}$:

$$\begin{aligned}
&\sum_{u\in\mathbb{R}} \mathbb{P}^*(y', x, u)\\
&= \sum_{u\in\mathcal{U}_\varepsilon} \mathbb{P}^*(y', x, u) + \mathbb{P}^*(y', x, u = u^*_{y'}) + \sum_{u\in(\mathcal{U}_\varepsilon\cup u^*_{y'})^c} \mathbb{P}^*(y', x, u)\\
&\overset{(122)}{=} \sum_{u\in\mathcal{U}_\varepsilon} \mathbb{P}(y' \mid x, u)\varepsilon + \mathbb{P}(y', x, u = u^*_{y'}) - \sum_{u\in\mathcal{U}_\varepsilon} \mathbb{P}(y' \mid u, x)(\varepsilon - \mathbb{P}(u, x)) + \sum_{u\in(\mathcal{U}_\varepsilon\cup u^*_{y'})^c} \mathbb{P}(y', x, u)\\
&= \sum_{u\in\mathcal{R}} \mathbb{P}(y', x, u) = \mathbb{P}(y', x).
\end{aligned} \tag{162}$$

For the second part, $\forall u \notin \{u^*_0, u^*_1\}$, we have that $\mathbb{P}^*(x, u) \in \{\varepsilon, \mathbb{P}(x, u)\} \subseteq [0, \mathbb{P}(u)]$. Otherwise, $\forall u \in \{u^*_0, u^*_1\}$, it could be verified that

$$\mathbb{P}^*(x, u) \in \Big[\mathbb{P}(x, u) - \sum_{u'\in\mathcal{U}_\varepsilon} (\varepsilon - \mathbb{P}(u', x)), \mathbb{P}(x, u)\Big] \subseteq [0, \mathbb{P}(u)]. \tag{163}$$

(163) is according to

$$\mathbb{P}(x, u) - \sum_{u'\in\mathcal{U}_\varepsilon} \big(\varepsilon - \mathbb{P}(u', x)\big) \geq \big(\mathbb{P}(x, y_0) \wedge \mathbb{P}(x, y_1)\big)/d_u - d_u\varepsilon \geq d_u\varepsilon > 0. \tag{164}$$

In combination with the above analysis, the construction $\mathbb{P}^{\mathcal{LB}}_\varepsilon(Y, X, U)$ is legitimate.

Stepping forwards, we aim to bound $\min \mathcal{P}^{\varepsilon}_{y|do(x)}$ and $\max \mathcal{P}^{\varepsilon}_{y|do(x)}$. According to the definition in (123), it directly holds that $\min \mathcal{P}^{\varepsilon}_{y|do(x)} \geq \mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U))$ and $\max \mathcal{P}^{\varepsilon}_{y|do(x)} \leq \mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U))$. On the other hand, under (122) we have that

$$\begin{aligned}
&\mathbb{P}^*(y \mid x, u)\mathbb{P}(u) - \mathbb{P}(y \mid x, u)\mathbb{P}(u) = 0,\ \ \forall u \notin \{u^*_0, u^*_1\}.\\
&\mid \mathbb{P}^*(y \mid x, u)\mathbb{P}(u) - \mathbb{P}(y \mid x, u)\mathbb{P}(u) \mid \leq \frac{d_u\varepsilon}{(\mathbb{P}(x, y_0) \wedge \mathbb{P}(x, y_1))/d_u - d_u\varepsilon} = \frac{\varepsilon}{2c - \varepsilon} \leq \frac{\varepsilon}{c},\ \ \forall u \in \{u^*_0, u^*_1\}.
\end{aligned} \tag{165}$$

Hence under the construction (122), we have

$$|\mathbb{P}^*(y \mid do(x)) - \mathbb{P}(y \mid do(x))| \leq \frac{2\varepsilon}{c}. \tag{166}$$

It indicates that

$$\min \mathcal{P}^{\varepsilon}_{y|do(x)} \leq \mathcal{LB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) + \frac{2}{c}\varepsilon,\ \max \mathcal{P}^{\varepsilon}_{y|do(x)} \geq \mathcal{UB}^{\text{mul}}_{x,y}(\mathbb{P}(U)) - \frac{2}{c}\varepsilon. \tag{167}$$

■

Lemma M.4 ((Rubin, 1981), Section 2)**.** *If we uniformly sample* $d_u - 1$ *points* $\{p_0, p_1, ...p_{d_u-1}\}$ *on the interval* $[0, 1]$ *and then re-order the* $d_u + 1$ *points* $\{0, p_0, p_1, ...p_{d_u-1}, 1\}$ *as*

$$p_{i(0)}, p_{i(1)}, ...p_{i(d_u)}.$$

Then the d_u-dimensional vector

$$(p_{i(1)} - p_{i(0)}, p_{i(2)} - p_{i(1)}, ...p_{i(d_u)} - p_{i(d_u-1)})$$

shares the same distribution with $\mathbb{P}(U)$, where $\mathbb{P}(U)$ is a uniformly sampled d_u-dimensional vector which is induced by $\{\mathbb{P}(U=0), \mathbb{P}(U=1), ...\mathbb{P}(U=d_u-1)\}$. Here $\sum_{i=0}^{d_u-1} \mathbb{P}(U=i) = 1$.

Lemma M.5 (Proof of (112))**.** *Consider the case $\mathbb{P}(U) \in (\mathcal{P}^L)^c$. If $\mathcal{U} \subseteq \mathbb{R}, t \in \mathcal{U}^c$ satisfies $\forall U \in \mathcal{U}$, $\mathbb{P}(\neg y, u, x) = 0$, $\forall (\mathcal{U} \cup \{t\})^c$, $\mathbb{P}(y, u, x) = 0$, then we have*

$$\mathbb{P}(U \in \mathcal{U}) \in \Big[\max\big\{0, \mathbb{P}(x) - \mathbb{P}(u_t)\big\}, \mathbb{P}(x,y) + \mathbb{P}(\neg x)\Big], \text{ where } t \in \mathcal{U}^c \subseteq \mathbb{R}.$$

Proof.

It holds that

$$\begin{aligned} \mathbb{P}(U \in \mathcal{U} \cup \{t\}) \geq \mathbb{P}(y, U \in \mathcal{U} \cup \{t\}, x) &\overset{(1)}{=} \mathbb{P}(y, U \in \mathcal{U} \cup \{t\}, x) + \mathbb{P}(y, U \in (\mathcal{U} \cup \{t\})^c, x) = \mathbb{P}(x, y). \\ \mathbb{P}(U \in \mathcal{U}) &\overset{(2)}{=} \mathbb{P}\Big(\{X, Y\} \neq \{x, \neg y\}, U \in \mathcal{U}\Big) \leq 1 - \mathbb{P}(x, \neg y) = \mathbb{P}(x,y) + \mathbb{P}(\neg x). \end{aligned} \tag{168}$$

Here (1)-(2) correspond to the properties $\forall u \in (\mathcal{U} \cup \{t\})^c$, $\mathbb{P}(y, u, x) = 0$ and $\forall u \in \mathcal{U}$, $\mathbb{P}(\neg y, u, x) = 0$, respectively. In combination with (168) and the fact $\mathbb{P}(U) \in (\mathcal{P}^L)^c$, the first conclusion could be strengthened as $\mathbb{P}(U \in \mathcal{U} \cup \{t\}) > \mathbb{P}(x)$. In sum, we derive that

$$\mathbb{P}(U \in \mathcal{U}) \in \Big[\max\big\{0, \mathbb{P}(x) - \mathbb{P}(u_t)\big\}, \mathbb{P}(x,y) + \mathbb{P}(\neg x)\Big], \text{ where } t \in \mathcal{U}^c \subseteq \mathbb{R}.$$

■

N. Auxiliary Experimental details

Due to the theoretical optimality of our *tight* identification region, additional experiments are, in general, not extremely necessary to provide further valuable information. This is the reason why we just focus on these two goals in our main text. We first show that the traditional entropy-based methods lose information compared with our oracle tight bound (Experiment N.1, N.2), then we additionally show that Proposition 4.7 is efficient (Experiment N.2); namely, it reveals more reliable information compared with traditional competitive bounds and additionally guide decision making.

N.1. Simulations

Experiment setting We follow the basic sampling method (Chickering & Meek, 2012) and replicate the setting in Jiang et al. (2023) to conduct Dirichlet sampling upon Figure 1. We assume that each generated data sample has only two parts of data information: $P(X,Y)$ and confounder information $\mathbb{P}(U)$. Specifically, we generate U with the same analogue as the previous: $U \sim Dir([0.1, 0.1, 0.1, 0.1, 0.1])$, $d_u = 5$. Moreover, following the famous sampling procedure (Chickering & Meek, 2012), X, Y is generated by $\mathbb{P}(X \mid u_i) \sim Dir(v'), \forall i = 0, ...d_u - 1$, $\mathbb{P}(Y \mid u_j, x_k) \sim Dir(s'), \forall j \in [0, 1, ...d_u - 1], k \in [0, 1, ...|X| - 1]$, where v' and s' are permutations of the vector $v := \frac{1}{\sum_1^{|X|} 1/i}[1, 1/2, 1/3, ...1/|X|]$ and $s := \frac{1}{\sum_1^{|Y|} 1/i}[1, 1/2, 1/3, ...1/|Y|]$ following Chickering & Meek (2012). Without loss of generalization, we consider the binary case; namely, $|X| = |Y| = 2$, and it is natural to extend to the multi-valued cases. For each sampling (10^6 in total), we select $\mathbb{P}(X,Y)$ and $\mathbb{P}(U)$ as our accessible data. We consider whether $\mathbb{P}(y' \mid do(x')), x', y' = 0, 1$ to be vanilla.

Experiment result We justify whether the PI region is vanilla according to the if and only if criteria (Theorem 4.1). We consider the case $H(U) \leq 1$ and separate it into ten groups corresponding to the confounder entropy $H(U) \in [i/10, i/10 + 0.1], i = 0, 1, ...0.9$. As illustrated in Table 5, our proposed PI bound is consistent with the ground truth blessed with its tightness guarantee. For comparison, the traditional entropy-based method (Jiang et al., 2023) exhibits an information loss. Such loss is significant when confounder entropy is relatively large. This is because for traditional entropy-based methods, $\mathbb{P}(y' \mid do(x'))$ degenerate to near $\mathbb{P}(y' \mid x')$ when the entropy is sufficiently small (smaller than the so-called "entropy threshold") although it is a relaxed optimization programming, which causes non-vanilla bound. In contrast, entropy-based methods lose this guarantee when the entropy is relatively large (which corresponds to many real-world scenarios). Our method, on the other hand, can accurately extract tight PI for any U-information and determine whether it is vanilla[6].

N.2. Real-world experiments

Experiment setting We also follow the setting of Jiang et al. (2023) for better comparison, where we reasonably simplify the graph within these two datasets into the paradigm of Figure 1, and we choose the same separating strategy of variables X, Y, U (Jiang et al., 2023). In the INSURANCE dataset, we treat car cost, property cost, and accident cost (other cars) as X, Y, U. Furthermore, in the ADULT dataset, we treat this triple as relationship (unmarried, in-family, etc), income and age. We follow the division method as before in Table 6. For instance, for "AGE" in the ADULT dataset, we choose 65 as the cutting point to separate it into two categories: "young" and "old". We treat other information as the protected feature but only the observations $\mathbb{P}(X, Y)$ and marginal information on the confounders are accessible. As we have comprehensively analyzed the quantitative performance of Li's bound (Li et al., 2023) in our main text (Theorem 3.4) and Appendix C, here we mainly focus on the comparison with Jiang et al. (2023) as our baseline.

Experiment result The results are shown in Table 6. Being blessed by our theoretical optimality, our PI bounds upon interventional probabilities are usually stricter than Jiang et al. (2023). Again, it validates our first goal. Noteworthy, there is no absolute guarantee under limited computational costs, as theoretical errors also exist when we compute approximate values based on the SSP problem, as mentioned in Theorem 4.5.

More importantly, for our second goal, apart from Table 6, we aim to argue the efficiency of our proposed valid ATE bound (Proposition 4.7). According to this proposition, we directly compute the valid ATE bound for the ADULT dataset across line $\{1, 2\}, \{3, 4\}, \{5, 6\}, \{7, 8\}, \{9, 10\}, \{11, 12\}$. We denote it as $\widehat{ATE}_{i-(i+1)}, i = 1, 3, 5, 7, 9, 11$. For instance, $\widehat{ATE}_{1-2} := \mathbb{P}(INCOME \le 50K \mid RELATIONSHIP = YES) - \mathbb{P}(INCOME \le 50K \mid RELATIONSHIP = NO)$. As illustrated in Table 7, our result via Proposition 4.7 is better than directly computing the upper and lower bounds of interventional probability and is also better (narrower) than Jiang's baseline; namely, our bound is more reliable. Noteworthy, compared with the baseline method (Tian & Pearl, 2000), Jiang et al. (2023) showed that $\widehat{ATE}_{11-12}$ is definitely greater than zero under this setting. It indicates the significant causal effect of the "relationship" to the high "income" among well-educated and "full-time" individuals. Stepping forward, our PI bound extends this observation to the "part-time" individuals. Namely, our bound via Proposition 4.7 additionally claims that $\widehat{ATE}_{7-8}$ is almost positive. It indicates the above-high-school and part-time individuals also exhibit a positive causal effect between "relationship" and "income". This phenomenon is in line with practical experience but has not been extracted in previous literature to our knowledge. This discovery will help guide relevant political and economic decision-making practices: we should advocate that the higher education population actively maintains their personal relationships and family situations in the pursuit of income, *regardless of whether full-time or part-time*, as "relationship" and "income" will have a positive causal relationship.

Similarly, for the INSURANCE dataset, we construct $\widehat{ATE}_{(2+i)-(1+i)}, \widehat{ATE}_{(3+i)-(2+i)}, \widehat{ATE}_{(3+i)-(1+i)}, i = 0, 3$. For instance, $\widehat{ATE}_{2-1} := \mathbb{P}(PROP\ COST = 100,000 \mid CAR\ COST = 100,000) - \mathbb{P}(PROP\ COST = 10,000 \mid CAR\ COST = 100,000)$. Our valid ATE bound (from Table 6 and Proposition 4.7) are also both stricter than the baseline.

N.3. Experiment 3

Finally, we also provide a visualization of the affiliation relationship in Corollary 4.3 (take $d_u = 3$ for brevity), which is a vivid supplement of Figure 3 in our manuscript.

[6] It is necessary to point out that when we only have estimations for confounder entropy but do not have knowledge of other side information, Jiang's method is effective. Our method sacrifices efficiency (by directly using confunder entropy) in exchange for accuracy (accurate vanilla-judgement for each possible U).

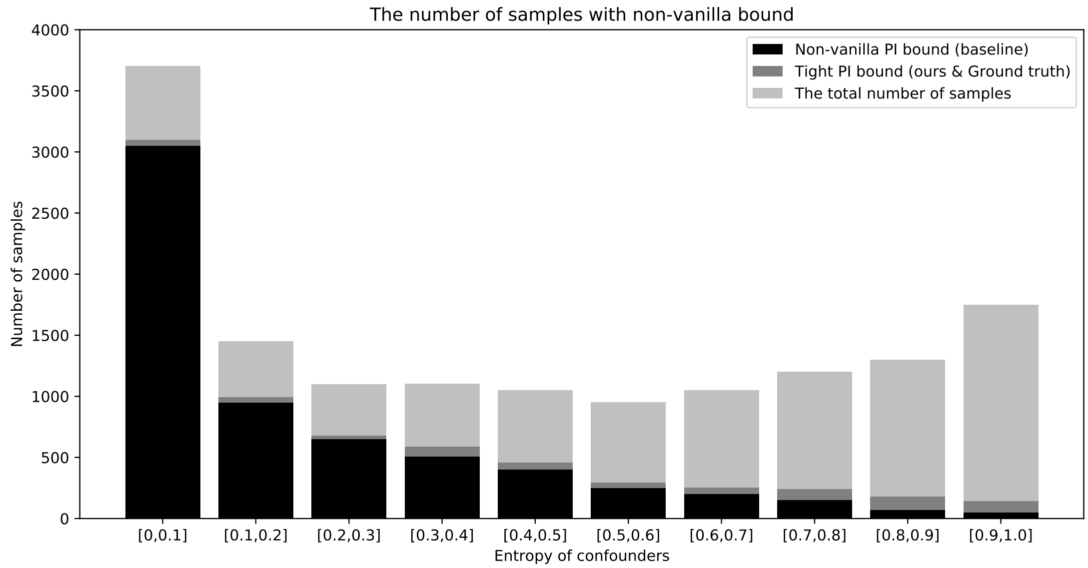

FIGURE 5: *Simulations (Experiment N.1). Tradition entropy-based optimization loss information of PI without taking full advantage of $\mathbb{P}(U)$, especially when $H(U)$ is relatively large, which is common in the real-world.*

dataset	SUBGROUP	X	Y	H(Z)	Baseline (Jiang et al., 2023)	Baseline (Tian & Pearl, 2000)	OUR BOUNDS
INSUR	UNDER 5000 MILES, NORMAL	CAR COST	PROP COST	ACCI			
		100,000	10,000	0.092	[0.000, 0.246]	[0.000, 0.800]	[0.000, 0.214]
		100,000	100,000	0.092	[0.699, 0.996]	[0.196, 0.996]	[0.703, 0.995]
		100,000	1,000,000	0.092	[0.004, 0.301]	[0.004, 0.804]	[0.004, 0.285]
		1,000,000	10,000	0.092	[0.000, 0.044]	[0.000, 0.249]	[0.000, 0.037]
		1,000,000	100,000	0.092	[0.000, 0.044]	[0.000, 0.249]	[0.000, 0.040]
		1,000,000	1,000,000	0.092	[0.956, 0.999]	[0.751, 0.999]	[0.964, 0.999]
ADULT		RELATIONSHIP	INCOME	AGE			
	BELOW HIGH SCHOOL, FULL-TIME	YES	$\leq$ 50K	0.21	[0.605, 0.934]	[0.423, 0.934]	[0.743, 0.924]
	BELOW HIGH SCHOOL, FULL-TIME	NO	$\leq$ 50K	0.21	[0.762, 0.985]	[0.496, 0.985]	[0.798, 0.982]
	BELOW HIGH SCHOOL, FULL-TIME	YES	>50K	0.21	[0.066, 0.395]	[0.066, 0.577]	[0.066, 0.388]
	BELOW HIGH SCHOOL, FULL-TIME	NO	>50K	0.21	[0.015, 0.238]	[0.015, 0.504]	[0.015,0.216]
	ABOVE HIGH SCHOOL, PART-TIME	YES	$\leq$ 50K	0.41	[0.186, 0.903]	[0.183, 0.903]	[0.192, 0.903]
	ABOVE HIGH SCHOOL, PART-TIME	NO	$\leq$ 50K	0.41	[0.779, 0.982]	[0.703, 0.983]	[0.832, 0.970]
	ABOVE HIGH SCHOOL, PART-TIME	YES	>50K	0.41	[0.017, 0.814]	[0.096, 0.817]	[0.097, 0.814]
	ABOVE HIGH SCHOOL, PART-TIME	NO	>50K	0.41	[0.017, 0.220]	[0.017, 0.297]	[0.017, 0.220]
	ABOVE HIGH SCHOOL, FULL-TIME	YES	$\leq$ 50K	0.12	[0.310, 0.664]	[0.250, 0.734]	[0.310, 0.664]
	ABOVE HIGH SCHOOL, FULL-TIME	NO	$\leq$ 50K	0.12	[0.725, 0.953]	[0.438, 0.953]	[0.752, 0.952]
	ABOVE HIGH SCHOOL, FULL-TIME	YES	>50K	0.12	[0.336, 0.690]	[0.266, 0.750]	[0.332, 0.677]
	ABOVE HIGH SCHOOL, FULL-TIME	NO	>50K	0.12	[0.046, 0.275]	[0.046, 0.562]	[0.048, 0.204]

TABLE 6: *Real-world experiment (Experiment N.2). Our proposed PI bounds are stricter than the competitive baselines.*

Dataset	ATE estimation	Baseline (Jiang et al., 2023)	Baseline (Tian & Pearl, 2000)	Our bounds via Table 6	Our bounds via Proposition 4.7
INSUR	$\widehat{ATE}_{2-1}$	[0.453,0.996]	[-0.604,0.996]	[0.489, 0.995]	[0.560, 0.995]
	$\widehat{ATE}_{3-1}$	[-0.242, 0.301]	[-0.796, 0.804]	[-0.210, 0.285]	[-0.210, 0.277]
	$\widehat{ATE}_{3-2}$	[-0.992, -0.398]	[-0.992, 0.608]	[-0.991, -0.418]	[0.991, 0.401]
	$\widehat{ATE}_{5-4}$	[-0.044, 0.044]	[-0.249, 0.249]	[-0.037, 0.040]	[-0.037, 0.038]
	$\widehat{ATE}_{6-4}$	[0.912, 0.999]	[0.502, 0.999]	[0.927, 0.999]	[0.920, 0.999]
	$\widehat{ATE}_{6-5}$	[0.912, 0.999]	[0.502, 0.999]	[0.924, 0.999]	[0.930, 0.999]
ADULT	$\widehat{ATE}_{1-2}$	[-0.380, 0.172]	[-0.562, 0.438]	[-0.239, 0.126]	[-0.218, 0.107]
	$\widehat{ATE}_{3-4}$	[-0.172, 0.38]	[-0.438, 0.562]	[-0.150,0.373]	[-0.102, 0.278]
	$\widehat{ATE}_{5-6}$	[-0.796, 0.124]	[-0.800,0.200]	[-0.778, 0.071]	[-0.770, 0.069]
	$\widehat{ATE}_{7-8}$	[-0.203, 0.797]	[-0.201,0.800]	[-0.123, 0.797]	[-0.003, 0.790]
	$\widehat{ATE}_{9-10}$	[-0.643, -0.061]	[-0.703, 0.296]	[-0.642, -0.088]	[-0.610, -0.102]
	$\widehat{ATE}_{11-12}$	[0.061, 0.644]	[-0.296, 0.704]	[0.128, 0.629]	[0.146, 0.602]

TABLE 7: *ATE estimation for the real-world dataset between the rows (Experiment N.2). The bounds from Table 6 means directly computing the difference between the lower (upper) bound of $\mathbb{P}(Y = 1 \mid do(x')), x' = 0, 1$. Compared with Jiang's bounds informatively indicating that $\widehat{ATE}_{11-12} > 0$, our proposed bounds additionally supplements that $\widehat{ATE}_{7-8}$ is almost positive. It indicates that the relationship significantly affects income among well-educated and high-income individuals, regardless of "full-time" or "part-time", which serves as our new observation..*

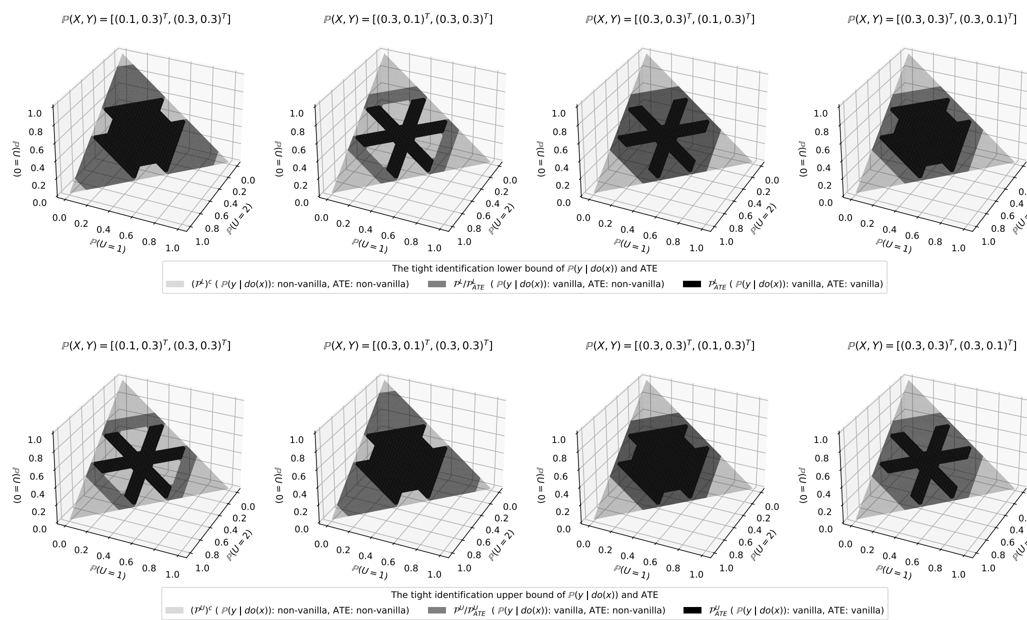

FIGURE 6: *Illustration on whether TPI would be vanilla (Theorem 4.1 under $d_u = 3$). Confounder information $\mathbb{P}(U)$ is shown as a 2-simplex, and each coordinate axis represents a $\mathbb{P}(u_i)$, where $i = 0, 1, 2$. Notice that $(\cdot)^c$ represents the complement of set $(\cdot)$, and $\mathbb{P}(X, Y) = [(\mathbb{P}(x, y), \mathbb{P}(x, \neg y))^T, (\mathbb{P}(\neg x, y), \mathbb{P}(\neg x, \neg y))^T], x = y = 1$. According to Theorem 4.1, we always have $\mathcal{P}^L_{ATE} \subseteq \mathcal{P}^L$, $\mathcal{P}^U_{ATE} \subseteq \mathcal{P}^U$, hence the total region of $\mathbb{P}(U)$ is separated into three disjoint partitions: $\{(\mathcal{P}^L)^c$, $\mathcal{P}^L/\mathcal{P}^L_{ATE}, \mathcal{P}^L_{ATE}\}$ or $\{(\mathcal{P}^U)^c$, $\mathcal{P}^U/\mathcal{P}^U_{ATE}, \mathcal{P}^U_{ATE}\}$. These practical examples clearly show that ATE is less prone to vanilla than $\mathbb{P}(y \mid do(x))$.*