Partial Identification with Proxy of Latent Confoundings via Sum-of-ratios Fractional Programming

Zhiheng Zhang[*1] **Xinyan Su**[2,3]

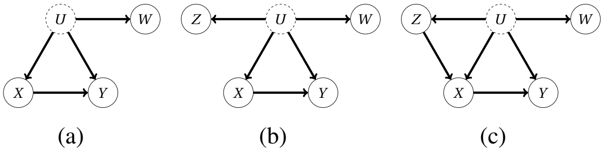
[1]Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, China
[2]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences, Beijing, China

Abstract

Causal effect estimation is a crucial theoretical tool in uncertainty analysis. The challenge of unobservable confoundings has raised concerns regarding quantitative causality computation. To address this issue, proxy control has become popular, employing auxiliary variables $\boldsymbol{W}$ as proxies for the confounding variables $\boldsymbol{U}$. However, proximal methods rely on strong assumptions, such as reversibility and completeness, that are challenging to interpret empirically and verify. Consequently, their applicability in real-world scenarios is limited, particularly when the proxies lack informativeness. In our paper, we have developed a novel optimization method named **P**artial **I**dentification with Proxy of Latent Confoundings via **S**um-of-Ratios **F**ractional **P**rogramming (PI-SFP). This method does not impose any additional restrictions upon proxies and only assumes the mild partial observability of the transition matrix $P(\boldsymbol{W} \mid \boldsymbol{U})$. We have theoretically proven the global convergence of PI-SFP to the valid bound of the causal effect and analyzed the conditions under which the bounds could be tight. Our synthetic and real-world experiments validate our theoretical framework.

1 INTRODUCTION

Causal inference is crucial in uncertainty analysis across various fields such as medicine [Castro et al., 2020], economics [Hicks et al., 1980, Zhang et al., 2023b, 2020], and education [Peng and Knowles, 2003]. However, extracting useful causal information from observational data is challenging due to latent confoundings that can impede statistical association studies [Pearl, 2009a]. To address this issue, researchers commonly rely on auxiliary variables for confounding adjustment. Representative auxiliaries include instrumental variables (IV) [Söderström and Stoica, 2002], proximal variables [Kuroki and Pearl, 2014, Tchetgen et al., 2020], or outcome-dependent variables [Gabriel et al., 2022].



(a) (b) (c)

FIGURE 1: *Causal identification with confoundings via single-proxy control (a) or double-proxy control (b,c). $\boldsymbol{W}$ or $\boldsymbol{Z}$ are so-called confonuder proxies when identifying the causal effect of treatment $\boldsymbol{X}$ on outcome $\boldsymbol{Y}$.*

In this paper, our primary focus is proximal causal identification, as it represents one of the most commonly utilized auxiliaries in our real world. The use of proxies for confounding adjustment has become a prominent topic of research both theoretically and empirically. Empirical work in this area dates back to [Wickens, 1972], which examined the potential benefits of proxies as an alternative to latent confounding in least square estimations. The concept has since been applied in observational studies such as Kolenikov and Angeles [2009], Wooldridge [2009] and further studied in other empirical works [Frost, 1979, Rothman et al., 2008]. On the other hand, theoretical research on this topic can be broadly categorized into two groups, which we illustrate in Fig.1: the "single-proxy scenario" (Figure1(a)) [Tchetgen et al., 2023, Park and Tchetgen, 2023] and the "double-proxy scenario" (Figures 1(b) and 1(c)) [Miao and Tchetgen, 2018, Shi et al., 2020, Singh, 2020, Kallus et al., 2021], where two confounding proxies $\boldsymbol{Z}, \boldsymbol{W}$ are available.

However, the application scope of proximal-based learning are still limited. For Figure 1(a), when both $\boldsymbol{W}$ and $\boldsymbol{U}$ are discrete random variables with a finite number of

[*] Correspondence author. Contacting Email: `zhiheng-20@mails.tsinghua.edu.cn`.

Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI 2024), PMLR 244:4123–4155.

choices, Pearl [2012], Tchetgen et al. [2023], Park and Tchetgen [2023] proved the point-wise identifiability of the true causal quantities when the proxies are informative enough, e.g., the probability transition matrix $P(\boldsymbol{W} \mid \boldsymbol{U})$ is fully observable and reversible, which is the so-called completeness assumption. Stepping forward, when $P(\boldsymbol{W} \mid \boldsymbol{U})$ is not observable, Pearl extended the point-wise double-proxy cases (Figures 1(b), 1(c)) from Cai and Kuroki [2012], where both the exposure proxy control $\boldsymbol{Z}$ and the outcome proxy control $\boldsymbol{W}$ exist. Regretfully, so-called "double-negative control" [Miao et al., 2018, Cui et al., 2020, Tchetgen et al., 2020, Deaner, 2018, Shi et al., 2020, Singh, 2020, Nagasawa, 2018] methods and "single-proxy control" methods [Pearl, 2012, Tchetgen et al., 2023, Park and Tchetgen, 2023] are still subject to strict bridge functions, completeness assumptions [1], or their weaker forms Ghassami et al. [2023], Kallus et al. [2021].

These untestable and impractical constraints surrogates an important motivation: excessively strong conditions on proxy are imposed to sufficiently achieve the point-wise value of the causal effect, which would be violated in general cases. For instance, in a recommender system, an item's exposure to users is often considered as treatment $\boldsymbol{X}(\boldsymbol{X} = 0, 1)$, while the observed feedback is treated as outcome $\boldsymbol{Y}$. In this process, the user's socio-economic status and the item characteristics are used as latent confounders [Sato et al., 2020], which affect both $\boldsymbol{X}$ and $\boldsymbol{Y}$, and some observations of confounders are considered as proxy (e.g., item popularity ranking) $\boldsymbol{W}$ [Zhang et al., 2023a]. Unfortunately, the measurement of $\boldsymbol{W}$ and $\boldsymbol{U}$ would be inaccurate, and even worse, $P(\boldsymbol{W} \mid \boldsymbol{U})$ is irreversible due to the low dimension of $\boldsymbol{W}$, rendering the previous approach invalid. Hence, a natural scientific question arises: *How to conduct partial identification via partially observed proxies?*

To address this question, due to its originality and difficulty of double-proxies collection, we mainly focus on the single proxy case (Fig.1(a)), which is mostly related to Pearl [2012], Tchetgen et al. [2023], Park and Tchetgen [2023], and our method could be naturally generalized to Fig.1(b)-Fig.1(c). We weaken the requirement of "total precise observability" of $P(\boldsymbol{W} \mid \boldsymbol{U})$ to "partial observability", and thus generalize the point-wise identification into a partial identification. Our method contributes to the traditional fractional programming methods [Stancu-Minasian, 2012] since we do not rely on strong concavity assumption. More importantly, our method also advances the state-of-the-art constrained-optimization-based literature upon partial identification Duarte et al. [2023], Li and Pearl [2022], since we additionally provide a non-parametric convergence rate to causal queries based on branch-and-bound strategy. Our contributions are summarized as follows:

- We generalize the traditional proximal learning literature from point-wise identification to partial identification, with a more reasonable and weaker partial observability assumption.
- We introduce a global optimization strategy called PI-SFP and theoretically prove that it can globally converge to the valid bound of casual queries. Moreover, we justify whether the bound is tight. Synthetic and real-world experiments have demonstrated our findings.
- We theoretically justify the necessity of the partial observability assumption in proximal control. It is supported by the negative result that traditional informative proxies might not sufficiently guarantee informative partial identification.

2 PRELIMINARIES AND FRAMEWORK

The concept of causal effect is closely linked with the 'do' operator, which can be viewed as an external intervention [Pearl et al., 2000, Pearl, 2009b]. In particular, the causal effect of treatment $\boldsymbol{X}$ on outcome $\boldsymbol{Y}$ is represented as $f(y \mid do(x))$ in Fig.1, where $do(x)$ signifies that the treatment $\boldsymbol{X}$ is fixed at a specific value x, and $f(\cdot)$ denotes the probability mass/density function for discrete/continuous variables. We use d_u to denote the cardinality of confounders. As per the back-door criteria [Pearl et al., 2000], the identification of $f(y \mid do(x))$ is given by $\sum_{i=1}^{d_u} f(y \mid u_i, x) f(u_i)$, namely $f(y, x) + \sum_{i=1}^{d_u} f(y, u_i, x) f(u_i, \neg x)/f(u_i, x)$.

Such decomposition results from $f(u_i) = f(u_i, x) + f(u_i, \neg x)$. In Kuroki and Pearl [2014], the authors assumed that the transition matrix $P(\boldsymbol{W} \mid \boldsymbol{U})$ is observable and reversible, and hence claimed that $f(y \mid do(x))$ is identifiable. In other words, the value of each item as above can be explicitly extracted as follows[2]:

$$\begin{bmatrix} f(y, \boldsymbol{U}, x) \\ f(\boldsymbol{U}, x) \\ f(\boldsymbol{U}, \neg x) \end{bmatrix} = P(\boldsymbol{W} \mid \boldsymbol{U})^{-1} \begin{bmatrix} f(y, \boldsymbol{W}, x) \\ f(\boldsymbol{W}, x) \\ f(\boldsymbol{W}, \neg x) \end{bmatrix}. \quad (1)$$

Our paper focuses on generalization, where we consider the partial identification of $f(y \mid do(x))$ instead of its unique form computation. This change stems from our relaxed assumption on $P(\boldsymbol{W} \mid \boldsymbol{U})$, where we move from total observability to partial observability and remove the guarantee

[1] The previous reversibility assumption of $P(\boldsymbol{W} \mid \boldsymbol{U})$ was strengthened to that of $P(\boldsymbol{Z}, \boldsymbol{W} \mid x)$ and $P(y, \boldsymbol{Z}, \boldsymbol{W} \mid x)$. Moreover, the path $\boldsymbol{W}$ to $\boldsymbol{Y}$ in Fig. 1(c) can be additionally permitted.

[2] In our paper, we use bold letters to denote column vectors of corresponding possible values. For instance, $f(y, \boldsymbol{U}, x) = [f(y, u_1, x), f(y, u_2, x), ...f(y, u_{d_u}, x)]^T$. Furthermore, if a symbol has two bold letters such as $P(\boldsymbol{W} \mid \boldsymbol{U})$, it denotes the matrix $[f(\boldsymbol{W} \mid u_1), f(\boldsymbol{W} \mid u_2), ...f(\boldsymbol{W} \mid u_{d_u)})]$, where $f(\boldsymbol{W} \mid u_i) = [f(w_1 \mid u_i), f(w_2 \mid u_i), ...f(w_{d_w} \mid u_i)]^T, i = 1, 2, ...d_u$.

for reversibility (thus $P(\boldsymbol{W} \mid \boldsymbol{U})^{-1}$ in Eqn (2) may not exist). Specifically, we expand the identification region of $P(\boldsymbol{W} \mid \boldsymbol{U})$ from a fixed distribution to the family $\mathscr{P}$, which contains all possible $P(\boldsymbol{W} \mid \boldsymbol{U})$ such that $P(\boldsymbol{W} \mid \boldsymbol{U}) - \underline{P(\boldsymbol{W} \mid \boldsymbol{U})}$ and $\overline{P(\boldsymbol{W} \mid \boldsymbol{U})} - P(\boldsymbol{W} \mid \boldsymbol{U})$ all both non-negative. Here $\underline{P(\boldsymbol{W} \mid \boldsymbol{U})}$ and $\overline{P(\boldsymbol{W} \mid \boldsymbol{U})}$ are two priori known matrices to bound $P(\boldsymbol{W} \mid \boldsymbol{U})$. This scenario is prevalent in the real world. While studies [Kuroki and Pearl, 2014, Greenland, 2005] have generally confirmed its verifiability, recent literature has not fully explored it. In our paper, we reiterate the condition $P(\boldsymbol{W} \mid \boldsymbol{U}) \in \mathscr{P}$ as the 'partial observability assumption' in our following text. Under this assumption, we can naturally set our original goal as seeking the lower bound of $f(y \mid do(x))$ (upper bound is symmetric) by solving the following partial identification problem: $f(y,x) + \min \sum_{i=1}^{d_u} f(y,u_i,x)f(u_i,\neg x)/f(u_i,x)$, subject to $f(y,\boldsymbol{W},\boldsymbol{U},\boldsymbol{X}) \in \mathcal{F}$. Here $f(y,\boldsymbol{W},\boldsymbol{U},\boldsymbol{X})$ is a three-order $d_w * d_u * d_x$ tensor indicating the joint probability distribution of each $w \in \boldsymbol{W}, u \in \boldsymbol{U}, x \in \boldsymbol{X}$ together with $\boldsymbol{Y} = y$. The set $\mathcal{F} = \{f(y,\boldsymbol{W},\boldsymbol{U},\boldsymbol{X}) : f(y,\boldsymbol{W},\boldsymbol{U},\boldsymbol{X})$ is compatible with $P(\boldsymbol{W} \mid \boldsymbol{U}) \in \mathscr{P}\}$.

Achieving this goal is challenging due to the difficulty in achieving its tight bound. Firstly, the feasible region $\mathcal{F}$ is challenging to represent in a closed form due to the boundary constraints that include the partially observable $P(\boldsymbol{W} \mid \boldsymbol{U})$. This constraint can be seen as a first-kind Fredholm integral equation, which is an ill-posed problem when $P(\boldsymbol{W} \mid \boldsymbol{U})$ is irreversible[3]. To address this issue, we propose relaxing the feasible region from $\mathcal{F}$ to $\widetilde{\mathcal{F}}$ ($\mathcal{F} \subseteq \widetilde{\mathcal{F}}$), which contains a closed-form expression. Specifically, the relaxed condition $f(y,\boldsymbol{W},\boldsymbol{U},\boldsymbol{X}) \in \widetilde{\mathcal{F}}$ ensures that the feasible regions of $f(y,\boldsymbol{U},x)$, $f(y,\boldsymbol{U},x)$, and $f(\boldsymbol{U},\neg x)$ can be represented in a calculable closed-form. We refer to these as $IR_{F(y,U,x)}$, $IR_{F(U,x)}$, and $IR_{F(U,\neg x)}$, respectively, in our final objective function in the following section.

Secondly, even if we retreat and seek its valid bound as above, it remains non-trivial due to the difficulty in finding an appropriate optimization method. Since the causal effect is expressed as a fractional summation, it is natural to explore techniques in sum-of-ratios fractional programming (SFP). The general form of SFP, as summarized in [Schaible and Shi, 2003], is represented as follows: $\min\{\sum_{i=1}^{M} \frac{g_{1i}(\boldsymbol{\phi})}{g_{2i}(\boldsymbol{\phi})}\}, \boldsymbol{\phi} \in S\}, g_{1i}(\boldsymbol{\phi})$ is convex, $g_{2i}(\boldsymbol{\phi})$ is concave, $g_{1i}(\boldsymbol{\phi}), g_{2i}(\boldsymbol{\phi}) > 0$. Here S is a convex set, and $M \geq 2$ is a integer. In order to ensure the global nature of optimal solutions, $g_{1i}(\Phi)$ and $g_{2i}(\Phi)$ are assumed to be convex and concave, respectively. In contrast with the above formulation, we should choose $\boldsymbol{\phi} = ((f(y,u_i,x),...)^T, (f(u_i,x),...)^T, (f(u_i,\neg x),...)^T)$, and $g_{1i}(\boldsymbol{\phi}) = f(y,u_i,x)f(u_i,\neg x), g_{2i}(\boldsymbol{\phi}) = f(u_i,x)$. Moreover, $M = d_u, i = 1,2,...d_u$. However, this construction violates the traditional convex-concave assumption, as $g_{1i}(\boldsymbol{\phi})$ is not convex. Therefore, here traditional SFP algorithms [Schaible and Shi, 2003] are not suitable.

To address these two challenges, we introduce the Partial Identification with Sum-of-Ratios Fractional Programming (PI-SFP) algorithm. Motivated by the branch and bound strategy [Dai et al., 2005, Lawler and Wood, 1966, Dur et al., 2001, Pei and Zhu, 2013] and DC programming [Horst and Thoai, 1999, Tao and An, 1997, Pei and Zhu, 2013], our algorithm iteratively searches the optimal bound through feasible region partitioning. Different from the closely-related partial identification literature Duarte et al. [2023], Li and Pearl [2022], we also provide a comprehensive new convergence analysis.

For supplement, due to these two challenges of solving the partial observability case, another line of recent literature has avoided further discussion on the observability of $P(\boldsymbol{W} \mid \boldsymbol{U})$. Instead, researchers have introduced an auxiliary variable $\boldsymbol{Z}$ and formalized the problem as the double negative control [Miao et al., 2018, Cui et al., 2020, Tchetgen et al., 2020, Deaner, 2018, Shi et al., 2020, Singh, 2020, Nagasawa, 2018, Kallus et al., 2021]. However, as illustrated in the introduction, there are no free lunch (Table. 2). These works are restricted by additional assumptions about proxies, such as the completeness condition and the bridge function condition. Importantly, all of these works still rely on the reversibility of $P(\boldsymbol{W} \mid \boldsymbol{U})$, except for Ghassami et al. [2023], Kallus et al. [2021], who substituted it as a weaker bridge function condition. Consequently, even if the transition matrix is still reversible, but with a large conditional number[4], numerical computations would already become extensively overwhelming. In conclusion, revisiting single-proxy control under the partial observability of $P(\boldsymbol{W} \mid \boldsymbol{U})$ is not only challenging but also necessary.

3 METHOD

We begin by presenting definitions and assumptions and then rigorously formulate the objective function.

DEFINITIONS We denote $Y, Y_0, Y_1 \in [Y^L, Y^U], Z \in [Z^L, Z^U], X \in [X^L, X^U], W \in [W^L, W^U], U \in [U^L, U^U]$. Moreover, we use d_z, d_u, d_w, d_x to denote the cardinality of variables $\boldsymbol{Z}, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{X}$. For instance, the set of confounder $\boldsymbol{U}$ is $\{u_1, u_2, ...u_{d_u}\}$. $X \neq x$ is simplified as $\neg x$. Y can be discrete or continuous.

Moreover, Y_x is the value of $\boldsymbol{Y}$ when X is forced to be x.

[3] Otherwise, the closed-form expression of $\mathcal{F}$ can only be approximated iteratively by complex numerical methods [Strand and Westwater, 1968], which is beyond our scope.

[4] The conditional number of matrix A is denoted as $\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$, where $\sigma_{max}(A)$ and $\sigma_{min}(A)$ denote the maximal and minimal singular values of A. If some rows/columns of A are similar (or equal), then $\kappa(A)$ is large (or $+\infty$), and A^{-1} is computationally hard (or may not exist).

$\pi(x)$ is a weight function[5] of $\boldsymbol{X}$. On this basis, $ACE_{\boldsymbol{X}\to\boldsymbol{Y}}$ denotes the average causal effect (ACE) from $\boldsymbol{X}$ to $\boldsymbol{Y}$, namely that $ACE_{\boldsymbol{X}\to\boldsymbol{Y}} = \int_x \int_y y f(Y_x = y)\pi(x)dxdy$.

Assumption 1 *(partial observability)* $P(\boldsymbol{W} \mid \boldsymbol{U}) \in \mathscr{P}$.

Here the set $\mathscr{P}$ is identified in Section 2, where $\underline{P(\boldsymbol{W} \mid \boldsymbol{U})}$ and $\overline{P(\boldsymbol{W} \mid \boldsymbol{U})}$ are two a priori known matrices serving as the partial order of $P(\boldsymbol{W} \mid \boldsymbol{U})$.

OBJECTIVE FUNCTION The objective of this section is to formalize the optimization problem of the single proxy control under Assumption. 1. During this process, we aim to tackle the two challenges introduced in the preliminaries. Our primary objective is (the maximum case is symmetric):

$$\begin{aligned} &\min f(y,x) + \sum_{i=1}^{d} \frac{f(y,u_i,x)f(u_i,\neg x)}{f(u_i,x)}, \\ &\text{subject to: } f(y,\boldsymbol{U},\boldsymbol{W},\boldsymbol{X}) \in \mathcal{F}. \end{aligned} \tag{2}$$

Here $\mathcal{F} = \{f(y,\boldsymbol{U},\boldsymbol{W},\boldsymbol{X}) : f(y,\boldsymbol{U},\boldsymbol{W},\boldsymbol{X})$ is compatible with Assumption. 1 and observed $f(y,\boldsymbol{W},\boldsymbol{X})\}$.

As we suggested in the preliminaries, *the first challenge* is the nonexistence of closed-form expression of $\mathcal{F}$. To solve it, we introduce the new symbol $\widetilde{\mathcal{F}}$ to formally describe the relaxation of the identification region of $f(y,\boldsymbol{U},\boldsymbol{W},\boldsymbol{X})$. For preparation, we introduce the symbol $\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\omega}$ and follow the previous notation $\boldsymbol{\phi}$: $\theta_i = f(y,u_i,x), \psi_i = f(u_i,x), \omega_i = f(u_i,\neg x), \boldsymbol{\theta} = (\theta_1,\theta_2,...\theta_d)^T, \boldsymbol{\psi} = (\psi_1,\psi_2,...\psi_d)^T, \boldsymbol{\omega} = (\omega_1,\omega_2,...\omega_d)^T,\ \boldsymbol{\phi} = (\boldsymbol{\theta}\ \boldsymbol{\psi}\ \boldsymbol{\omega})$.

Then we construct a broader set $\widetilde{\mathcal{F}}$ as follows:

$$\widetilde{\mathcal{F}} = \left\{f(y,\boldsymbol{U},\boldsymbol{W},\boldsymbol{X}) : \boldsymbol{\phi} \in IR_{\boldsymbol{\Phi}}, IR_{\boldsymbol{\Phi}} = IR^1_{\boldsymbol{\Phi}} \cap IR^2_{\boldsymbol{\Phi}}\right\},$$

where the set $IR^1_{\boldsymbol{\Phi}}$ denotes the set of $\boldsymbol{\Phi}$ satisfies $\overline{P(\boldsymbol{W} \mid \boldsymbol{U})}\boldsymbol{\phi} \geq f(y,\boldsymbol{W},x), f(\boldsymbol{W},x), f(\boldsymbol{W},\neg x) \geq \underline{P(\boldsymbol{W} \mid \boldsymbol{U})}\boldsymbol{\phi}$. Here we use $\boldsymbol{S}_1 \geq \boldsymbol{S}_2$ to denote $\boldsymbol{S}_1 - \boldsymbol{S}_2$ is a non-negative matrix, and $\boldsymbol{I_{d*d}}$ denotes the $d*d$ identity matrix. Moreover, the set $IR^2_{\boldsymbol{\Phi}}$ indicates all possible $\boldsymbol{\phi}$ such that $\mathbf{1} * \boldsymbol{\theta} = f(y,x), \mathbf{1} * \boldsymbol{\phi} = f(x), \mathbf{1} * \boldsymbol{\omega} = f(\neg x)$, and $\forall i, \theta_i \in [0, f(y,x)], \phi_i \in (0, f(x)], \omega_i \in [0, f(\neg x)]$.

Here $\mathbf{1}$ denotes the corresponding all-ones vector. By this construction, the enclosure property $\mathcal{F} \subseteq \widetilde{\mathcal{F}}$ is guaranteed.

Proposition 1 *$\mathcal{F}$ is enclosed by $\widetilde{\mathcal{F}}$, namely that $\mathcal{F} \subseteq \widetilde{\mathcal{F}}$.*

The proof is shown in the Appendix A.1. Proposition. (1) provides the extension of the feasible region of $f(y,\boldsymbol{U},\boldsymbol{W},\boldsymbol{X})$ from $\mathcal{F}$ to $\widetilde{\mathcal{F}}$. On this basis, Eqn (2) is relaxed as follows: $\underline{f(Y_x = y)} = \min f(y,x) + \sum_{i=1}^{d} \theta_i\omega_i/\psi_i$, subject to: $f(y,\boldsymbol{U},\boldsymbol{W},\boldsymbol{X}) \in \widetilde{\mathcal{F}}, i.e., \boldsymbol{\phi} \in IR_{\boldsymbol{\Phi}}$.

Symmetrically, the optimal value is denoted as $\overline{f(Y_x = y)}$ for the maximum case. Moreover, the corresponding set of optimal solutions are denoted as $\boldsymbol{\Phi_{opt}}$. The following proposition discuss the tightness of $\underline{f(Y_x = y)}$:

Proposition 2 *The outcome $\underline{f(Y_x = y)}$ serves as the valid lower bound of $f(Y_x = y)$. Moreover, this bound is tight if and only if the following set is not empty: $\Big\{f(y,\boldsymbol{U},\boldsymbol{W},\boldsymbol{X}) : f(y,\boldsymbol{U},\boldsymbol{W},\boldsymbol{X}) \in \mathcal{F}$ and is compatible with some $\boldsymbol{\phi_{opt}} \in \boldsymbol{\Phi_{opt}}\Big\} \neq \emptyset$, where $\boldsymbol{\phi_{opt}}$ is an element of the set $\boldsymbol{\Phi_{opt}}$. The maximum case $\overline{f(Y_x = y)}$ is symmetric.*

As discussed in the preliminaries, we have explained the reasons why guaranteeing that $\underline{f(Y_x = y)}$ is a tight bound is beyond the community scope, hence Proposition 2 is already the optimal result. In practice, it could be verified that Eqn (2) holds in a number of cases, whose details are detailed in Appendix A.2.

We now aim to tackle *the second challenge*: the non-trivial nature of the fractional programming problem due to the invalidation of the convex-concave condition. To address this issue, we adopt the *difference-in-convex (DC)* decomposition strategy to formally describe how we relax the above formulation into a linear programming problem. To prepare for this, we first transform the fractional form and introduce the knockoff variable $\boldsymbol{\psi^o}$ to replace the denominator. Next, we introduce the $4d-$ dimensional vector $\boldsymbol{\gamma}$: $\boldsymbol{\psi^o} = (\psi^o_1, \psi^o_2, ...\psi_d)^T,\ \boldsymbol{\gamma} = \left((\boldsymbol{\psi^o})^T, \boldsymbol{\theta}^T, \boldsymbol{\psi}^T, \boldsymbol{\omega}^T\right)^T$, where $(\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\omega})$ is copied from $\boldsymbol{\phi}$.

Then our original function is equivalently transformed to

$$\begin{aligned} &\underline{f(Y_x = y)} = \min f(y,x) + \sum_{i=1}^{d} \psi^o_i\theta_i\omega_i \\ &\text{subject to} : \boldsymbol{\gamma} \in IR_\Gamma := \{\boldsymbol{\gamma} : \boldsymbol{\phi} \in IR_\Phi, \psi^o_i\psi_i = 1\}. \end{aligned} \tag{3}$$

Here $i = 1, ...d$. Although the knock-off trick has been implemented, achieving the final goal remains challenging in practice. Firstly, the objective function and constraints are both non-convex and nonlinear. Secondly, relying on local optimal algorithms alone is not viable, as it may not guarantee the validity of the bound $\underline{f(Y_x = y)}$. Thus, our motivation is to construct a weaker linear programming form that can approximate the global optimal value of (3). To achieve this, we propose applying the *difference-in-convex (DC)* decomposition as our core idea: $\forall\boldsymbol{\gamma}$,

$$\sum_{i=1}^{d} \psi^o_i\theta_i\omega_i = C_1(\boldsymbol{\gamma}) - C_2(\boldsymbol{\gamma}),\ \psi^o_i\psi_i = D_{i1}(\boldsymbol{\gamma}) - D_{i2}(\boldsymbol{\gamma}), \tag{4}$$

[5] In [Kallus et al., 2021], it is called as generalized average causal effect. It can degenerate to the traditional form [Pearl, 2013] as $ACE_{\boldsymbol{X}\to\boldsymbol{Y}} = E(Y_1) - E(Y_0)$ if we choose $d_x = 2, X = \{0,1\}$, and $\pi(x) = \boldsymbol{sgn}(x)$, where $\boldsymbol{sgn}(\cdot)$ is the sign function.

where $i = 1, 2, \cdots, d, C_1(\boldsymbol{\gamma}), C_2(\boldsymbol{\gamma}), D_{i1}(\boldsymbol{\gamma}), D_{i2}(\boldsymbol{\gamma})$ [6] are all convex functions (see Appendix A.5) satisfying that $C_1(\boldsymbol{\gamma}), C_2(\boldsymbol{\gamma}), D_{i1}(\boldsymbol{\gamma}) :=$

$$\begin{aligned}&\sum_{i=1}^{d}\frac{1}{6}(\sum_{cyc}\psi_i^o)^3+\frac{1}{2}\sum_{cyc}(\psi_i^o)^4+\frac{1}{2}\sum_{cyc}(\psi_i^o)^2,\\&\sum_{i=1}^{d}\frac{1}{6}\sum_{cyc}(\psi_i^o)^3+\frac{1}{4}\sum_{cyc}[(\psi_i^o)^2+\theta_i]^2+[\psi_i^o+\theta_i^2]^2,\\&\frac{1}{2}(\psi_i^o+\psi_i)^2,\ D_{i2}(\boldsymbol{\gamma})=\frac{1}{2}[(\psi_i^o)^2+(\psi_i)^2],\end{aligned}\tag{5}$$

respectively. Exploiting their convexity, we bound them by the following linear functions, which are constructed by secants and tangents of the original function:

$$\begin{aligned}&C_1(\boldsymbol{\gamma})-C_2(\boldsymbol{\gamma})\geq C_1^{\tan}(\boldsymbol{\gamma})-C_2^{\sec}(\boldsymbol{\gamma}),\\&D_{i1}(\boldsymbol{\gamma})-D_{i2}(\boldsymbol{\gamma})\in[D_{i1}^{\tan}(\boldsymbol{\gamma})-D_{i2}^{\sec}(\boldsymbol{\gamma}),D_{i1}^{\sec}(\boldsymbol{\gamma})-D_{i2}^{\tan}(\boldsymbol{\gamma})]\end{aligned}\tag{6}$$

For their explicit form solutions, we refer the readers to (10). This allows us to relax the original problem in (3) into the following linear program:

$$\begin{aligned}&\min f(y,x)+C_1^{\tan}(\boldsymbol{\gamma})-C_2^{\sec}(\boldsymbol{\gamma})\\&\text{subject to : }\boldsymbol{\phi}\in IR_{\Phi},D_{i1}^{\tan}(\boldsymbol{\gamma})-D_{i2}^{\sec}(\boldsymbol{\gamma})\leq 1,\\&D_{i1}^{\sec}(\boldsymbol{\gamma})-D_{i2}^{\tan}(\boldsymbol{\gamma})\geq 1,i=1,2,\cdots,d.\end{aligned}\tag{7}$$

It is clear that this shift causes the estimation error. In order to eliminate it in practice, we iteratively do DC within simplicial partitioned feasible regions. Details will be shown in the following section. In conclusion, we already address these two challenges in the preliminaries. In addition, Our framework is natural to extend to the ACE cases, and we refer readers to Appendix A.4 for details.

4 ALGORITHM

In this section, we demonstrate how to compute $\underline{f(Y_x = y)}$ in (3) practically. As mentioned earlier, this involves optimizing a non-convex function, which requires new optimization techniques to find the global optimum. To this end, we propose Partial Identification via Sum-of-ratios Fractional Programming (PI-SFP), a fractional programming-based method that optimizes the objective through iterative approximation. Specifically, we begin by constructing a simplex S_0 that encloses the feasible region of (3). We then use S_0 to identify a lower bound of $\underline{f(Y_x = y)}$ using the difference-in-convex (DC) decomposition strategy. In each iteration, we partition S_0 into multiple simplices to refine the lower bound constructed in the initial step.

Algorithm 1: Partial Identification via Sum-of-ratios Fractional Programming (PI-SFP).

```
Input: Observational distribution $f(y, \boldsymbol{W}, \boldsymbol{X})$,
    $\underline{P(\boldsymbol{W} \mid \boldsymbol{U})}$, $\overline{P(\boldsymbol{W} \mid \boldsymbol{U})}$, a prespecified error
    bound $\delta > 0$.
Output: A lower bound estimate $\underline{f_{opt}^k(Y_x = y)}$.
1 Let $k = 0$, construct an original simplex $S_0 =$
    Initialization($f(y, \boldsymbol{W}, \boldsymbol{X}), \underline{P(\boldsymbol{W} \mid \boldsymbol{U})}, \overline{P(\boldsymbol{W} \mid \boldsymbol{U})}$);
2 Calculate a lower bound of $\underline{f_{S_0}(Y_x = y)}$ via the
    Bounding function: $\underline{\underline{f_{S_0}(Y_x = y)}} =$ Bounding($S_0$);
3 Set the collection of simplices at the 0-th iteration as
    $\mathcal{S}_0 = \{S_0\}$;
4 while $PI\text{-}SFP_{error} \leq \delta$ do
5     Let $\tilde{S}_k = \text{argmin}_{S \in \mathcal{S}_k} \underline{\underline{f_S(Y_x = y)}}$, where
        $\underline{\underline{f_S(Y_x = y)}}$ denotes the output of Bounding($S$)
        with input $S$;
6     Split $\tilde{S}_k$ into two simplicies $\tilde{S}_{k1}$ and $\tilde{S}_{k2}$ via the
        Bisection function: $\tilde{S}_{k1}, \tilde{S}_{k2} =$ Bisection($\tilde{S}_k$)
        and set $\mathcal{S}_{k+1} = \left(\mathcal{S}_k \setminus \tilde{S}_k\right) \cup \{\tilde{S}_{k1}, \tilde{S}_{k2}\}$;
7     Calculate the estimation error bound via
        $PI\text{-}SFP_{error}$ = Global_error($\tilde{S}_0, \tilde{S}_1, \cdots, \tilde{S}_{k+1}$);
8     Set $k = k + 1$;
9 Return $\underline{f_{opt}^k(Y_x = y)} = \max_{i \in \{0,1,\ldots k\}} \underline{\underline{f_{\tilde{S}_i}(Y_x = y)}}$.
```

The remainder of this section is organized as follows. In Section. 4.1, we introduce the main framework of our algorithm, which we divided into four modules: 1) **Initialization**, 2)**Bisection**, 3) **Bounding**, and 4) **Global_error**. We then elaborate on each module in detail in Section. 4.2. For ease of notation, we introduce the following symbols for algorithm description: • For a simplex S, $dia(S) := \max_{s_1,s_2 \in S} \|s_1 - s_2\|_2$ denotes its diameter, and S^i denotes its $i-$th supporting vector, $i = 0, 1, ...4d$. • $\underline{f_S(Y_x = y)}$ denotes the optimal value of (3) when its feasible region is strengthened to $\boldsymbol{\gamma} \in IR_\Gamma \cap S$.

4.1 OVERVIEW OF PI-SFP

The framework of PI-SFP to solve (3) is as follows.

Step 1 involves pre-processing, where the **Initialization** function is used to construct a baseline simplex S_0 enclosing the original feasible region, such that $IR_\Gamma \subseteq S_0$ and $\underline{f(Y_x = y)} = \underline{f_{S_0}(Y_x = y)}$ (see Lemma 2 in Appendix A.5). This equivalent transformation enables the computation of $\underline{f(Y_x = y)}$ via $\underline{f_{S_0}(Y_x = y)}$. In **Step 2**, we use the DC decomposition strategy as in (4)-(7) to find a lower bound of $\underline{f_{S_0}(Y_x = y)}$, which is denoted by $\underline{\underline{f_{S_0}(Y_x = y)}}$.

In **Steps 4-9**, we employ a bisection-like approach to iteratively partition S_0 into a set of simplices $\mathcal{S}_k$ in the k-th

[6] Note that the sub-script cyc in (5) is an abbreviation of cyclic sum following [Du et al., 2012], which cycles through $\{\psi_i^o, \theta_i, \omega_i\}$ in the corresponding function and take the sum. For instance, we have $\sum_{cyc}[\psi_i^o + \theta_i^2]^2 = [\psi_i^o + \theta_i^2]^2 + [\theta_i + \omega_i^2]^2 + [\omega_i + (\psi_i^o)^2]^2$.

Algorithm 2: Recursive procedure to split the simplicial partitions (**Bisection**).

```
Input: Simplex $S$ with vertices $\{S^0, \cdots, S^{4d}\}$.
Output: Two new simplices $S_1, S_2$.
Set $S^{t_1}, S^{t_2}$ as the vertices incident to the longest edge
    of $S$: $\{t_1, t_2\} = \underset{\{a,b\} \in \{0,1,\cdots,4d\}}{\text{argmax}} \|S^a - S^b\|_2$;
Construct $S_1, S_2$ based on the following two sets of
    vertices: $\{S^0, \cdots, S^{t_1 - 1}, v, S^{t_1 + 1}, \cdots, S^{4d}\}$,
    $\{S^0, \cdots, S^{t_2 - 1}, v, S^{t_2 + 1}, \cdots, S^{4d}\}$, where $v$
    corresponds to the midpoint the longest edge.
```

iteration. Next, we reapply the DC decomposition strategy to new simplices to obtain a more accurate estimate. We stop and return the lower bound estimate in **Step 10** once the bounding error calculated by **Global_error** reaches the prespecified threshold δ. Otherwise, we make more delicate partitions and iterate this step.

4.2 IMPLEMENTATION OF PI-SFP

In this section, the above four functions are illustrated in detail. **1) Initialization()**: The objective of this function is to construct an original simplex S_0 that encloses the feasible region IR_Γ. To achieve this, we draw inspiration from [Horst et al., 2000, Pei and Zhu, 2013] and use the following approach to construct S_0. The justification of such construction is given in lemma. (2).Here S_0 denotes the set of $\boldsymbol{\gamma} = (\gamma_1, \cdots \gamma_{4d})$ satisfying $1 \leq i \leq 4d_u, \mathbf{1} * \boldsymbol{\gamma} \leq \alpha, \gamma_i \geq \gamma_i^l := \min_{\boldsymbol{\gamma} \in IR_\Gamma} \gamma_i$, and $\alpha = 1 + f(y, x) + \frac{d^2(\psi^l + \psi^u)^2}{4f(x)\psi^l\psi^u}$, where

$$\psi^l = \min_{i \in [2d+1, 3d]} \gamma_i^l, \psi^u = \min_{i \in [2d+1, 3d]} \gamma_i^u, \gamma_i^u := \max_{\boldsymbol{\gamma} \in IR_\Gamma} \gamma. \quad (8)$$

2) Bisection(): Motivated by the approach proposed in Rivara [1984], the goal of this function is to partition an input simplex S into two simplices S_1 and S_2 using the longest-edge (LE) bisection strategy. The partitioning details are outlined in Algorithm 2.

3) Bounding(): The purpose of this function is to derive a lower bound of $\underline{f_S(Y_x = y)}$ using the input S. This is the most crucial element of the algorithm. It is worth remembering that $\underline{f_S(Y_x = y)}$ can be expressed as the solution of the optimization program (3) with an additional constraint $\boldsymbol{\gamma} \in S$. With the derivations shown in (4)-(7), we can easily obtain a lower bound of $\underline{f_S(Y_x = y)}$ by solving the following optimization problem[7]:

$$\underline{\underline{f_S(Y_x = y)}} = \min f(y, x) + C_1^{\tan}(\boldsymbol{\gamma}) - C_2^{\sec}(\boldsymbol{\gamma})$$
$$\text{subject to: } \boldsymbol{\phi} \in IR_\Phi, \boldsymbol{\gamma} \in S; D_{i1}^{\tan}(\boldsymbol{\gamma}) - D_{i2}^{\sec}(\boldsymbol{\gamma}) \leq 1,$$
$$D_{i1}^{\sec}(\boldsymbol{\gamma}) - D_{i2}^{\tan}(\boldsymbol{\gamma}) \geq 1, i = 1, ...d. \quad (9)$$

As demonstrated in lemma. (5) of Appendix A.5, the functions $C_1^{\tan}(\boldsymbol{\gamma})$, $C_2^{\sec}(\boldsymbol{\gamma})$, $D_{i1}^{\tan}(\boldsymbol{\gamma})$, $D_{i2}^{\sec}(\boldsymbol{\gamma})$, $D_{i1}^{\sec}(\boldsymbol{\gamma})$ and $D_{i2}^{\tan}(\boldsymbol{\gamma})$ are constructed from $C_1(\boldsymbol{\gamma}), C_2(\boldsymbol{\gamma})$, $D_{i1}(\boldsymbol{\gamma}), D_{i2}(\boldsymbol{\gamma})$'s in (5) based on **sec**ants and **tan**gents within the simplex S:

$$\begin{bmatrix} C_k^{\tan}(\boldsymbol{\gamma}) \\ D_{ik}^{\tan}(\boldsymbol{\gamma}) \end{bmatrix} := \begin{bmatrix} C_k(\boldsymbol{\gamma_0}) \\ D_{ik}(\boldsymbol{\gamma_0}) \end{bmatrix} + \begin{bmatrix} \frac{\partial C_k(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \mid_{\boldsymbol{\gamma} = \boldsymbol{\gamma_0}} \\ \frac{\partial D_{ik}(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \mid_{\boldsymbol{\gamma} = \boldsymbol{\gamma_0}} \end{bmatrix} (\boldsymbol{\gamma} - \boldsymbol{\gamma_0})$$
$$\begin{bmatrix} C_k^{\sec}(\boldsymbol{\gamma}) \\ D_{ik}^{\sec}(\boldsymbol{\gamma}) \end{bmatrix} := \begin{bmatrix} C_k(S^0), ...C_k(S^{4d}) \\ D_{ik}(S^0), ...D_{ik}(S^{4d}) \end{bmatrix} \begin{bmatrix} S^0, ..., S^{4d} \\ 1, ..., 1 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\gamma} \\ 1 \end{bmatrix}. \quad (10)$$

Here $k = 1, 2, \forall \boldsymbol{\gamma_0} \in S$. As shown above, (9) is a linear programming problem that can be solved using various methods, including the simplex algorithm [Klee and Minty, 1972] and the interior algorithm [Kojima et al., 1989, Nesterov and Nemirovskii, 1994].

4) Global_error(): This function is to terminate PI-SFP via estimating the order of the error with respect to n. Recall that in **Step 5** of Algorithm. 1, we always select the $\tilde{S}_k$ with the lowest $\underline{\underline{f_S(Y_x = y)}}$ in the $k-$th iteration. This strategy guarantees (see Appendix A.5 for more details)

$$\underline{\underline{f_{\tilde{S}_k}(Y_x = y)}} \leq \min_{S \in \mathcal{S}_k} \underline{f_S(Y_x = y)} = \underline{f(Y_x = y)}, \quad (11)$$

i.e., all the $\underline{\underline{f_{\tilde{S}_k}(Y_x = y)}}$'s are lower bounds of $\underline{f(Y_x = y)}$, and thus $\underline{f_{opt}^n(Y_x = y)} \leq \underline{f(Y_x = y)}$. From this, we further have that, in the n-th iteration, for any $k \in \{0, \cdots, n\}$,

$$\begin{aligned} 0 &\leq \underline{f(Y_x = y)} - \underline{f_{opt}^n(Y_x = y)} \\ &\leq \min_{S \in \mathcal{S}_k} \underline{f_S(Y_x = y)} - \underline{\underline{f_{\tilde{S}_k}(Y_x = y)}} \\ &\leq \underline{f_{\tilde{S}_k}(Y_x = y)} - \underline{\underline{f_{\tilde{S}_k}(Y_x = y)}}. \end{aligned} \quad (12)$$

Also see Appendix A.5 for details. This allows us to calculate an error bound via targeting

$$\min_{0 \leq k \leq n} \left\{ \underline{f_{\tilde{S}_k}(Y_x = y)} - \underline{\underline{f_{\tilde{S}_k}(Y_x = y)}} \right\}. \quad (13)$$

Since the bound of $\underline{f_{\tilde{S}_k}(Y_x = y)} - \underline{\underline{f_{\tilde{S}_k}(Y_x = y)}}$ is dominated by the diameter of the simplex $\tilde{S}_k$, i.e., $dia(\tilde{S}_k)$, we aim to get an order of (13) based on the order of the smallest $dia(\tilde{S}_k)$ with respect to n. As shown in Eqn (A.48) in Appendix A.5, this order is controlled by the length L_n of the longest nested subsequence of $\{\tilde{S}_k\}_{k=0}^n$, which is summarized as Algorithm 3.

[7] If $IR_\Gamma \cap S = \emptyset$, then $\underline{\underline{f_S(Y_x = y)}} = +\infty$.

```
Algorithm 3: Procedure to estimate the current convergence (Global_error).
Input: Collections of simplex partitions in each
        iteration till $n-$th iteration: $\tilde{S}_0, \cdots, \tilde{S}_n$.
Output: An estimate of the global error.
Let $\{\tilde{S}_{i_k}\}_{k=1}^{L_n}$ be the (longest) subsequence of $\{\tilde{S}_k\}_{k=0}^{n}$
  such that each $\tilde{S}_{i_{j+1}}$ is partitioned from $\tilde{S}_{i_j}$ for
  $j = 0, 1, \cdots, L_n - 1$, where $L_n$ is the length of this
  subsequence;
Return the global error estimate $(\frac{\sqrt{3}}{2})^{\lfloor \frac{L_n}{4d} \rfloor}$.
```

5 THEORETICAL ANALYSIS

This section delves into the theoretical properties of PI-SFP. First, we examine the general convergence rate of PI-SFP concerning L_n (see Theorem. 1). Then, we demonstrate that PI-SFP can be extended from computing $\underline{f(Y_x = y)}$ to the general ACE case.

Assumption 2 *(Positivity) $\mathscr{P}$ is a set of $P(\boldsymbol{W} \mid \boldsymbol{U})$ guaranteeing each compatible solution $P(\boldsymbol{U}, \boldsymbol{X} = x)$ to be positive definite. Namely, $\exists \delta > 0$, such that $\forall \boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\omega}) \in IR_{\boldsymbol{\phi}}$, we have $\boldsymbol{\psi} \geq \delta * \mathbf{1}_{1*d} > \mathbf{0}_{1*d}$.*

It is a fairly broad and reasonable assumption in practice, just in order to ensure that the denominator in the original formulation is not too small to facilitate the calculation. Under this assumption, we have $\psi_i^o < \frac{1}{\delta}$ in (3) and $\psi^l > \delta$ in (8). Hence we have $sup_{\gamma \in IR_\Gamma} \|\gamma\|_{+\infty} < +\infty$ and $dia(S_0) < +\infty$ respectively. In addition, when Assumption 2 is violated, we propose an alternative PI-SFP refer readers to Appendix A.3 for more information.On this basis, we formally collate the previous analysis as follows:

Theorem 1 *Under Assumption. 1–2, PI-SFP concentrates around the target value $\underline{f(Y_x = y)}$ at the $O((\frac{3}{4})^{\lfloor \frac{L_n}{4d} \rfloor})$ rate. Specifically, $| \quad \underline{f_{opt}^n(Y_x = y)} - \underline{f(Y_x = y)} \quad | \leq A(\frac{3}{4})^{\lfloor \frac{L_n}{4d} \rfloor} dia(S_0)^2$, where $A = A_1 + A_2 + A_3 < +\infty$, $A_1 = \max\limits_{\boldsymbol{\gamma} \in S_0} \frac{2(\sqrt{2}+1)\sqrt{d}}{\delta} \| \frac{\partial (C_1(\boldsymbol{\gamma}) - C_2(\boldsymbol{\gamma}))}{\partial \boldsymbol{\gamma}} \|$, $A_2 = \max\limits_{\boldsymbol{\gamma} \in S_0} \| \frac{\partial^2 C_1(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F$, $A_3 = \frac{1}{2} \max\limits_{\boldsymbol{\gamma} \in S_0} \| \frac{\partial^2 C_2(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F$. Here $\| \cdot \|$ denotes the Euclidean norm, and $\| \cdot \|_F$ denotes the Frobenius norm. $L_n \in [\lfloor log(n) \rfloor + 1, n]$ is the length of the longest nested sequence till $n-$th iteration. Moreover, $\lim\limits_{n \to +\infty} \underline{f_{opt}^n(Y_x = y)} = \underline{f(Y_x = y)}$.*

Theorem. 1 states that PI-SFP converges to $\underline{f(Y_x = y)}$ with the growing length of the longest nested sequence, and will approach it in the infinite case. We relegate the proof to Appendix A.5 and reserve a brief summary. First, $\underline{f(Y_x = y)}$ is equal to $\underline{f_{S_0}(Y_x = y)}$ via constructing an original enclosure S_0 in (8). Second, $\underline{f_{S_0}(Y_x = y)}$ is substituted with $\min_{S \in \mathcal{S}_k} \underline{f_S(Y_x = y)}$ in the k-th iteration by bisection. Third, each $\underline{f_S(Y_x = y)}$ is lower bounded by (9), namely we have $\forall S \in \mathcal{S}_k, \underline{f_S(Y_x = y)} \geq \underline{\underline{f_S(Y_x = y)}}$. Finally, $\tilde{S}_k$ with the lowest bound $\min\limits_{S \in \mathcal{S}_k} \underline{\underline{f_S(Y_x = y)}}$ is gathered as $\{\tilde{S}_k\}_{k=0}^n$ in order to formulate $\underline{f_{opt}^n(Y_x = y)}$ (see **Step 10** in Algorithm 1). The asymptotic error can be bounded by (12)-(13). In conclusion, these four steps correspond to the four functions in the above section in order.

Noteworthy, it is well beyond the scope of this paper to theoretically estimate L_n w.r.t n, both empirically and theoretically. We refer readers to Appendix A.7.5 for detailed comment. In this comment, we also figure out a conjecture upon finiteness of regular simplicial partitions, which is our extra contribution. Moreover, our method and theorem could be naturally extended to the ACE case, which is detailed in Appendix A.6 for space limitation.

6 SIMULATIONS AND REAL-WORLD EXPERIMENTS

In this section, we perform experiments to demonstrate the efficacy of PI-SFP, aiming to address two key questions: 1) Can PI-SFP effectively manage the partially observable $P(\boldsymbol{W} \mid \boldsymbol{U})$, a scenario not previously explored in the literature, and generate informative bounds? 2) How does PI-SFP's convergence rate manifest in practical applications? Due to space constraints, some visualizations are deferred to Appendix A.8.

6.1 SIMULATIONS

Experiment settings We refer to the case presented in Section. 3, as shown in Equation. (A.5), and generalize our findings, with a specific focus on Fig. 1(a). We address an intriguing and universal situation referred to as 'information leakage,' where the information of $\boldsymbol{U}$ is regularly retained by $\boldsymbol{W}$ but incurs loss during transmission. Formally, we claim $P(W = w_i \mid U = u_i) \geq 1 - \varepsilon, \varepsilon \in (0, 1)$. To make the experiment simple and representative, we consider the binary cases of $\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X}$. On this basis, the construction is $\overline{P(\boldsymbol{W} \mid \boldsymbol{U})} := (1 - \varepsilon)\boldsymbol{I_{2*2}} + \varepsilon \boldsymbol{J_{2*2}}$ and $\underline{P(\boldsymbol{W} \mid \boldsymbol{U})} := (1 - \varepsilon)\boldsymbol{I_{2*2}}, \varepsilon \in (0, 1)$. The construction of $f(\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{X})$ still follows (A.5)[8]. Moreover, we set the iteration number as 1000.

Experiment result The simulation results, shown in Table. 1 and Fig.3, indicate that PI-SFP successfully finds optimal solutions and values, with a fast convergence rate within 1000 iterations. At the beginning step, estimation errors increase as ε increases, but they remain under control

[8] In order to avoid the ill-conditioned case for PYTHON 3.8.5, we make a rather broad restriction that elements of $P(\boldsymbol{U}, x)$ are at least $1e^{-2}$ in all cases (Assumption. 2)

ε	Φ						PI-SFP result	$\underline{f(Y_x = y)}$ (Ground Truth)	Error
	θ_1	θ_2	ψ_1	ψ_2	ω_1	ω_2			
0.1	0.067	0.133	0.261	0.239	0.333	0.167	0.370	0.372	0.548%
0.2	0.050	0.150	0.262	0.238	0.375	0.125	0.350	0.351	0.285%
0.3	0.029	0.171	0.264	0.236	0.429	0.072	0.298	0.301	0.997%
$\geq$ 0.4	0.001	0.199	0.310	0.190	0.500	0.000	0.200	0.205	2.439%

TABLE 1: *Simulation results (the upper bound is symmetric). The middle column Φ denotes the optimal solution within iteration 10^6. The ground truth is approached via 10^6 Monte-Carlo sampling. Our PI-SFP result decreases monotonically with the increasing ε, since the feasible region of latent variables Φ is gradually enlarged with ε. Detailed visulization of convergence rate is shown in Fig. 3 (Appendix A.8).*

by the theoretical error guaranteed by Theorem.1. During iterations, estimation errors converge quickly to the real $\underline{f(Y_x = y)}$. The ground truth decreases as the feasible region of Φ increases, i.e., when ε increases. Moreover, when $\varepsilon \geq 0.4$, we observe that $\underline{f(Y_x = y)}$ achieves its minimum value Pearl [2009b] of $\underline{f(y, x) = 0.2}00$.

6.2 REAL-WORLD APPLICATIONS

In the simulation experiments, we have demonstrated that PI-SFP can quickly converge to the valid bound; furthermore, in the real experiments in this section, we show that the valid bound generated by PI-SFP can more effectively substantiate the causal relationships in the real world compared with previous methods. Specifically, we re-analyze the Zika Virus outbreak dataset [Taddeo et al., 2022, Tchetgen et al., 2024] in the most-related literature [Tchetgen et al., 2023]. Our PI-SFP result exhibits a more significant adverse effect from Zika Virus to the birth rate, which is more aligned with the well-known scientific hypothesis [Castro et al., 2018] compared with the previous literature [Tchetgen et al., 2023]. We defer the experimental details to Appendix A.8 due to space limitation.

7 JUSTIFICATION OF ASSUMPTIONS AND FURTHER DISCUSSIONS

In this section, for the core partial observability assumption (Assumption.1), we analyze its necessity, generalizability, and verifiability. First, the necessity is supported by the following lemma. It leads to an interesting and counter-intuitive *negative result*: informative proxies (namely transition matrix $P(\boldsymbol{W} \mid \boldsymbol{U})$ is reversible) *do not guarantee* informative bounds (instead of the vanilla bound);

lemma 1 *Assume that $[\underline{P(\boldsymbol{W} \mid \boldsymbol{U})}, \overline{P(\boldsymbol{W} \mid \boldsymbol{U})}] = [\mathbf{0}_{d_w * d_u}, \mathbf{1}_{d_w * d_u}]$, and $f(\boldsymbol{U}, x) > \mathbf{0}$. We consider the whole set of $f(y, \boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X})$ which is within $\widetilde{\mathcal{F}}$ and is additionally compatible with two observed distributions $f(\boldsymbol{W}, \neg x) > \mathbf{0}_{d_w * 1}$, $f(y, \boldsymbol{W}, x) > \mathbf{0}_{d_w * 1}$ by an unknown $P(\boldsymbol{W} \mid \boldsymbol{U})$. Then (i) The tight lower bound of $f(Y_x = y)$ is vanilla, namely $f(y, x)$. (ii) If $P(\boldsymbol{W} \mid \boldsymbol{U})$ is restricted to be left-reversible and $f(\boldsymbol{W} \mid \neg x) \neq f(\boldsymbol{W} \mid x, y)$, then the tight lower bound of $f(Y_x = y)$ is still the vanilla $f(y, x)$. (iii) If $P(\boldsymbol{W} \mid \boldsymbol{U})$ is restricted to be left-reversible and $f(\boldsymbol{W} \mid \neg x) = f(\boldsymbol{W} \mid x, y)$, then $f(Y_x = y)$ is lower bounded by another vanilla bound $f(y \mid x)$.*

The proof is in Appendix A.7.1. This lemma extends the well-known inequality $f(Y_x = y) \geq f(y, x)$ [Pearl, 2009b] to single proxy control. Lemma 1 sufficiently indicates that partial observability Assumption 1 ([Kuroki and Pearl, 2014, Greenland, 2005]), instead of the reversibility assumption in the previous literature (e.g., [Miao and Tchetgen, 2018]), is more necessary for partial identification in most cases.

Furthermore, for genealizability, verifiability and practical correspondence of Assumption 1, we refer readers to arguments in Appendix A.7.1. Moreover, we provide algorithm comparison and algorithm acceleration in Appendix A.7.2, discuss graph structure extension in Appendix A.7.3, and then extend to the continuous case of confounders in Appendix A.7.4.

8 CONCLUSIONS

In this paper, we highlight the limitations and strict assumptions of the transfer matrix $P(\boldsymbol{W} \mid \boldsymbol{U})$ through practical examples, emphasizing that exact observability and reversibility are often not feasible in real-world scenarios. Based on this, we propose a novel PI-SFP framework that achieves a valid bound for the causal effect, even with only partial observability of $P(\boldsymbol{W} \mid \boldsymbol{U})$. To achieve it, we employ deformation techniques in DC programming and implement a branch-and-bound method. We offer a theoretical analysis of the mathematical reasons behind the lack of tight bounds and provide sufficient and necessary conditions to determine if the bounds are tight. We also conduct a convergence rate analysis of PI-SFP. Furthermore, we provide specific convergence rate analysis for these methods. We also provide a fundamental negative result that informative proxies might not yield informative partial identification bounds.

Our paper has initiated new research trajectories, specifically focusing on the proximal partial identification with broader confounding proxy information. An additional avenue for exploration could involve evaluating the performance of PI-SFP under more intricate partial observability assumptions. Furthermore, it would be promising for in-depth investigation to extend our single-proxy control scheme to encompass double-proxy control and other causal graphs. These avenues remain further exploration in our future research.

ACKNOWLEDGEMENT

The research was partially completed while Zhiheng Zhang was a student intern at Shanghai Qi Zhi Institute. He is grateful to Professor Yuhao Wang for suggestions.

References

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2009.

Alexander Balke and Judea Pearl. Counterfactual probabilities: Computational methods, bounds and applications. In *Uncertainty Proceedings 1994*, pages 46–54. Elsevier, 1994.

Ivan Budimir, Sever S Dragomir, and Josep Pecaric. Further reverse results for jensen's discrete inequality and applications in information theory. *J. Inequal. Pure Appl. Math*, 2(1):5, 2001.

Zhihong Cai and Manabu Kuroki. On identifying total effects in the presence of latent variables and selection bias. *arXiv preprint arXiv:1206.3239*, 2012.

David Card and Alan B Krueger. Minimum wages and employment: A case study of the fast food industry in new jersey and pennsylvania, 1993.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. CRC press, 2006.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11 (1):3673, 2020.

Marcia C Castro, Qiuyi C Han, Lucas R Carvalho, Cesar G Victora, and Giovanny VA França. Implications of zika virus and congenital zika syndrome for the number of live births in brazil. *Proceedings of the National Academy of Sciences*, 115(24):6177–6182, 2018.

Philippe G Ciarlet. *The finite element method for elliptic problems*. SIAM, 2002.

Yifan Cui, Hongming Pu, Xu Shi, Wang Miao, and Eric Tchetgen Tchetgen. Semiparametric proximal causal inference. *arXiv preprint arXiv:2011.08411*, 2020.

Yang Dai, Jianming Shi, and Shouyang Wang. Conical partition algorithm for maximizing the sum of dc ratios. *Journal of Global Optimization*, 31(2):253–270, 2005.

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

Yi-Jian Du, Bo Feng, and Chih-Hao Fu. Note on cyclic sum and combination sum of color-ordered gluon amplitudes. *Journal of High Energy Physics*, 2012(3):1–23, 2012.

Guilherme Duarte, Noam Finkelstein, Dean Knox, Jonathan Mummolo, and Ilya Shpitser. An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, (just-accepted):1–25, 2023.

Mirjam. Dur, Reiner. Horst, and Nguyen. Van Thoai. Solving sum-of-ratios fractional programs using efficient points. *Optimization*, 49(5-6):447–466, 2001. doi: 10.1080/02331930108844543. URL `https://doi.org/10.1080/02331930108844543`.

Peter A Frost. Proxy variables and specification bias. *The review of economics and Statistics*, pages 323–325, 1979.

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

AmirEmad Ghassami, Ilya Shpitser, and Eric Tchetgen Tchetgen. Partial identification of causal effects using proxy variables. *arXiv preprint arXiv:2304.04374*, 2023.

Sander Greenland. Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168(2):267–306, 2005.

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

Reiner Horst and Nguyen V Thoai. Dc programming: overview. *Journal of Optimization Theory and Applications*, 103(1):1–43, 1999.

Reiner Horst, Panos M Pardalos, and Nguyen Van Thoai. *Introduction to global optimization*. Springer Science & Business Media, 2000.

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

Baker Kearfott. A proof of convergence and an error bound for the method of bisection in $\boldsymbol{R}^n$. *Mathematics of Computation*, 32(144):1147–1153, 1978.

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

Victor Klee and George J Minty. How good is the simplex algorithm. *Inequalities*, 3(3):159–175, 1972.

Masakazu Kojima, Shinji Mizuno, and Akiko Yoshise. A primal-dual interior point algorithm for linear programming. In *Progress in mathematical programming*, pages 29–47. Springer, 1989.

Stanislav Kolenikov and Gustavo Angeles. Socioeconomic status measurement with discrete proxy variables: Is principal component analysis a reliable answer? *Review of Income and Wealth*, 55(1):128–165, 2009.

Sergey Korotov, Ángel Plaza, and José P Suárez. Longest-edge n-section algorithms: properties and open problems. *Journal of Computational and Applied Mathematics*, 293: 139–146, 2016.

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

Eugene L Lawler and David E Wood. Branch-and-bound methods: A survey. *Operations research*, 14(4):699–719, 1966.

Hoai An Le Thi, Tao Pham Dinh, et al. Dc programming and dca for general dc programs. In *Advanced Computational Methods for Knowledge Engineering*, pages 15–35. Springer, 2014.

Sanghack Lee and Elias Bareinboim. Causal identification with matrix equations. *Columbia CausalAI Laboratory Technical Report*, 2020.

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

George Matsaglia and George PH Styan. Equalities and inequalities for ranks of matrices. *Linear and multilinear Algebra*, 2(3):269–292, 1974.

Wang Miao and Eric Tchetgen Tchetgen. A confounding bridge approach for double negative control inference on causal effects (supplement and sample codes are included). *arXiv preprint arXiv:1808.04945*, 2018.

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

Kenichi Nagasawa. Identification and estimation of partial effects with proxy variables. *arXiv preprint arXiv:1811.00667*, 2018.

Yurii Nesterov and Arkadii Nemirovskii. *Interior-point polynomial algorithms in convex programming*. SIAM, 1994.

Chan Park and Eric Tchetgen Tchetgen. Single proxy synthetic control. *arXiv preprint arXiv:2307.16353*, 2023.

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

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009b.

Judea Pearl. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*, 2012.

Judea Pearl. On the testability of causal models with latent and instrumental variables. *arXiv preprint arXiv:1302.4976*, 2013.

Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000.

Yonggang Pei and Detong Zhu. Global optimization method for maximizing the sum of difference of convex functions ratios over nonconvex region. *Journal of Applied Mathematics and Computing*, 41(1):153–169, 2013.

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

Maria-Cecilia Rivara. Mesh refinement processes based on the generalized bisection of simplices. *SIAM Journal on Numerical Analysis*, 21(3):604–613, 1984.

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

Masahiro Sato, Sho Takemori, Janmajay Singh, and Tomoko Ohkuma. Unbiased learning for the causal effect of recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 378–387, 2020.

Siegfried Schaible and Jianming Shi. Fractional programming: the sum-of-ratios case. *Optimization Methods and Software*, 18(2):219–229, 2003.

Jan Selén. Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data. *Journal of the American Statistical Association*, 81 (393):75–81, 1986.

Peiping Shen, Tongli Zhang, and Chunfeng Wang. Solving a class of generalized fractional programming problems using the feasibility of linear programs. *Journal of inequalities and applications*, 2017(1):1–16, 2017.

Xu Shi, Wang Miao, Jennifer C Nelson, and Eric J Tchetgen Tchetgen. 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.

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

Torsten Söderström and Petre Stoica. Instrumental variable methods for system identification. *Circuits, Systems and Signal Processing*, 21(1):1–9, 2002.

Ioan M Stancu-Minasian. *Fractional programming: theory, methods and applications*, volume 409. Springer Science & Business Media, 2012.

Otto Neall Strand and Ed R Westwater. Statistical estimation of the numerical solution of a fredholm integral equation of the first kind. *Journal of the ACM (JACM)*, 15(1): 100–114, 1968.

Marcelo M Taddeo, Leila D Amorim, and Rosana Aquino. Causal measures using generalized difference-in-difference approach with nonlinear models. *Statistics and Its Interface*, 15(4):399–413, 2022.

Pham Dinh Tao and Le Thi Hoai An. Convex analysis approach to dc programming: theory, algorithms and applications. *Acta mathematica vietnamica*, 22(1):289–355, 1997.

Eric J Tchetgen Tchetgen, Andrew Ying, Yifan Cui, Xu Shi, and Wang Miao. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Eric J Tchetgen Tchetgen, Chan Park, and David B Richardson. Universal difference-in-differences for causal inference in epidemiology. *Epidemiology*, 35(1):16–22, 2024.

Eric Tchetgen Tchetgen, Chan Park, and David Richardson. Single proxy control. *arXiv preprint arXiv:2302.06054*, 2023.

Michael R Wickens. A note on the use of proxy variables. *Econometrica: Journal of the Econometric Society*, pages 759–761, 1972.

Jeffrey M Wooldridge. On estimating firm-level production functions using proxy variables to control for unobservables. *Economics letters*, 104(3):112–114, 2009.

Qing Zhang, Xiaoying Zhang, Yang Liu, Hongning Wang, Min Gao, Jiheng Zhang, and Ruocheng Guo. Debiasing recommendation by learning identifiable latent confounders. *arXiv preprint arXiv:2302.05052*, 2023a.

Zhiheng Zhang, Wenbo Hu, Tian Tian, and Jun Zhu. Dynamic window-level granger causality of multi-channel time series. *arXiv preprint arXiv:2006.07788*, 2020.

Zhiheng Zhang, Quanyu Dai, Xu Chen, Zhenhua Dong, and Ruiming Tang. Robust causal inference for recommender system to overcome noisy confounders. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2349–2353, 2023b.

Literature	Tools			Assumptions		
	Valid Instrument	Negative exposure	Negative outcome	Reversibility Completeness	Bridge function	Observability of $P(\boldsymbol{W} \mid \boldsymbol{U})$
[Balke and Pearl, 1994] [Kitagawa, 2009]	✓	✗	✗	✗	✗	✗
[Kuroki and Pearl, 2014](1) [Rothman et al., 2008] [Lee and Bareinboim, 2020]	✗	✗	✓	✓	✗	✓[a]
[Kuroki and Pearl, 2014](2) [Nagasawa, 2018]	✗	✓	✓	✓	✗	✗
[Miao et al., 2018] [Shi et al., 2020] [Singh, 2020] [Cui et al., 2020] [Tchetgen et al., 2020] [Deaner, 2018]	✗	✓	✓	✓	✓	✗
[Kallus et al., 2021]	✗	✓	✓	✗	✓	✗
Our paper	✖	✖	✔	✖	✖	✔[b]

TABLE 2: *Tools and assumptions of previous literature on partial identification. [Kuroki and Pearl, 2014](1) is without $\boldsymbol{Z}$, while (2) is with $\boldsymbol{Z}$.*

[a] $P(\boldsymbol{W} \mid \boldsymbol{U})$ is assumed to be reversible and explicitly, totally observed.

[b] In our paper, $P(\boldsymbol{W} \mid \boldsymbol{U})$ only needs be partially bounded.

In the appendices, we provide supplementary material and proofs for our main text. Appendices A.1-A.2 contain proofs for propositions. In Appendix A.1, we establish that $\mathcal{F} \subseteq \widetilde{\mathcal{F}}$. In Appendix A.2, we demonstrate that the bound is tight under certain conditions.

Appendix A.3 is dedicated to discussing the assumption. We explore cases where Assumption 2 does not hold.

Appendix A.5 contains the main results. Firstly, we show that the original simplex S_0 encloses our identification region. Secondly, we prove that the original optimization problem can be transformed into a set of sub-problems in the reduced space. Thirdly, we demonstrate our construction to transfer the original nonlinear optimization problem to the weaker linear case. Finally, we prove that our algorithm converges to the global optimal solution at an exponential rate.

Appendix A.6 extends our result from $f(Y_x = y)$ to the more general ACE.

Appendix A.7 is dedicated to extensions. We discuss 1) the previous assumptions in the original literature, 2) auxiliary acceleration strategies, 3) extension of graph structure, and 4) extension to the continuous confounding.

A.1 THE PROOF OF PROPOSITION 1

According to Assumption. 1, by integration, we can also directly claim that if $f(y, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{X}) \in \mathcal{F}$, then

$$\begin{aligned}
\underline{P(\boldsymbol{W} \mid \boldsymbol{U})}\boldsymbol{\theta} &\leq P(y, \boldsymbol{W}, x) \leq \overline{P(\boldsymbol{W} \mid \boldsymbol{U})}\boldsymbol{\theta}, \forall x \in X. \\
\underline{P(\boldsymbol{W} \mid \boldsymbol{U})}\boldsymbol{\psi} &\leq P(\boldsymbol{W}, x) \leq \overline{P(\boldsymbol{W} \mid \boldsymbol{U})}\boldsymbol{\psi}, \forall x \in X. \\
\underline{P(\boldsymbol{W} \mid \boldsymbol{U})}\boldsymbol{\omega} &\leq f(\boldsymbol{W}, \neg x) \leq \overline{P(\boldsymbol{W} \mid \boldsymbol{U})}\boldsymbol{\omega}, \forall x \in X.
\end{aligned} \tag{A.1}$$

Thus

$$\begin{bmatrix} -\boldsymbol{I_{d*d}} \\ \boldsymbol{I_{d*d}} \end{bmatrix} \begin{bmatrix} f(y, \boldsymbol{W}, x)^T \\ f(\boldsymbol{W}, x)^T \\ f(\boldsymbol{W}, \neg x)^T \end{bmatrix}^T - \begin{bmatrix} -\overline{P(\boldsymbol{W} \mid \boldsymbol{U})} \\ \underline{P(\boldsymbol{W} \mid \boldsymbol{U})} \end{bmatrix} \boldsymbol{\phi} \geq \boldsymbol{0}. \tag{A.2}$$

Combined with the natural that $\theta_i \in [0, P(y, x)], \psi_i \in [0, P(x)], \omega_i \in [0, P(\neg x)], i = 1, 2, ...d$, we have $f(y, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{X}) \in \widetilde{\mathcal{F}}$. In conclusion, we claim $\mathcal{F} \subseteq \widetilde{\mathcal{F}}$.

A.2 THE PROOF OF PROPOSITION 2

As the optimal solution $\boldsymbol{\phi_{opt}}$ satisfies the constraint (2) in Proposition. (2), we can equivalently claim that $\boldsymbol{\phi_{opt}}$ is compatible with some $f(y, \boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X})$ which satisfies $f(y, \boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X}) \in \mathcal{F}$. On this basis, the original formulation can be transformed with stricter constraints but equal minimum optimal value, namely that from

$$\min f(y,x) + \sum_{i=1}^{d} \frac{1}{\psi_i}\theta_i\omega_i, \text{ subject to: } f(y, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{X}) \in \widetilde{\mathcal{F}} \tag{A.3}$$

to

$$\min f(y,x) + \sum_{i=1}^{d} \frac{1}{\psi_i}\theta_i\omega_i, \text{ subject to: } f(y, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{X}) \in \widetilde{\mathcal{F}} \cap \mathcal{F} = \mathcal{F}. \tag{A.4}$$

This is equal to the original (2). Hence $f(Y_x = y)$ is the tight lower bound of $f(Y_x = y)$ under constraint. 2.

By contrast, if the constraint (2) does not hold, then any $f(y, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{X})$ compatible with $\boldsymbol{\phi_{opt}}$ will be within $\mathcal{F}^c \cap \widetilde{\mathcal{F}}$. In another word, the minimum value of the original formulation will be lower than that of Eqn (2), and the bound $\underline{f(Y_x = y)}$ is not tight. Proved.

We further provide an instance that our bound is tight:

$$\begin{bmatrix} f(Y = y, \boldsymbol{W}, x)^T \\ f(Y = y, \boldsymbol{W}, \neg x)^T \\ f(Y \neq y, \boldsymbol{W}, x)^T \\ f(Y \neq y, \boldsymbol{W}, \neg x)^T \end{bmatrix} = \begin{bmatrix} 0.08 & 0.12 \\ 0.15 & 0.1 \\ 0.18 & 0.12 \\ 0.15 & 0.1 \end{bmatrix}, \begin{bmatrix} \overline{P(\boldsymbol{W} \mid \boldsymbol{U})} \\ \underline{P(\boldsymbol{W} \mid \boldsymbol{U})} \end{bmatrix} = \begin{bmatrix} 0.6\boldsymbol{I_{2*2}} + 0.4\boldsymbol{J_{2*2}} \\ 0.6\boldsymbol{I_{2*2}} \end{bmatrix}. \tag{A.5}$$

Here $\boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X}$ are all binary, and $\boldsymbol{I_{n*n}}, \boldsymbol{J_{n*n}}$ denote the n dimensional identity matrix and all-ones matrix respectively. we can verify one of the optimal solutions $\boldsymbol{\phi_{opt}} = [0\ 0.2\ 0.3\ 0.2\ 0.5\ 0]^T$. The corresponding $f(y, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{X})$ satisfying Eqn (2) exists, whose explicit form is detailed as follows:

$$\begin{bmatrix} f(Y = y, W = w_1, \boldsymbol{U}, \boldsymbol{X}) & f(Y \neq y, W = w_1, \boldsymbol{U}, \boldsymbol{X}) \\ f(Y = y, W = w_2, \boldsymbol{U}, \boldsymbol{X}) & f(Y \neq y, W = w_2, \boldsymbol{U}, \boldsymbol{X}) \end{bmatrix} = \begin{bmatrix} 0 & 0.15 & 0.18 & 0.15 \\ 0.08 & 0 & 0 & 0 \\ 0 & 0.1 & 0.12 & 0.1 \\ 0.12 & 0 & 0 & 0 \end{bmatrix} \tag{A.6}$$

A.3 FURTHER DISCUSSION ON ASSUMPTION. 2

In this section, we consider the case when Assumption. 2 does not hold. We propose a new version of PI-SFP. Recall that our objective function is:

$$\underline{f(Y_x = y)} = \min f(y,x) + \sum_{i=1}^{d} \frac{1}{\psi_i}\theta_i\omega_i, \text{ subject to: } f(y, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{X}) \in \widetilde{\mathcal{F}}, i.e., \boldsymbol{\phi} \in IR_{\boldsymbol{\Phi}}. \tag{A.7}$$

In our main text, we let $\psi_i^o = \frac{1}{\psi_i}$. However, when we can not guarantee that $\exists\delta$, $\forall i \in \{1, 2, ...d\}, \psi_i \geq \delta$ (without Assumption. 2,), then ψ_i^o may turn to infinity. On this basis, we introduce another algebraic distortion $\psi_i^o = \frac{\theta_i\omega_i}{\psi_i}$. Then the above programming can be transformed to:

$$\underline{f(Y_x = y)} = \min f(y,x) + \sum_{i=1}^{d} \psi_i^o$$
$$\text{subject to: } \boldsymbol{\gamma} \in IR_\Gamma, \text{where } IR_\Gamma = \{\boldsymbol{\gamma} : \boldsymbol{\phi} \in IR_\Phi, \psi_i^o\psi_i = \theta_i\omega_i, \psi_i^o \leq C, i = 1, ...d.\}. \tag{A.8}$$

where C is a local optimal value (a priori computed) of $\underline{f(Y_x = y)}$. On this basis, we can adopt the analogous strategy as in the traditional PI-SFP. Here the original S_0 is easy to be constructed since $\|\boldsymbol{\gamma}\|_{+\infty} < +\infty$.

Programming (A.8) can also be adopted under Assumption. 2. Compared with the traditional PI-SFP, firstly, programming (A.8) needs an a priori computed C to upper bound ψ_i^o. Secondly, we will do linearization on $\psi_i^o\psi_i = \theta_i\omega_i$ instead of $\psi_i^o\psi_i = 1$, which is more complex. There is no guarantee of which version is better and we will explore it in the future work.

A.4 VALID BOUND OF AVERAGE CAUSAL EFFECT (ACE)

The identification region of $f(Y_x)$ is constructed as follows.

$$\begin{aligned} &IR_{F(Y_x)} = \{f(Y_x) : \int_{Y^L}^{Y^U} f(Y_x = y)dy = 1, \\ &\forall y \in [Y^L, Y^U], f(y, \boldsymbol{U}, \boldsymbol{W}, \boldsymbol{X}) \in \mathcal{F}\}. \end{aligned} \tag{A.9}$$

Then the valid bound of $ACE_{\boldsymbol{X}\to\boldsymbol{Y}}$ can be denoted as $[\underline{ACE_{\boldsymbol{X}\to\boldsymbol{Y}}}, \overline{ACE_{\boldsymbol{X}\to\boldsymbol{Y}}}]$:

$$\begin{aligned} &\underline{ACE_{\boldsymbol{X}\to\boldsymbol{Y}}} \leq \min\{ACE_{\boldsymbol{X}\to\boldsymbol{Y}} \\ &= \int_{X^L}^{X^U} \int_{Y^L}^{Y^U} f(Y_x = y)\pi(x)dxdy : f(Y_x) \in IR_{F(Y_x)}\}, \\ &\overline{ACE_{\boldsymbol{X}\to\boldsymbol{Y}}} \geq \max\{ACE_{\boldsymbol{X}\to\boldsymbol{Y}} \\ &= \int_{X^L}^{X^U} \int_{Y^L}^{Y^U} f(Y_x = y)\pi(x)dxdy : f(Y_x) \in IR_{F(Y_x)}\}. \end{aligned} \tag{A.10}$$

$[\underline{ACE_{\boldsymbol{X}\to\boldsymbol{Y}}}, \overline{ACE_{\boldsymbol{X}\to\boldsymbol{Y}}}]$ is the valid bound of ACE. In our paper, we aim to design an algorithm to seek the valid bound of $f(Y_x = y)$, and then extend our strategy from bounding $f(Y_x = y)$ to bounding ACE. Homoplastically, we only need to consider the optimization technique on the minimum case, and the maximum case will be symmetric.

A.5 THE PROOF OF THEOREM. 1

The sketch of proof This is the main result of our paper. The main procedure are as follows:

$$\begin{aligned} &\mid \underline{f(Y_x = y)} - \underline{f_{opt}^n(Y_x = y)} \mid \\ &= \mid \underline{f(Y_x = y)} - \max_{k\in\{0,1,...n\}} \underline{\underline{f_{\tilde{S}_k}(Y_x = y)}} \mid && \text{Definition of } \underline{f_{opt}^n(Y_x = y)} \\ &= \min_{k\in\{0,1,...n\}} \mid \underline{f(Y_x = y)} - \underline{\underline{f_{\tilde{S}_k}(Y_x = y)}} \mid \\ &\overset{\textbf{(1)}}{=} \min_{k\in\{0,1,...n\}} \mid \underline{f_{S_0}(Y_x = y)} - \underline{\underline{f_{\tilde{S}_k}(Y_x = y)}} \mid && \textbf{Initialization} \\ &\overset{\textbf{(2)}}{=} \min_{k\in\{0,1,...n\}} \mid \min_{S\in\mathcal{S}_k} \underline{f_S(Y_x = y)} - \underline{\underline{f_{\tilde{S}_k}(Y_x = y)}} \mid && \textbf{Bisection} \\ &\overset{*}{\leq} \min_{k\in\{0,1,...n\}} \mid \underline{f_{\tilde{S}_k}(Y_x = y)} - \underline{\underline{f_{\tilde{S}_k}(Y_x = y)}} \mid \\ &\leq \mid \underline{f_{\tilde{S}_{i_{L_n}}}(Y_x = y)} - \underline{\underline{f_{\tilde{S}_{i_{L_n}}}(Y_x = y)}} \mid \\ &\overset{\textbf{(3)}}{=} O(dia(\tilde{S}_{i_{L_n}})) && \textbf{Bounding} \\ &\overset{\textbf{(4)}}{=} O((\frac{\sqrt{3}}{2})^{\lfloor \frac{L_n}{4d} \rfloor}). && \textbf{Global\_error} \end{aligned} \tag{A.11}$$

$*$ is directly by **(2)** and we have previously mentioned it in Eqn (12). In the following demonstration, we mainly focus on procedure **(1)(2)(3)(4)**, corresponding to the algorithm part **Initialization**, **Bisection**, **Bounding**, **Global_error** in order.

(1) Initialization() We will claim that $IR_\Gamma \subseteq S_0$.

lemma 2 *The original S_0 satisfies $IR_\Gamma \subseteq S_0$, and thus $\underline{f(Y_x = y)} = \underline{f_{S_0}(Y_x = y)}$.*

The proof of lemma. (2) The simplex construction is as follows. S_0 is spanned by $\{S_0^0, S_0^1, ...S_0^{4d}\}$, where

$$S_0^i = \begin{cases} \boldsymbol{\gamma^l}, \ i = 0 \\ \boldsymbol{\gamma^l} + (\alpha - \boldsymbol{1_{1*4d}}\boldsymbol{\gamma^l}) * \boldsymbol{\vec{e_i}}, \ i \in \{1, 2, ...4d\} \end{cases}, \text{ where } \boldsymbol{\gamma^l} = (\gamma_1^l, \gamma_2^l, ...\gamma_{4d}^l)^T, \tag{A.12}$$

where S_0^i is the supporting vertices set described in our main text. For each $\boldsymbol{\gamma} \in IR_\Gamma$, we attempt to provide a direct construction as follows:

$$\forall \boldsymbol{\gamma} \in IR_\Gamma, \text{we have } \boldsymbol{\gamma} \overset{*}{=} \sum_{i=0}^{4d} \beta_i S_0^i,\ \beta_i = \begin{cases} 1 - \sum_{i=1}^{4d} \beta_i, i = 0 \\ \frac{\boldsymbol{\gamma}\vec{\boldsymbol{e_i}} - \gamma_i^l}{\alpha - \mathbf{1_{1*4d}}\boldsymbol{\gamma^l}}, i = 1, 2, ...4d \end{cases}, \beta_i \in [0, 1], \tag{A.13}$$

To prove (A.13), we only need to prove the correctness of the equality $*$ and the fact $\beta_i \in [0, 1], \forall i = 0, 1, ...4d$.

First, we demonstrate the correctness of this construction.

$$\begin{aligned} \sum_{i=0}^{4d} \beta_i S_0^i &= \beta_0 S_0 + \sum_{i=1}^{4d} \beta_i S_0^i \\ &= \beta_0 \boldsymbol{\gamma^l} + \sum_{i=1}^{4d} \beta_i \left(\boldsymbol{\gamma^l} + (\alpha - \mathbf{1_{1*4d}}\boldsymbol{\gamma^l})\vec{\boldsymbol{e_i}} \right) && \text{(definition of } S_0^i\text{)} \\ &= (1 - \sum_{i=1}^{4d} \beta_i)\boldsymbol{\gamma^l} + \sum_{i=1}^{4d} \beta_i \left(\boldsymbol{\gamma^l} + (\alpha - \mathbf{1_{1*4d}}\boldsymbol{\gamma^l})\vec{\boldsymbol{e_i}} \right) && \text{(definition of } \beta_i\text{)} \\ &= \boldsymbol{\gamma^l} + \sum_{i=1}^{4d} \frac{\boldsymbol{\gamma}\vec{\boldsymbol{e_i}} - \gamma_i^l}{\alpha - \mathbf{1_{1*4d}}\boldsymbol{\gamma^l}} \left((\alpha - \mathbf{1_{1*4d}}\boldsymbol{\gamma^l})\vec{\boldsymbol{e_i}} \right) && \text{(definition of } \beta_i\text{)} \\ &= \boldsymbol{\gamma^l} + \left(\sum_{i=1}^{4d} \boldsymbol{\gamma}\vec{\boldsymbol{e_i}} - \gamma_i^l \right) \vec{\boldsymbol{e_i}} = \boldsymbol{\gamma}. \end{aligned} \tag{A.14}$$

Second, we claim $\forall i \in \{1, ...4d\}, \beta_i \in [0, 1]$. Since we already have $\beta_i > 0, i = 1, 2, ...4d$ according to the construction of $\{\alpha, \boldsymbol{\gamma^l}\}$, we only need to prove the left: $\beta_0 > 0$. Notice that

$$\sum_{i=1}^{4d} \beta_i = \sum_{i=1}^{4d} \frac{\boldsymbol{\gamma}\vec{\boldsymbol{e_i}} - \gamma_i^l}{\alpha - \mathbf{1_{1*4d}}\boldsymbol{\gamma^l}} = \frac{\mathbf{1_{1*4d}}\boldsymbol{\gamma} - \mathbf{1_{1*4d}}\boldsymbol{\gamma^l}}{\alpha - \mathbf{1_{1*4d}}\boldsymbol{\gamma^l}}. \tag{A.15}$$

Due to $\beta_0 = 1 - \sum_{i=1}^{4d} \beta_i$, it is equal to prove

$$\mathbf{1_{1*4d}}\boldsymbol{\gamma} \leq \alpha = 1 + f(y, x) + \frac{d^2(\psi^l + \psi^u)^2}{4f(x)\psi^l\psi^u}, \tag{A.16}$$

where ψ^l, ψ^u are identified in the main text. It is equivalent to

$$\sum_{i=1}^{d} \psi_i^o + \sum_{i=1}^{d} \theta_i + \sum_{i=1}^{d} \psi_i + \sum_{i=1}^{d} \omega_i \leq 1 + f(y, x) + \frac{d^2(\psi^l + \psi^u)^2}{4f(x)\psi^l\psi^u}, \tag{A.17}$$

namely that

$$\sum_{i=1}^{d} \psi_i^o \leq \frac{d^2(\psi^l + \psi^u)^2}{4f(x)\psi^l\psi^u}. \tag{A.18}$$

We only need prove the inequality (A.18). It is due to the fact $(\psi_i - \psi^l)(\frac{1}{\psi_i} - \frac{1}{\psi^u}) \geq 0$, namely $1 + \frac{\psi^l}{\psi^u} \geq \frac{\psi^l}{\psi_i} + \frac{\psi_i}{\psi^u}$. By which we have

$$(1 + \frac{\psi^l}{\psi^u})d \geq \psi^l \sum_{i=1}^{d} \frac{1}{\psi_i} + \frac{1}{\psi^u} \sum_{i=1}^{d} \psi_i \geq 2\sqrt{\frac{\psi^l}{\psi^u}} \sqrt{\sum_{i=1}^{d} \frac{1}{\psi_i}} \sqrt{f(x)}. \tag{A.19}$$

It is equal to

$$\sum_{i=1}^{d} \psi_i^o = \sum_{i=1}^{d} \frac{1}{\psi_i} \leq \frac{(\psi^u + \psi^l)^2 d^2}{4f(x)\psi^u\psi^l}, \text{and thus} \sum_{i=1}^{4d} \beta_i \in [0, 1]. \tag{A.20}$$

On this basis, $\beta_0 = 1 - \sum_{i=1}^{4d} \beta_i \in [0, 1]$. Combining with $\beta_i \geq 0, i \in \{0, 1, ...4d\}$ and Eqn (A.13), we claim that $\forall \boldsymbol{\gamma} \in IR_\Gamma$, we have $\boldsymbol{\gamma} \in S_0$. Due to the arbitrary of $\boldsymbol{\gamma}$, we have $IR_\Gamma \subseteq S_0$, and thus $\underline{f(Y_x = y)} = \underline{f_{S_0}(Y_x = y)}$.

(2) Bisection() We introduce the following lemma:

lemma 3 *The partitioning set $\mathcal{S}_k$ satisfies $\underline{f_{S_0}(Y_x = y)} = \min_{S \in \mathcal{S}_k} \underline{f_S(Y_x = y)}$.*

The proof of lemma. (3) By definition of bisection process, $\tilde{S}_k$ is bisectioned into $\tilde{S}_{k1}, \tilde{S}_{k2}$. Then

$$\mathcal{S}_{k+1} := \left(\mathcal{S}_k \setminus \tilde{S}_k\right) \cup \{\tilde{S}_{k1}, \tilde{S}_{k2}\} \tag{A.21}$$

Hence we have $\cup_{S \in \mathcal{S}_k} S = \cup_{S \in \mathcal{S}_{k+1}} S, \forall k = 0, 1, ...$ Thus $S_0 = \cup_{S \in \mathcal{S}_k} S$, and we have

$$\underline{f_{S_0}(Y_x = y)} = \underline{f_{(\cup_{S \in \mathcal{S}_k} S)}(Y_x = y)} = \min_{S \in \mathcal{S}_k} \underline{f_S(Y_x = y)}. \tag{A.22}$$

Hence we have proved.

(3) Bounding() We first introduce lemma. (4) and lemma. (5) for preparation, then the procedure **(3)** is proved by lemma. (6).

lemma 4 *The decomposition of* (3) *can be established as Eqn* (5).

The proof of lemma. (4) Specifically, we give the explicit decomposition as follows, and the sub-script *cyc* means the cycle of symbol set $[\psi_i^o, \theta_i, \omega_i]$:

$$\begin{aligned}
&\psi_i^o \theta_i \omega_i \\
&= \left[\frac{1}{2}\sum_{cyc}(\psi_i^o)^2\theta_i + \frac{1}{2}\sum_{cyc}(\psi_i^o)\theta_i^2 + \sum_{i=1}^{d}\psi_i^o\theta_i\omega_i\right] - \frac{1}{2}\sum_{cyc}(\psi_i^o)^2\theta_i - \frac{1}{2}\sum_{cyc}(\psi_i^o)\theta_i^2 \\
&= \left[\frac{1}{6}(\sum_{cyc}\psi_i^o)^3 - \frac{1}{6}(\sum_{cyc}(\psi_i^o)^3)\right] - \frac{1}{2}\sum_{cyc}(\psi_i^o)^2\theta_i - \frac{1}{2}\sum_{cyc}(\psi_i^o)\theta_i^2 \\
&= \left[\frac{1}{6}(\sum_{cyc}\psi_i^o)^3 - \frac{1}{6}(\sum_{cyc}(\psi_i^o)^3)\right] + \frac{1}{2}\sum_{cyc}(\psi_i^o)^4 - \frac{1}{4}\sum_{cyc}(\psi_i^o)^4 - \frac{1}{4}\sum_{cyc}(\theta_i)^2 - \frac{1}{2}\sum_{cyc}(\psi_i^o)^2\theta_i \\
&\quad + \frac{1}{2}\sum_{cyc}(\psi_i^o)^2 - \frac{1}{4}\sum_{cyc}(\psi_i^o)^2 - \frac{1}{4}\sum_{cyc}(\theta_i)^4 - \frac{1}{2}\sum_{cyc}\psi_i^o\theta_i^2 \\
&= \left[\frac{1}{6}(\sum_{cyc}\psi_i^o)^3 - \frac{1}{6}(\sum_{cyc}(\psi_i^o)^3)\right] + \frac{1}{2}\sum_{cyc}(\psi_i^o)^4 - \frac{1}{4}\sum_{cyc}((\psi_i^o)^2 + \theta_i)^2 + \frac{1}{2}\sum_{cyc}(\psi_i^o)^2 - \\
&\frac{1}{4}\sum_{cyc}(\psi_i^o + \theta_i^2)^2.
\end{aligned} \tag{A.23}$$

On this basis, if we choose

$$\begin{aligned}
C_1(\boldsymbol{\gamma}) &= \sum_{i=1}^{d}\left[\frac{1}{6}(\sum_{cyc}\psi_i^o)^3 + \frac{1}{2}\sum_{cyc}(\psi_i^o)^4 + \frac{1}{2}\sum_{cyc}(\psi_i^o)^2\right], \\
C_2(\boldsymbol{\gamma}) &= \sum_{i=1}^{d}\left[\frac{1}{6}\sum_{cyc}(\psi_i^o)^3 + \frac{1}{4}\sum_{cyc}[(\psi_i^o)^2 + \theta_i]^2 + \frac{1}{4}\sum_{cyc}[\psi_i^o + \theta_i^2]^2\right],
\end{aligned} \tag{A.24}$$

then we have

$$\sum_{i=1}^{d}\psi_i^o\theta_i\omega_i = C_1(\boldsymbol{\gamma}) - C_2(\boldsymbol{\gamma}). \tag{A.25}$$

Here the Hessian matrix $\frac{\partial^2 C_1(\boldsymbol{\gamma})}{\partial^2(\boldsymbol{\gamma})}$ and $\frac{\partial^2 C_2(\boldsymbol{\gamma})}{\partial^2(\boldsymbol{\gamma})}$ is positive semi-definite:

$$\frac{\partial^2 C_1(\boldsymbol{\gamma})}{\partial^2(\boldsymbol{\gamma})} = \frac{\partial^2 C_2(\boldsymbol{\gamma})}{\partial^2(\boldsymbol{\gamma})} = [\boldsymbol{\gamma} + 6\boldsymbol{\gamma} \circ \boldsymbol{\gamma} + \mathbf{1_{4d*1}}] \circ \begin{bmatrix}\mathbf{1_{1*2d}} & \mathbf{0_{1*d}} & \mathbf{1_{1*d}}\end{bmatrix}^T \geq \mathbf{0_{4d*1}}, \tag{A.26}$$

where $\circ$ denotes the Hadamard product. Moreover,

$$\frac{\partial^2 D_{i1}(\gamma)}{\partial^2(\gamma)} = \frac{\partial^2 D_{i2}(\gamma)}{\partial^2(\gamma)} = \begin{bmatrix}\mathbf{1_{1*d}} & \mathbf{0_{1*d}} & \mathbf{1_{1*d}} & \mathbf{0_{1*d}}\end{bmatrix}^T \geq \mathbf{0_{4d*1}}. \tag{A.27}$$

$D_{i1}(\gamma), D_{i2}(\gamma)$ are also positive semi-definite.

On this basis, we further give the upper and lower bound of the convex function as follows:

lemma 5 *If function $F(\gamma)$ is differential and convex restricted by any simplex S, then $F^{tan}(\gamma) \leq F(\gamma) \leq F^{sec}(\gamma)$, where $\boldsymbol{\gamma_0} \in S$. In our paper, function $F(\cdot)$ can be chosen as $C_1(\cdot), C_2(\cdot), D_{i1}(\cdot), D_{i2}(\cdot)$, and $F^{tan}(\cdot), F^{sec}(\cdot)$ hold the same construction as in Formulation 10.* [9]

The proof of lemma. (5) The left part is intuitive. It is the tangent line equation of $F(\gamma)$. We only consider the right part by the convex property of $F(\gamma)$, whose construction is motivated by Pei and Zhu [2013]. We use $\gamma_i, i = 1, 2, ...4d$ to denote the value of γ on each dimension ($\lambda_i \in [0, 1]$, $\sum_{i=0}^{4d} \lambda_i = 1$):

$$\begin{aligned} F(\gamma) = F(\sum_{i=0}^{4d} \lambda_i S^i) \leq \sum_{i=0}^{4d} \lambda_i F(S^i) &= \sum_{i=0}^{4d} \lambda_i [F(S^0), F(S^1), ..., F(S^{4d})] \begin{bmatrix} S^0, ..., S^{4d} \\ 1, ..., 1 \end{bmatrix}^{-1} [\begin{smallmatrix} S^i \\ 1 \end{smallmatrix}] \\ &= [F(S^0), F(S^2), ..., F(S^{4d})] \begin{bmatrix} S^0, ..., S^{4d} \\ 1, ..., 1 \end{bmatrix}^{-1} [\begin{smallmatrix} \gamma \\ 1 \end{smallmatrix}] = F^{\text{sec}}(\gamma). \end{aligned} \tag{A.28}$$

Hence we have proved our lemma.

On this basis, we can claim (9) provides the lower bound of $\underline{f_S(Y_x = y)}$, namely $\underline{\underline{f_S(Y_x = y)}} \leq \underline{f_S(Y_x = y)}$. After the above difference-in-convex linear construction, we introduce the following lemma to approximate $\underline{f_S(Y_x = y)}$ by $\underline{\underline{f_S(Y_x = y)}}$:

lemma 6 $\forall S, \mid \underline{f_S(Y_x = y)} - \underline{\underline{f_S(Y_x = y)}} \mid \leq A * dia(S)^2$, *where* $A = \max_{\boldsymbol{\gamma} \in S_0} \| \frac{\partial(C_1(\boldsymbol{\gamma}) - C_2(\boldsymbol{\gamma}))}{\partial \boldsymbol{\gamma}} \| \frac{2(\sqrt{2}+1)\sqrt{d}}{\delta} + \max_{\boldsymbol{\gamma} \in S_0} \| \frac{\partial^2 C_1(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F + \frac{1}{2} \max_{\boldsymbol{\gamma} \in S_0} \| \frac{\partial^2 C_2(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F < +\infty.$

The proof of lemma. (6) Since $dia(S_0) < +\infty$, we have that each element of $\gamma \in S_0$ can be bounded, namely $\|\gamma\|_{+\infty} < +\infty$. Then $\| \frac{\partial C_1(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \|$, $\| \frac{\partial C_2(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}} \|$, $\| \frac{\partial^2 C_1(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F$, $\| \frac{\partial^2 C_2(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F$ are all finite. Here $\| \cdot \|$ denotes the Euclidean norm, and $\| \cdot \|_F$ denotes the Frobenius norm.

If the corresponding optimal solution of $\underline{f_S(Y_x = y)}$ and $\underline{\underline{f_S(Y_x = y)}}$ are denoted as $\underline{\boldsymbol{\gamma}}$ and $\underline{\underline{\boldsymbol{\gamma}}}$ ($\underline{\boldsymbol{\gamma}}, \underline{\underline{\boldsymbol{\gamma}}} \in S$). Then according to lemma. (5), $\mid \underline{f_S(Y_x = y)} - \underline{\underline{f_S(Y_x = y)}} \mid$ can be bounded as follows:

$$\begin{aligned} 0 \leq & \underline{f_S(Y_x = y)} - \underline{\underline{f_S(Y_x = y)}} \\ = & \mid C_1(\underline{\boldsymbol{\gamma}}) - C_1^{\text{tan}}(\underline{\underline{\boldsymbol{\gamma}}}) - C_2(\underline{\boldsymbol{\gamma}}) + C_2^{\text{sec}}(\underline{\underline{\boldsymbol{\gamma}}}) \mid \\ \leq & \mid C_1(\underline{\underline{\boldsymbol{\gamma}}}) - C_1^{\text{tan}}(\underline{\underline{\boldsymbol{\gamma}}}) - C_2(\underline{\underline{\boldsymbol{\gamma}}}) + C_2^{\text{sec}}(\underline{\underline{\boldsymbol{\gamma}}}) \mid + \mid C_1(\underline{\boldsymbol{\gamma}}) - C_1(\underline{\underline{\boldsymbol{\gamma}}}) - C_2(\underline{\boldsymbol{\gamma}}) + C_2(\underline{\underline{\boldsymbol{\gamma}}}) \mid \\ \overset{*}{\leq} & \underbrace{\mid C_1(\underline{\underline{\boldsymbol{\gamma}}}) - C_1^{\text{tan}}(\underline{\underline{\boldsymbol{\gamma}}}) \mid}_{(1)} + \underbrace{\mid C_2(\underline{\underline{\boldsymbol{\gamma}}}) - C_2^{\text{sec}}(\underline{\underline{\boldsymbol{\gamma}}}) \mid}_{(2)} + \underbrace{\mid (C_1(\underline{\boldsymbol{\gamma}}) - C_2(\underline{\boldsymbol{\gamma}})) - (C_1(\underline{\underline{\boldsymbol{\gamma}}}) - C_2(\underline{\underline{\boldsymbol{\gamma}}})) \mid}_{(3)} \end{aligned} \tag{A.29}$$

item (1):We consider the last line. The tangent line equation satisfies the following bound by Taylor expansion:

$$\mid C_1(\boldsymbol{\gamma}) - C_1^{\text{tan}}(\boldsymbol{\gamma}) \mid = O(\max_{\boldsymbol{\gamma} \in S_0} \| \frac{\partial^2 C_1(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F (dia(S))^2) = O(dia(S)^2), \tag{A.30}$$

[9] The matrix of the starting simplex $\begin{bmatrix} S_0^0, ..., S_0^{4d} \\ 1, ..., 1 \end{bmatrix}$ is reversible by the construction in lemma. (2). Moreover, the reversibility of $\begin{bmatrix} S^0, ..., S^{4d} \\ 1, ..., 1 \end{bmatrix}$, $S \in \mathcal{S}_k, k = 0, 1, ...$ still holds during bisection, since each bisection can be seen as a linear transformation between different columns.

item (2): On the other hand, note that $\boldsymbol{\gamma} = \sum\limits_{j=0}^{4d} \lambda_j S^j$, here $\sum\limits_{j=0}^{4d} \lambda_j = 1, \lambda_j \geq 0$:

$$\begin{aligned} \mid C_2(\boldsymbol{\gamma}) - C_2^{\text{sec}}(\boldsymbol{\gamma}) \mid =& - [C_2(S^0), C_2(S^1), ...C_2(S^{4d})] \begin{bmatrix} S^0, ..., S^{4d} \\ 1, ..., 1 \end{bmatrix}^{-1} \begin{bmatrix} \sum\limits_{i=0}^{4d} \lambda_i S^i \\ 1 \end{bmatrix} + C_2(\sum_{j=0}^{4d} \lambda_j S^j) \\ =& \sum_{j=0}^{4d} \lambda_j C_2(S^j) - C_2(\sum_{j=0}^{4d} \lambda_j S^j). \end{aligned} \tag{A.31}$$

We now aim to bound Eqn (A.31), inspired by [Budimir et al., 2001]. For simplicity, we use $\triangledown$ to denote the derivative of a vector. Notice that the convex function has the property:

$$C_2(\sum_{i=0}^{4d} \lambda_j S^j) - C_2(S^j) \geq \langle \triangledown C_2(S^j), \sum_{j=0}^{4d} \lambda_j S^j - S^j \rangle \tag{A.32}$$

By summation, we have

$$\begin{aligned} \text{(A.31)} =& \sum_{j=0}^{4d} \lambda_j C_2(S^j) - C_2(\sum_{i=0}^{4d} \lambda_j S^j) \\ \leq& \sum_{j=0}^{4d} \lambda_j \langle \triangledown C_2(S^j), -\sum_{j=0}^{4d} \lambda_j S^j + S^j \rangle \\ =& \sum_{j=0}^{4d} \lambda_j \langle \triangledown C_2(S^j), S^j \rangle - \langle \sum_{j=0}^{4d} \lambda_j S^j, \sum_{j=0}^{4d} \lambda_j \triangledown C_2(S^j) \rangle \end{aligned} \tag{A.33}$$

(A.33) equals to

$$\begin{aligned} &\frac{1}{2} \sum_{i=0}^{4d} \sum_{j=0}^{4d} \lambda_i \lambda_j \left[[\langle \triangledown C_2(S^j), S^j \rangle + \langle \triangledown C_2(S^i), S^i \rangle] - [\langle \triangledown C_2(S^j), S^i \rangle + \langle \triangledown C_2(S^i), S^j \rangle] \right] \\ =& \frac{1}{2} \sum_{i=0}^{4d} \sum_{j=0}^{4d} \lambda_i \lambda_j \langle S^i - S^j, \triangledown C_2(S^i) - \triangledown C_2(S^j) \rangle \\ \leq& \frac{1}{2} \sum_{i=0}^{4d} \sum_{j=0}^{4d} \lambda_i \lambda_j \| S^i - S^j \| \| \triangledown C_2(S^i) - \triangledown C_2(S^j) \| \\ \leq& \frac{1}{2} \sum_{i=0}^{4d} \sum_{j=0}^{4d} \lambda_i \lambda_j (\max_{\boldsymbol{\gamma} \in S} \| \frac{\partial^2 C_2(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F) \| S^i - S^j \|^2 \\ \leq& \frac{1}{2} \sum_{i=0}^{4d} \sum_{j=0}^{4d} \lambda_i \lambda_j (\max_{\boldsymbol{\gamma} \in S} \| \frac{\partial^2 C_2(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F) dia(S)^2 \\ \leq& \frac{1}{2} (\max_{\boldsymbol{\gamma} \in S_0} \| \frac{\partial^2 C_2(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F) dia(S)^2. \end{aligned} \tag{A.34}$$

We have

$$0 \leq \sum_{j=0}^{4d} \lambda_j C_2(S^j) - C_2(\sum_{j=0}^{4d} \lambda_j S^j) \leq \frac{1}{2} \max_{\boldsymbol{\gamma} \in S} \| \frac{\partial^2 C_2(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^2} \|_F (dia(S))^2 = O((dia(S))^2). \tag{A.35}$$

Thus

$$\text{Eqn (A.31)} = \sum_{j=0}^{4d} \lambda_j C_2(S^j) - C_2(\sum_{j=0}^{4d} \lambda_j S^j) = O((dia(S))^2). \tag{A.36}$$

item (3) We introduce an auxiliary optimization problem as follows:

$$\begin{aligned} &\min f(y,x) + C_1(\gamma) - C_2(\gamma) \\ &\text{subject to} : \boldsymbol{\phi} \in IR_{\Phi}, D_{i1}(\gamma) - D_{i2}(\gamma) = 1, \\ &\qquad \gamma \in \{\gamma^{'} : \exists \gamma^{''} \in S, \|\gamma^{'} - \gamma^{''}\| \leq \frac{(\sqrt{2}+1)\sqrt{d}}{\delta} dia(S)^2\} \cap S_0. \end{aligned} \tag{A.37}$$

Compared with the optimization problem of $\underline{f_S(Y_x = y)}$ (by (A.37) with an additional constraint $\gamma \in S$), (A.37) provides a relaxed constraint on γ. We denote the optimal solution of (A.37) as $\utilde{\gamma}$, and the optimal value as $\utilde{f_S(Y_x = y)}$.

On the one hand, (A.37) slightly relaxes the constraint $\gamma \in S_0$. Namely for each $\utilde{\gamma}$, there exists a corresponding $\gamma^{''} \in S$ with a distance less than $\frac{Q}{\delta} dia(S)^2$. Hence

$$\begin{aligned} &[C_1(\underline{\gamma}) - C_2(\underline{\gamma})] - [C_1(\utilde{\gamma}) - C_2(\utilde{\gamma})] \\ \leq &[C_1(\gamma^{''}) - C_2(\gamma^{''})] - [C_1(\utilde{\gamma}) - C_2(\utilde{\gamma})] \\ \leq &\max_{\gamma \in S_0} \| \frac{\partial (C_1(\gamma) - C_2(\gamma))}{\partial \gamma} \| \left(\frac{(\sqrt{2}+1)\sqrt{d}}{\delta} dia(S)^2 \right). \end{aligned} \tag{A.38}$$

On the other hand, we consider the optimal solution $\underline{\underline{\gamma}}$ of $\underline{\underline{f_S(Y_x = y)}}$. We identify the elements $\underline{\underline{\gamma}}^T = (\underline{\underline{(\boldsymbol{\psi^o})}}^T, \underline{\underline{\boldsymbol{\theta}}}^T, \underline{\underline{\boldsymbol{\psi}}}^T, \underline{\underline{\boldsymbol{\omega}^T}})$. Then we introduce an auxiliary solution as follows:

$$\utilde{\psi^o_i} = \frac{1}{\underline{\underline{\psi_i}}}, \utilde{\boldsymbol{\psi^o}} = (\utilde{\psi^o_1}, ...\utilde{\psi^o_d}), \utilde{\gamma}^{'} = ((\utilde{\boldsymbol{\psi^o}})^T, \underline{\underline{\boldsymbol{\theta}}}^T, \underline{\underline{\boldsymbol{\psi}}}^T, \underline{\underline{\boldsymbol{\omega}}}^T). \tag{A.39}$$

We will show that $\utilde{\gamma}^{'}$ is within the feasible region of (A.37). By identification in (A.39), the first row of constraints in (A.37) can be directly satisfied. Moreover, by Assumption. 2, we have

$$\begin{aligned} \|\utilde{\gamma}^{'} - \underline{\underline{\gamma}}\| = &\left(\sum_{i=1}^{d} (\frac{1}{\underline{\underline{\psi_i}}} - \underline{\underline{\psi^o_i}})^2 \right)^{\frac{1}{2}} \\ \leq &\frac{1}{\delta} (\sum_{i=1}^{d} (\underline{\underline{\psi_i}}\underline{\underline{\psi^o_i}} - 1)^2)^{\frac{1}{2}} \\ = &\frac{1}{\delta} (\sum_{i=1}^{d} (D_{i1}(\underline{\underline{\gamma}}) - D_{i2}(\underline{\underline{\gamma}}) - 1)^2)^{\frac{1}{2}} \\ \leq &\frac{\sqrt{d}}{\delta} \max_{i=1,...d} \left(1 - \left(D_{i1}(\underline{\underline{\gamma}}) - D_{i2}(\underline{\underline{\gamma}}) \right) \right) \\ \leq &\frac{\sqrt{d}}{\delta} \max_{i=1,...d} \left[\left(D^{sec}_{i1}(\underline{\underline{\gamma}}) - D^{tan}_{i2}(\underline{\underline{\gamma}}) \right) - \left(D_{i1}(\underline{\underline{\gamma}}) - D_{i2}(\underline{\underline{\gamma}}) \right) \right]. \end{aligned} \tag{A.40}$$

Symmetrically, we have

$$\|\utilde{\gamma}^{'} - \underline{\underline{\gamma}}\| \leq \frac{\sqrt{d}}{\delta} \max_{i=1,...d} \left[\left(D_{i1}(\underline{\underline{\gamma}}) - D_{i2}(\underline{\underline{\gamma}}) \right) - \left(D^{tan}_{i1}(\underline{\underline{\gamma}}) - D^{sec}_{i2}(\underline{\underline{\gamma}}) \right) \right]. \tag{A.41}$$

By the same strategy in **item(1)-(2)**, and noticing the fact that

$$\max_{\gamma \in S} \| \frac{\partial^2 D_{i1}(\gamma)}{\partial \gamma^2} \|_F = 2, \max_{\gamma \in S} \| \frac{\partial^2 D_{i2}(\gamma)}{\partial \gamma^2} \|_F = \sqrt{2}, \tag{A.42}$$

(A.40) and (A.41) can be combined as

$$\|\underset{\sim}{\gamma}' - \underline{\underline{\gamma}}\| \leq \frac{\sqrt{d}}{\delta}\min\{\frac{1}{2}*2+\sqrt{2}, 2+\frac{1}{2}\sqrt{2}\}dia(S)^2 = \frac{\sqrt{d}}{\delta}(\sqrt{2}+1)dia(S)^2. \tag{A.43}$$

Hence we claim this $\underset{\sim}{\gamma}'$ is within the feasible region of (A.37). Then

$$\begin{aligned}\Big[C_1(\underset{\sim}{\gamma}) - C_2(\underset{\sim}{\gamma})\Big] - \Big[C_1(\underline{\underline{\gamma}}) - C_2(\underline{\underline{\gamma}})\Big] &\leq \Big[C_1(\underset{\sim}{\gamma}') - C_2(\underset{\sim}{\gamma}')\Big] - \Big[C_1(\underline{\underline{\gamma}}) - C_2(\underline{\underline{\gamma}})\Big] \\ &\leq \max_{\gamma\in S}\|\frac{\partial(C_1(\gamma)-C_2(\gamma))}{\partial\gamma}\|\frac{\sqrt{d}}{\delta}(\sqrt{2}+1)dia(S)^2.\end{aligned} \tag{A.44}$$

Combining (A.38) and (A.44), we have

$$\mid (C_1(\underline{\gamma}) - C_2(\underline{\gamma})) - (C_1(\underline{\underline{\gamma}}) - C_2(\underline{\underline{\gamma}})) \mid\leq \max_{\gamma\in S_0}\|\frac{\partial(C_1(\gamma)-C_2(\gamma))}{\partial\gamma}\|\frac{2\sqrt{d}}{\delta}(\sqrt{2}+1)dia(S)^2. \tag{A.45}$$

Combination of item(1)-(3) Combining with Eqn (A.30) and Eqn (A.36), and recalling the bound in (A.29), we have:

$$\underline{\underline{f_S(Y_x=y)}} \leq \underline{f_S(Y_x=y)} \leq \underline{\underline{f_S(Y_x=y)}} + A*dia(S)^2. \tag{A.46}$$

In brief, we have $\mid \underline{\underline{f_S(Y_x=y)}} - \underline{f_S(Y_x=y)} \mid = O(dia(S)^2)$. Thus we have proved our lemma.

Remark 1 *We can do enhancement in **Bounding** as follows. It is through taking advantage of the information from the parent simplex pa(S)* [10] *and encapsulating the above bounding strategy into a recursive form during partitioning.*

$$\underline{\underline{f_S(Y_x=y)}} = \max\{\textbf{\textit{Bounding}}(pa(S)), \underline{\underline{f_S(Y_x=y)}}\} \tag{A.47}$$

(4) Global_error For the final preparation, we introduce the bisection theorem:

Theorem 2 *(Kearfott [1978], Theorem 3.1) When $\tilde{S}_{i_k}$ is bisectioned from S_0 by k times, we have $dia(\tilde{S}_{i_k}) \leq (\frac{\sqrt{3}}{2})^{\lfloor\frac{k}{4d}\rfloor}dia(S_0)$.*

On this basis, notice that lemma. (6) holds on each iteration, and $dia(S_0) < +\infty$, then we have

$$\mid \underline{f_{\tilde{S}_{i_{L_n}}}(Y_x=y)} - \underline{\underline{f_{\tilde{S}_{i_{L_n}}}(Y_x=y)}} \mid\leq A((\frac{3}{4})^{\lfloor\frac{L_n}{4d}\rfloor}) = O((\frac{3}{4})^{\frac{L_n}{4d}}), \tag{A.48}$$

where A is identified in our main text.

Until here we have proved procedure **(1)-(4)**, thus the main part of Theorem. 1 has been proved.

Additionally, consider the infinite case. Due to $L_n \geq log(n)$ (the worst case is that simplices set is bisectioned like a complete binary tree), we have $L \to +\infty$ when $n \to +\infty$, thus $\lim_{n\to+\infty} \mid \underline{f(Y_x=y)} - \underline{f^n_{opt}(f(Y_x=y))} \mid = 0$. Done.

A.6 EXTENSION TO THE ACE CASE

Taking advantage of PI-SFP, we can further achieve the valid bound of $ACE_{\boldsymbol{X}\to\boldsymbol{Y}}$. The above PI-SFP algorithm is to seek $\underline{f(Y_x=y)}$ when x is fixed. We do further extension to consider all values of $\boldsymbol{X}$ simultaneously. In this sense, we reorganize (2) to bound ACE as follows:

[10] $S_1 = \text{pa}(S_2)$ denotes S_2 is bisectioned from S_1.

$$
\begin{aligned}
&\min \sum_{x} \pi(x) \int_{Y^L}^{Y^U} y f(y,x) dy \\
&+\sum_{x} \pi(x) \sum_{i=1}^{d} \frac{\left(\int_{Y^L}^{Y^U} y f(y,u_i,x) dy\right) f(u_i,\neg x)}{f(u_i,x)} \\
&\text{subject to: } f(y,\boldsymbol{U},\boldsymbol{W},\boldsymbol{X}) \in \mathcal{F}.
\end{aligned}
\tag{A.49}
$$

Using the same strategy as in Section. 3-4, we can achieve the valid bound of $ACE_{\boldsymbol{X}\to\boldsymbol{Y}}$ in (A.10),

We first illustrate the construction of (A.49):

$$
\begin{aligned}
&\int_{X^L}^{X^U}\int_{Y^L}^{Y^U} f(Y_x = y)\pi(x) dx dy \\
=&\int_{X^L}^{X^U}\int_{Y^L}^{Y^U} \sum_{i=1}^{d}\left(\frac{f(y,u_i,x) f(u_i,\neg x)}{f(u_i,x)} + f(y,x)\right)\pi(x) dx dy \\
=&\sum_{x}\pi(x)\int_{Y^L}^{Y^U} y f(y,x) dy + \sum_{x}\pi(x)\sum_{i=1}^{d}\frac{\left(\int_{Y^L}^{Y^U} y f(y,u_i,x)dy\right) f(u_i,\neg x)}{f(u_i,x)}.
\end{aligned}
\tag{A.50}
$$

Let $X = \{x_1, x_2, ...x_{d_x}\}$. In this section, we extend PI-SFP method from bounding $f(Y_x = y)$ to bounding ACE. For simplicity, we extend the denotations in our main text as follows:

$$
\begin{aligned}
&\theta_{i|x} = \int_{Y^L}^{Y^U} y f(y, U = u_i, x) dy, && \boldsymbol{\theta_x} = (\theta_{1|x}, \theta_{2|x}, ...\theta_{d|x})^T, \\
&\psi_{i|x} = f(U = u_i, x), && \boldsymbol{\psi_x} = (\psi_{1|x}, \psi_{2|x}, ...\psi_{d|x})^T, \\
&\omega_{i|x} = f(U = u_i, \neg x), && \boldsymbol{\omega_x} = (\omega_{1|x}, \omega_{2|x}, ...\omega_{d|x})^T, \\
&\psi_{i|x}\psi^o_{i|x} = 1, && \boldsymbol{\psi^o_x} = (\psi^o_{1|x}, \psi^o_{2|x}, ...\psi^o_{d|x})^T, \\
&\boldsymbol{\phi_x} = (\boldsymbol{\theta_x}, \boldsymbol{\psi_x}, \boldsymbol{\omega_x}), && \gamma_x = \left((\boldsymbol{\psi^o_x})^T, \boldsymbol{\theta^T_x}, \boldsymbol{\psi^T_x}, \boldsymbol{\omega^T_x}\right)^T.
\end{aligned}
\tag{A.51}
$$

On this basis, the independent variables are transformed to $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_{x_1}, \boldsymbol{\gamma}_{x_2}, ...\boldsymbol{\gamma}_{x_{d_x}})$. Following the same strategy as in Section. 3 and Section. 4, we can relax the programming (A.49) in our main text as follows. It is a natural extension of (3) in Section. 3, by which we seek the valid bound of ACE:

$$
\begin{aligned}
\underline{ACE_{\boldsymbol{X}\to\boldsymbol{Y}}} = \min \sum_{x}\int_{Y^L}^{Y^U} \pi(x) y f(y,x) dy + \sum_{x}\sum_{i=1}^{d} \psi^o_{i|x}\theta_{i|x}\omega_{i|x}\pi(x),\\
\text{subject to } \forall x \in X, \psi^o_{i|x}\psi_{i|x} = 1, \boldsymbol{\phi_x} \in IR_{\boldsymbol{\Phi_x}} = IR^1_{\boldsymbol{\Phi_x}} \cap IR^2_{\boldsymbol{\Phi_x}},
\end{aligned}
\tag{A.52}
$$

where the set $IR^1_{\boldsymbol{\Phi}}$ is constructed as

$$
IR^1_{\boldsymbol{\Phi_x}} = \{\boldsymbol{\phi_x} : \begin{bmatrix} -\boldsymbol{I_{d*d}} \\ \boldsymbol{I_{d*d}} \end{bmatrix} \begin{bmatrix} (\int_{Y^L}^{Y^U} y f(y,\boldsymbol{W},x)dy)^T \\ f(\boldsymbol{W},x)^T \\ f(\boldsymbol{W},\neg x)^T \end{bmatrix}^T - \begin{bmatrix} -\overline{P(\boldsymbol{W}\mid\boldsymbol{U})} \\ \underline{P(\boldsymbol{W}\mid\boldsymbol{U})} \end{bmatrix} \boldsymbol{\phi_x} \geq \boldsymbol{0}\}.
\tag{A.53}
$$

$\boldsymbol{I_{d*d}}$ is the $d*d$ identity matrix. Moreover, the set $IR^2_{\boldsymbol{\Phi}}$ indicates the natural constraints by default:

$$
IR^2_{\boldsymbol{\Phi}_x} = \left\{ \boldsymbol{\phi}_x : \begin{bmatrix} \boldsymbol{1_{1*d}\theta_x} \\ \boldsymbol{1_{1*d}\phi_x} \\ \boldsymbol{1_{1*d}\omega_x} \end{bmatrix} = \begin{bmatrix} \int_{Y^L}^{Y^U} y f(y,x) dy \\ f(x) \\ f(\neg x)\} \end{bmatrix}, \forall i, \left\{ \begin{matrix} \theta_{i|x} \in [0, f(y,x)] \\ \phi_{i|x} \in (0, f(x)] \\ \omega_{i|x} \in [0, f(\neg x)] \end{matrix} \right\} \right\}.
\tag{A.54}
$$

$\mathbf{1_{1*d}}$ is the $1 * d$ all-ones vector. Then (7) in our main text is extended as

$$\begin{aligned}&\min \sum_x \pi(x)\{[C_1^{\tan}(\gamma_x) - C_2^{\sec}(\gamma_x)]\mathbb{1}_{\pi(x)>0} + [C_1^{\sec}(\gamma_x) - C_2^{\tan}(\gamma_x)]\mathbb{1}_{\pi(x)<0}\},\\ &\text{subject to } D_i^l(\gamma_x) \leq 1, D_i^u(\gamma_x) \geq 1, \forall i = 1, 2, ...d, \boldsymbol{\phi_x} \in IR_{\boldsymbol{\Phi_x}}.\end{aligned} \tag{A.55}$$

Here the function $C_k^{\tan}(\gamma_x), C_k^{\sec}(\gamma_x), k = 1, 2, D_i^l(\gamma_x), D_i^u(\gamma_x), i = 1, 2, ...d$ are all following (10) in our main text. After this construction, we adopt the same simplicial partition strategy as in our main text.

A.7 THE PROOF OF FURTHER DISCUSSIONS AND EXTENSIONS

This part is the supplement of the discussion in the main paper.

A.7.1 Discussion 1: the proof of lemma. (1) and the justification of our Assumption 1

For simplification, the denotations $Y = y, x$ are simplified as y and x, the denotation $\neg x$ is simplified as x^c, and d_w is simplified as $\mathscr{W}$. Samely, we use $\boldsymbol{E_{i*i}}$ to denote the $i * i$ identity matrix, $\boldsymbol{J_{i,j}}$ to denote the $i * j$ all-ones matrix, and $\boldsymbol{0_{i*j}}$ to denote the $i * j$ all-zero matrix.

- **Conclusion 1:** The tight lower bound of $f(Y_x = y)$ is $f(y, x)$.

We divide it into two parts. On the one hand, if $\mathscr{W} \geq d$, $P(\boldsymbol{W} \mid \boldsymbol{U})$ can be constructed as follows.

$$P(\boldsymbol{W} \mid \boldsymbol{U}) = \left[\begin{array}{c:c} \overbrace{\boldsymbol{P_{11}}}^{m*m} & \overbrace{\boldsymbol{P_{12}}}^{(d-m)*(d-m)} \\ \overbrace{\boldsymbol{P_{21}}}^{(\mathscr{W}-m)*m} & \overbrace{\boldsymbol{P_{22}}}^{(\mathscr{W}-d+m)*(d-m)} \end{array}\right], \tag{A.56}$$

where $\boldsymbol{P_{11}}, \boldsymbol{P_{12}}, \boldsymbol{P_{21}}, \boldsymbol{P_{22}}$ are matrices whose upper brackets indicate their rows and columns ($m \in [1, d-1]$). Specifically,

$$\begin{aligned}&\boldsymbol{P_{11}} = \sum_{i=1}^{m} f(W = w_i \mid y, x)\boldsymbol{E_{m*m}},\ \boldsymbol{P_{12}} = \sum_{i=1}^{d-m} f(W = w_i \mid x^c)\boldsymbol{E_{(d-m)*(d-m)}},\\ &\boldsymbol{P_{21}} = \begin{bmatrix} f(W = w_{m+1} \mid y, x) \\ ... \\ f(W = w_{\mathscr{W}} \mid y, x) \end{bmatrix}\boldsymbol{J_{1*m}},\ \boldsymbol{P_{22}} = \begin{bmatrix} f(W = w_{d-m+1} \mid x^c) \\ ... \\ f(W = w_{\mathscr{W}} \mid x^c) \end{bmatrix}\boldsymbol{J_{1*(d-m)}}.\end{aligned} \tag{A.57}$$

There is a solution for $f(y, \boldsymbol{U}, x)$, $f(\boldsymbol{U}, x^c)$ respectively as

$$\frac{1}{\sum\limits_{i=1}^{m} f(W = w_i \mid y, x)} \begin{bmatrix} f(y, W = w_1, x) \\ ... \\ f(y, W = w_m, x) \\ \boldsymbol{0_{(d-m)*1}} \end{bmatrix}, \quad \frac{1}{\sum\limits_{i=1}^{d-m} P(W = w_i \mid x^c)} \begin{bmatrix} \boldsymbol{0_{m*1}} \\ P(W = w_1, x^c) \\ ... \\ P(W = w_{d-m}, x^c) \end{bmatrix}. \tag{A.58}$$

Due to $f(y, \boldsymbol{U}, x) \circ f(\boldsymbol{U}, x^c) = 0$ and the condition $f(\boldsymbol{U}, x) > \boldsymbol{0}$, we have

$$f(Y_x = y) = f(y, x) + \sum_{i=1}^{d} \frac{f(y, u_i, x)}{P(u_i, x)} P(u_i, x^c) = f(y, x). \tag{A.59}$$

On the other hand, if $\mathscr{W} < d$, we make adjustments on (A.56) ($m_1 + m_2 \leq \mathscr{W}$):

$$\left[\begin{array}{c:c:c} \overbrace{\boldsymbol{P_{11}}}^{m_1*m_1} & \overbrace{\boldsymbol{P_{12}}}^{m_2*m_2} & \overbrace{\boldsymbol{P_3}}^{\mathscr{W}*(d-m_1-m_2)} \\ \overbrace{\boldsymbol{P_{21}}}^{(\mathscr{W}-m_1)*m_1} & \overbrace{\boldsymbol{P_{22}}}^{(\mathscr{W}-m_2)*m_2} & \end{array}\right], \tag{A.60}$$

Specifically,

$$\boldsymbol{P_{11}} = \sum_{i=1}^{m_1} P(W = w_i \mid y, x)\boldsymbol{E_{m_1*m_1}},\ \boldsymbol{P_{12}} = \sum_{i=1}^{m_2} P(W = w_i \mid x^c)\boldsymbol{E_{m_2*m_2}}.$$
$$\boldsymbol{P_{21}} = \begin{bmatrix} f(W = w_{m_1+1} \mid y, x) \\ ... \\ f(W = w_{\mathscr{W}} \mid y, x) \end{bmatrix} \boldsymbol{J_{1,m_1}},\ \boldsymbol{P_{22}} = \begin{bmatrix} f(W = w_{m_2+1} \mid x^c) \\ ... \\ f(W = w_{\mathscr{W}} \mid x^c) \end{bmatrix} \boldsymbol{J_{1*m_2}}. \tag{A.61}$$
$$\boldsymbol{P_3} = \frac{1}{\mathscr{W}} \boldsymbol{J_{\mathscr{W}*(d-m_1-m_2)}}.$$

Analogously, there is a solution for $f(y, \boldsymbol{U}, x)$, $f(\boldsymbol{U}, x^c)$ as follows respectively:

$$\frac{1}{\sum\limits_{i=1}^{m_1} f(W = w_i \mid y, x)} \begin{bmatrix} f(y, W = w_1, x) \\ ... \\ f(y, W = w_{m_1}, x) \\ \boldsymbol{0_{(d-m_1)*1}} \end{bmatrix},\ \frac{1}{\sum\limits_{i=1}^{m_2} P(W = w_i \mid x^c)} \begin{bmatrix} \boldsymbol{0_{m_1*1}} \\ P(W = w_1, x^c) \\ ... \\ P(W = w_{m_2}, x^c) \\ \boldsymbol{0_{(d-m_1-m_2)*1}} \end{bmatrix}. \tag{A.62}$$

In this case, we also have $f(Y_x = y) = f(y, x) + \sum\limits_{i=1}^{d} \frac{f(y,u_i,x)}{f(u_i,x)} f(u_i, x^c) = f(y, x)$. In conclusion, if no assumptions are imposed, we have $\min f(Y_x = y) = f(y, x)$. Proved.

- **Conclusion 2:** If $P(\boldsymbol{W} \mid \boldsymbol{U})$ is restricted to be left-reversible and $f(\boldsymbol{W} \mid \neg x) \neq f(\boldsymbol{W} \mid x, y)$, then the tight lower bound of $f(Y_x = y)$ is $f(y, x)$.

Without loss of generalization, we can assume that $\exists i_0 \in \{d, d+1, ...\mathscr{W}\}$, such that $f(W = w_{i_0} \mid \neg x) \neq f(W = w_{i_0} \mid x, y)$, or else we just need to relabel $\boldsymbol{W}$ in another order.

On this basis, we still follow the Construction. A.56 in the first part. The tight lower bound has already been proved as $f(y, x)$, thus we only need demonstrate that with some choice of m, $P(\boldsymbol{W} \mid \boldsymbol{U})$ is left-reversible with the above assumption. In practice, we choose $m = d - 1$. Then the $P(\boldsymbol{W} \mid \boldsymbol{U})$ is reformulated as

$$\left[\begin{array}{c:c} \overbrace{\boldsymbol{P_{11}}}^{(d-1)*(d-1)} & \overbrace{\boldsymbol{P_{12}}}^{1*1} \\ \overbrace{\boldsymbol{P_{21}}}^{(\mathscr{W}-d+1)*(d-1)} & \overbrace{\boldsymbol{P_{22}}}^{(\mathscr{W}-1)*1} \end{array}\right] := \left[\begin{array}{c:c} \sum\limits_{i=1}^{d-1} P(W = w_i \mid y, x)\boldsymbol{E_{(d-1)*(d-1)}} & \begin{bmatrix} f(W = w_1 \mid x^c) \\ f(W = w_2 \mid x^c) \\ ... \\ f(W = w_{\mathscr{W}} \mid x^c) \end{bmatrix} \\ \begin{bmatrix} f(W = w_d \mid y, x) \\ f(W = w_{d+1} \mid y, x) \\ ... \\ f(W = w_{\mathscr{W}} \mid y, x) \end{bmatrix} \boldsymbol{J_{1*(d-1)}} & \end{array}\right]. \tag{A.63}$$

We make equivalent denotations:

$$\begin{bmatrix} \overbrace{\boldsymbol{P_{12}^{'}}}^{(d-1)*1} \\ \overbrace{\boldsymbol{P_{22}^{'}}}^{(\mathscr{W}-d+1)*1} \end{bmatrix} := \begin{bmatrix} \overbrace{\boldsymbol{P_{12}}}^{1*1} \\ \overbrace{\boldsymbol{P_{22}}}^{(\mathscr{W}-1)*1} \end{bmatrix} \tag{A.64}$$

In the following part, we claim that we only need to prove $\boldsymbol{P_{22}^{'}} - \boldsymbol{P_{21}}\boldsymbol{P_{11}^{-1}}\boldsymbol{P_{12}^{'}} \neq \boldsymbol{0}$. We do the following algebraic distortion:

$$\begin{bmatrix} \boldsymbol{E_{(d-1)*(d-1)}} & \boldsymbol{0_{(d-1)*(\mathscr{W}-d+1)}} \\ -\boldsymbol{P_{21}}\boldsymbol{P_{11}^{-1}} & \boldsymbol{E_{(\mathscr{W}-d+1)*(\mathscr{W}-d+1)}} \end{bmatrix} * \begin{bmatrix} \boldsymbol{P_{11}} & \boldsymbol{P_{12}^{'}} \\ \boldsymbol{P_{21}} & \boldsymbol{P_{22}^{'}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{P_{11}} & \boldsymbol{P_{12}^{'}} \\ \boldsymbol{0_{(\mathscr{W}-d+1)*(d-1)}} & \boldsymbol{P_{22}^{'}} - \boldsymbol{P_{21}}\boldsymbol{P_{11}^{-1}}\boldsymbol{P_{12}^{'}} \end{bmatrix}. \tag{A.65}$$

According to the well-known Sylvester's inequality [Matsaglia and PH Styan, 1974]: $\forall \boldsymbol{A_{m*n}}, \boldsymbol{B_{n*p}}$, we have $\min\{rank(\boldsymbol{A}), rank(\boldsymbol{B})\} \geq rank(\boldsymbol{AB}) \geq rank(\boldsymbol{A}) + rank(\boldsymbol{B}) - n$. Then we have

$$rank\left(\begin{bmatrix} \boldsymbol{P_{11}} & \boldsymbol{P_{12}^{'}} \\ \boldsymbol{P_{21}} & \boldsymbol{P_{22}^{'}} \end{bmatrix}\right) = rank\left(\begin{bmatrix} \boldsymbol{P_{11}} & \boldsymbol{P_{12}^{'}} \\ \boldsymbol{0_{(\mathscr{W}-d+1)*(d-1)}} & \boldsymbol{P_{22}^{'}} - \boldsymbol{P_{21}}\boldsymbol{P_{11}^{-1}}\boldsymbol{P_{12}^{'}} \end{bmatrix}\right). \tag{A.66}$$

If $\boldsymbol{P}_{\boldsymbol{22}}^{'} - \boldsymbol{P}_{\boldsymbol{21}}\boldsymbol{P}_{\boldsymbol{11}}^{-\boldsymbol{1}}\boldsymbol{P}_{\boldsymbol{12}}^{'} = \mathbf{0}_{(\mathscr{W}-d+1)*(d-1)}$, then the right side of $rank()$ will be equal to $rank(\begin{bmatrix}\boldsymbol{P}_{\boldsymbol{11}}, \boldsymbol{P}_{\boldsymbol{12}}^{'}\end{bmatrix}) = d-1 < d$. On the other hand, if $\boldsymbol{P}_{\boldsymbol{22}}^{'} - \boldsymbol{P}_{\boldsymbol{21}}\boldsymbol{P}_{\boldsymbol{11}}^{-\boldsymbol{1}}\boldsymbol{P}_{\boldsymbol{12}}^{'} \neq \mathbf{0}_{(\mathscr{W}-d+1)*(d-1)}$, then it will turn to be d (full column rank). In conclusion, to demonstrate the left-reversibility of $P(\boldsymbol{W} \mid \boldsymbol{U})$, $\boldsymbol{P}_{\boldsymbol{22}}^{'} - \boldsymbol{P}_{\boldsymbol{21}}\boldsymbol{P}_{\boldsymbol{11}}^{-\boldsymbol{1}}\boldsymbol{P}_{\boldsymbol{12}}^{'} \neq \mathbf{0}$ is all we need.

If we use $[\cdot]_{(i)}$ to denote the i-th element of vector $i = d, ...\mathscr{W}$, then

$$[\boldsymbol{P}_{\boldsymbol{22}}^{'} - \boldsymbol{P}_{\boldsymbol{21}}\boldsymbol{P}_{\boldsymbol{11}}^{-\boldsymbol{1}}\boldsymbol{P}_{\boldsymbol{12}}^{'}]_{(i)} = \sum_{i=1}^{d-1} f(W = w_i \mid x^c) \left[\frac{f(W = w_i, x^c)}{\sum\limits_{i=1}^{d-1} f(W = w_i, x^c)} - \frac{f(y, W = w_i, x)}{\sum\limits_{i=1}^{d-1} f(y, W = w_i, x)}\right]. \tag{A.67}$$

We make the contradiction. If we have $\boldsymbol{P}_{\boldsymbol{22}}^{'} - \boldsymbol{P}_{\boldsymbol{21}}\boldsymbol{P}_{\boldsymbol{11}}^{-\boldsymbol{1}}\boldsymbol{P}_{\boldsymbol{12}}^{'} = \mathbf{0}_{(\mathscr{W}-d+1)*(d-1)}$, then

$$\begin{aligned} \|\boldsymbol{P}_{\boldsymbol{22}}^{'} - \boldsymbol{P}_{\boldsymbol{21}}\boldsymbol{P}_{\boldsymbol{11}}^{-\boldsymbol{1}}\boldsymbol{P}_{\boldsymbol{12}}^{'}\|_1 &= \sum_{i=1}^{d-1} f(W = w_i \mid x^c) \left[\frac{\sum\limits_{i=d}^{\mathscr{W}} f(W = w_i, x^c)}{\sum\limits_{i=1}^{d-1} f(W = w_i, x^c)} - \frac{\sum\limits_{i=d}^{\mathscr{W}} f(y, W = w_i, x)}{\sum\limits_{i=1}^{d-1} f(y, W = w_i, x)}\right] \\ &= \sum_{i=1}^{d-1} f(W = w_i \mid x^c) \left[\frac{f(x^c)}{\sum\limits_{i=1}^{d-1} f(W = w_i, x^c)} - \frac{f(y, x)}{\sum\limits_{i=1}^{d-1} f(y, W = w_i, x)}\right] \\ &= \sum_{i=1}^{d-1} f(W = w_i \mid x^c) \left[\frac{1}{\sum\limits_{i=1}^{d-1} f(W = w_i \mid x^c)} - \frac{1}{\sum\limits_{i=1}^{d-1} f(W = w_i \mid y, x)}\right] = 0. \end{aligned} \tag{A.68}$$

Thus we have $\sum\limits_{i=1}^{d-1} f(W = w_i \mid x^c) = \sum\limits_{i=1}^{d-1} f(W = w_i \mid y, x)$. Then we substitute it into Eqn (A.67), we have

$$f(W = w_i \mid x^c) - f(W = w_i \mid x, y) = 0, \forall i \in \{d, ...\mathscr{W}\}. \tag{A.69}$$

Contradiction! Hence we have $\boldsymbol{P}_{\boldsymbol{22}} - \boldsymbol{P}_{\boldsymbol{21}}\boldsymbol{P}_{\boldsymbol{11}}^{-\boldsymbol{1}}\boldsymbol{P}_{\boldsymbol{12}} \neq \mathbf{0}_{(\mathscr{W}-d+1)*(d-1)}$, and then $P(\boldsymbol{W} \mid \boldsymbol{U})$ in Construction. A.63 is left-reversible. Proved.

- **Conclusion 3:** If $P(\boldsymbol{W} \mid \boldsymbol{U})$ is restricted to be left-reversible and $f(\boldsymbol{W} \mid \neg x) = f(\boldsymbol{W} \mid x, y)$, then the tight lower bound of $f(Y_x = y)$ is $f(y \mid x)$.

If this assumption holds, we will have $\boldsymbol{P}_{\boldsymbol{22}}^{'} - \boldsymbol{P}_{\boldsymbol{21}}\boldsymbol{P}_{\boldsymbol{11}}^{-\boldsymbol{1}}\boldsymbol{P}_{\boldsymbol{12}}^{'} = \mathbf{0}$ in the above construction, thus $P(\boldsymbol{W} \mid \boldsymbol{U})$ will be irreversible and validates the condintion here. Hence we need another way.

According to the left-reversibility of $P(\boldsymbol{W} \mid \boldsymbol{U})$, we have

$$f(\boldsymbol{U} \mid x, y) = P(\boldsymbol{W} \mid \boldsymbol{U})^{-1} f(\boldsymbol{W} \mid x, y) = P(\boldsymbol{W} \mid \boldsymbol{U})^{-1} f(\boldsymbol{W} \mid x^c) = f(\boldsymbol{U} \mid x^c) \tag{A.70}$$

Then we have

$$
\begin{aligned}
f(Y_x = y) &= f(x,y) + \sum_{i=1}^{d} \frac{f(x,y,u_i)}{f(x,u_i)} f(u_i, x^c) \\
&= f(x,y) + f(x,y)f(x^c) \sum_{i=1}^{d} \frac{f(u_i \mid x,y)}{f(x,u_i)} f(u_i \mid x^c) \\
&= f(x,y) + f(x,y)f(x^c) \sum_{i=1}^{d} \frac{f(u_i \mid x,y)^2}{f(x,u_i)} \\
&\overset{*}{\geq} f(x,y) + f(x,y)f(x^c) \frac{(\sum_{i=1}^{d} f(u_i \mid x,y))^2}{\sum_{i=1}^{d} f(x,u_i)} \\
&= f(x,y) \left(1 + \frac{f(x^c)}{f(x)}\right) \\
&= f(y \mid x).
\end{aligned}
\tag{A.71}
$$

According to the Chauchy's inequality, the $' \geq '$ $(*)$ turns to be $' = '$ if and only if $f(\boldsymbol{U} \mid x,y) = f(\boldsymbol{U} \mid x)$. Combining with Eqn (A.70), we have $f(\boldsymbol{U} \mid x,y) = f(\boldsymbol{U} \mid x) = f(\boldsymbol{U} \mid x^c) = f(\boldsymbol{U})$. It holds if and only if $f(\boldsymbol{W} \mid x,y) = f(\boldsymbol{W} \mid x) = f(\boldsymbol{W} \mid x^c) = f(\boldsymbol{W})$, or else the lower bound is not tight.

Generalisability Our PI-SFP approach's generalizability can be highlighted in two ways: 1) PI-SFP can handle cases where either reversibility or total observability, or both, do not exist, which renders the literature on single-proxy control ineffective. 2) Extending PI-SFP to incorporate negative control (as shown in Fig.1(b) and Fig.1(c)) is an optional add-on and not a necessity. This simplicity eliminates the need for numerous assumptions, such as completeness and bridge function, which are present in previous double negative control literature [Miao et al., 2018, Cui et al., 2020, Tchetgen et al., 2020, Deaner, 2018, Shi et al., 2020, Singh, 2020, Nagasawa, 2018, Kallus et al., 2021]. The negative control extension to Fig. 1(b)-1(c) will be discussed in the next subsection.

Verifiability The feasibility of Assumption 1 has been suggested in previous work. Kuroki et al. Kuroki and Pearl [2014] suggested that the bounds $\underline{P(\boldsymbol{W} \mid \boldsymbol{U})}$ and $\overline{P(\boldsymbol{W} \mid \boldsymbol{U})}$ can be determined a priori through the Bayesian strategy [Greenland, 2005] and some re-calibration methods [Rothman et al., 2008, Selén, 1986]. In their "Head Start Program," they provide a detailed estimation of $P(\boldsymbol{W} \mid \boldsymbol{U})$ to support this claim.

Practical correspondence

- Some general cases (just conduct sampling upon $\boldsymbol{U}$): This hypothesis is commonly encountered in real situations, with Kuroki, Judea Pearl Kuroki and Pearl [2014] (page 4) and Li, Judea Pearl Li and Pearl [2022] specifically illustrating how to sample U to infer approximate/bounded estimates of $P(\boldsymbol{W} \mid \boldsymbol{U})$, and mentioning such sampling method has been previously and commonly used, such as fundamental work Greenland [2005], Rothman et al. [2008], Carroll et al. [2006], Selén [1986].
- Concrete example 1 for $P(\boldsymbol{W} \mid \boldsymbol{U})$ (recommendation system: $\boldsymbol{W}$ denotes the popularity of $\boldsymbol{U}$): There are also some more practical examples in our life. For instance, in the context of recommendation systems, a significant amount of work uses the representation of product-user features as a confounder $\boldsymbol{U}$. However, this representation often includes sensitive information, leading to incomplete observations of $\boldsymbol{U}$. Building upon this, the popularity ranking of products in different regions and time periods is publicly available information, which can be used as a proxy variable $\boldsymbol{W}$ for products. By analyzing the different purchasing tendencies of various demographic groups, we can obtain upper and lower bounds estimates for the transition matrix $P(\boldsymbol{W} \mid \boldsymbol{U})$. In this scenario, firstly, $P(\boldsymbol{W} \mid \boldsymbol{U})$ is often irreversible because popularity ranking information itself can be seen as an indicator/projection, and the information it carries is not as rich as the product features. Secondly, we typically can only estimate the upper and lower bounds of $P(\boldsymbol{W} \mid \boldsymbol{U})$ (through methods like Bayesian estimation), as the sampling estimation process for $\boldsymbol{U}$ is likely to be biased.
- Concrete example 2 for $P(\boldsymbol{W} \mid \boldsymbol{U})$ (privacy protection: $\boldsymbol{W}$ is the de-identified $\boldsymbol{U}$): Our PI-SFP framework is also related to privacy-protecting scenario. Survey collectors often need to gather some sensitive information $\boldsymbol{U}$. To obtain more accurate responses and avoid the risk of disclosing personal privacy, they often ask survey respondents to answer some yes-or-no questions using the Randomized Response method. We consider the survey results as $\boldsymbol{W}$, with the following steps: First, the survey respondent flips a coin (with equal probability of heads or tails), and only they know the result. If it lands heads, they answer the question truthfully; if it lands tails, they flip the coin again (with only them knowing the result); if the second toss is headed, they answer "Yes"; if it is tails, they answer "No". With this setup,

even if we do not know $\boldsymbol{U}$, we can deduce $P(W)$ and $P(\boldsymbol{W}|\boldsymbol{U}) = [3/4, 1/4; 1/4, 3/4]$. Of course, this simple setup may still expose other sensitive information, such as the joint distribution of $P(\boldsymbol{Y}, \boldsymbol{U}, \boldsymbol{X})$, which is something the government would not want to see or make public. Therefore, in practical use, for highly confidential information which requires the strongest privacy protection, we tend to develop a more complex/dynamic/irreversible $P(\boldsymbol{W}|\boldsymbol{U})$ (i.e., privacy-protecting algorithms) and manually set its upper and lower bounds.

A.7.2 Discussion 2: algorithm comparison and acceleration

In this section, we discuss two additional optimization methods which are potential to solve our partial observability problem. We subsequently illustrate the superior performance of PI-SFP compared to these methods. Furthermore, we introduce a novel pruning strategy supported by a local optimization method, which accelerates the optimization process.

Algorithm comparison The author of Shen et al. [2017] derived an $\varepsilon-$approximation method that can be utilized in our problem, and the outcome will lie within $[\underline{f(Y_x = y)}, (1+\varepsilon)\underline{f(Y_x = y)}]$. However, this algorithm exhibits an exponential time complexity for the dimension d_u, making it challenging to operate effectively in high-dimensional confoundings. Additionally, an iterative algorithm was developed in Le Thi et al. [2014] to search the Karush-Kuhn-Tucker (KKT) point of the difference-in-convex (DC) problem, which can be applied to (3). Nevertheless, KKT theory cannot guarantee global optimality compared to our PI-SFP.

Algorithm acceleration To expedite the PI-SFP process, we propose setting sufficient criteria to evaluate whether the current partition contains the optimal solution. If the criteria are not met, we can delete the branch online and narrow our search. To this end, we propose a new auxiliary algorithm specifically designed to search for the local minimum of $f(Y_x = y)$, which serves as an upper-bound of $\underline{f(Y_x = y)}$. We achieve this by implementing the algorithm in the sub-simplex S. Specifically, if the optimal value $\underline{\underline{f_S(Y_x = y)}}$ is even larger than the local minimum, then it will be larger than $\underline{f(Y_x = y)}$. Therefore, we claim that this partition must not include the optimal solutions and can be removed permanently. The auxiliary local optimization algorithm is provided as follows. The principle of our algorithm is based on the lemma:

lemma 7 *$\forall i, j$, if we make adjustment:*

$$\begin{bmatrix} \breve{\theta}_i & \breve{\theta}_j \\ \breve{\psi}_i & \breve{\psi}_j \\ \breve{\omega}_i & \breve{\omega}_j \end{bmatrix} = \begin{bmatrix} \theta_i & \theta_j \\ \psi_i & \psi_j \\ \omega_i & \omega_j \end{bmatrix} \begin{bmatrix} \alpha & 1-\alpha \\ 1-\alpha & \alpha \end{bmatrix}, \textit{ where } \alpha \in \begin{cases} (0,1] \textit{ if } (\frac{\theta_i}{\theta_j} - \frac{\psi_i}{\psi_j})(\frac{\omega_i}{\omega_j} - \frac{\psi_i}{\psi_j}) \geq 0. \\ [1, +\infty) \textit{ if } (\frac{\theta_i}{\theta_j} - \frac{\psi_i}{\psi_j})(\frac{\omega_i}{\omega_j} - \frac{\psi_i}{\psi_j}) \leq 0, \end{cases} \tag{A.72}$$

Then we have

$$\sum_{m=i,j} \frac{\breve{\theta}_m}{\breve{\psi}_m} \breve{\omega}_m \leq \sum_{m=i,j} \frac{\theta_m}{\psi_m} \omega_m. \tag{A.73}$$

We consider the case $\alpha \in (0,1)$, and the second case is symmetric. Due to $(\frac{\theta_i}{\theta_j} - \frac{\psi_i}{\psi_j})(\frac{\omega_i}{\omega_j} - \frac{\psi_i}{\psi_j}) \geq 0$, we have

$$(\theta_i - \theta_j)(\omega_j - \omega_i)\psi_i\psi_j + (\theta_j\omega_j\psi_i - \theta_i\omega_i\psi_j)(\psi_j - \psi_i) \leq 0. \tag{A.74}$$

If we denote that

$$Q_{ij} := \alpha\psi_i\psi_j(\theta_i - \theta_j)(\omega_j - \omega_i)\left[(1-\alpha)\psi_i + \alpha\psi_j\right] + (\theta_j\omega_j\psi_i - \theta_i\omega_i\psi_j)\left[(1-\alpha)\psi_i + \alpha\psi_j\right]\psi_j. \tag{A.75}$$

Then Eqn (A.74) is equal to

$$Q_{ij} \leq -Q_{ji}. \tag{A.76}$$

Furthermore, we find

$$Q_{ij} = \psi_j\breve{\psi}_j\left[\alpha\psi_i(\theta_i - \theta_j)(\omega_j - \omega_i) + \theta_j\omega_j\psi_i - \theta_i\omega_i\psi_j\right] = \frac{1}{1-\alpha}\psi_j\breve{\psi}_j\left[\breve{\theta}_i\breve{\omega}_i\psi_i - \theta_i\omega_i\breve{\psi}_i\right]. \tag{A.77}$$

Hence Eqn (A.76) can be transformed as

$$\psi_j\breve{\psi}_j\left[\breve{\theta}_i\breve{\omega}_i\psi_i - \theta_i\omega_i\breve{\psi}_i\right] < -\psi_i\breve{\psi}_i\left[\breve{\theta}_j\breve{\omega}_j\psi_j - \theta_j\omega_j\breve{\psi}_j\right] \tag{A.78}$$

Hence

$$\sum_{m=i,j}\frac{\breve{\theta}_m}{\breve{\psi}_m}\breve{\omega}_m \leq \sum_{m=i,j}\frac{\theta_m}{\psi_m}\omega_m. \tag{A.79}$$

Thus we have proved. By this strategy, we should choose suitable α to satisfy $\boldsymbol{\phi} \in IR_{\boldsymbol{\Phi}}$, namely that $f(y, \boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X}) \in \widetilde{\mathcal{F}}$.

A.7.3 Discussion 3: Extension to Fig. 1(b) and 1(c)

Fig. 1(a) Our algorithm PI-SFP mainly focuses on Fig. 1(a). Moreover, if $\boldsymbol{W} \to \boldsymbol{Y}$ is added, the optimization problem will be transferred under Assumption. 1.

$$\begin{aligned}
&\min f(y \mid \boldsymbol{U}, \boldsymbol{X}=x) f(\boldsymbol{U}),\\
&\text{subject to } f(y \mid \boldsymbol{U}, \boldsymbol{X}=x) f(\boldsymbol{U} \mid \boldsymbol{X}=x) = f(y \mid x),\\
&\text{where } f(\boldsymbol{U} \mid \boldsymbol{X}=x) \text{ satisfies}\\
&\begin{bmatrix} \overline{f(\boldsymbol{W} \mid \boldsymbol{U})} f(\boldsymbol{U} \mid \boldsymbol{X}=x) - f(\boldsymbol{W} \mid \boldsymbol{X}=x) \\ f(\boldsymbol{W} \mid \boldsymbol{X}=x) - \underline{f(\boldsymbol{W} \mid \boldsymbol{U})} f(\boldsymbol{U} \mid \boldsymbol{X}=x) \end{bmatrix} \geq 0,\\
&\begin{bmatrix} \overline{f(\boldsymbol{W} \mid \boldsymbol{U})} f(\boldsymbol{U}) - f(\boldsymbol{W}) \\ f(\boldsymbol{W}) - \underline{f(\boldsymbol{W} \mid \boldsymbol{U})} f(\boldsymbol{U}) \end{bmatrix} \geq 0.
\end{aligned} \tag{A.80}$$

Notice that the feasible region of $f(y \mid \boldsymbol{U}, \boldsymbol{X}=x)$ and $f(\boldsymbol{U})$ is even more irregular than in (3). Nevertheless, we can still adopt an analogous strategy to PI-SFP to approximate its optimal value. This remains a topic for future research.

Fig. 1(b) and 1(c) The double negative control via introducing auxiliary exposure $\boldsymbol{Z}$ can enhance our estimation. Due to the fact $f(y \mid u, x) = f(y \mid u, x, Z)$, we have:

$$\begin{aligned}
&\underline{f(Y_x = y)} := f(y, x) + \max_{\mathcal{Z} \subseteq \boldsymbol{Z}}\\
&\min_{f(y, \boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X}, \mathcal{Z}) \in \widetilde{\mathcal{F}}_Z} \sum_{u=1}^{d} \frac{f(y, u_i, x, z \in \mathcal{Z}) f(u_i, \neg x)}{f(u_i, x, z \in \mathcal{Z})}.
\end{aligned} \tag{A.81}$$

The feasible region $\widetilde{\mathcal{F}}_Z$ of $f(y, \boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X}, \mathcal{Z})$ is constructed analogously to that of $\widetilde{\mathcal{F}}$. Specifically, for each subset $\mathcal{Z} \subseteq Z$, we can apply PI-SFP and select the maximum of them as the best lower bound of $\underline{f(Y_x = y)}$. We will conduct a more detailed analysis in our subsequent work, particularly in comparison to the performance of the single-proxy control.

$\widetilde{\mathcal{F}}_Z$ is identified as follows:

We denote

$$\begin{array}{ll} \theta_i = f(y, u_i, x, z \in \mathcal{Z}) & \boldsymbol{\theta}_{\boldsymbol{\mathcal{Z}}} \;= (\theta_1, \theta_2, \ldots \theta_d)^T \\ \psi_i = f(u_i, x, z \in \mathcal{Z}) \;\;, & \boldsymbol{\psi}_{\boldsymbol{\mathcal{Z}}} \;= (\psi_1, \psi_2, \ldots \psi_d)^T, \;\boldsymbol{\phi}_{\boldsymbol{\mathcal{Z}}} = \left(\boldsymbol{\theta}_{\boldsymbol{\mathcal{Z}}}\; \boldsymbol{\psi}_{\boldsymbol{\mathcal{Z}}}\; \boldsymbol{\omega}_{\boldsymbol{\mathcal{Z}}}\right). \\ \omega_i = f(u_i, \neg x) & \boldsymbol{\omega}_{\boldsymbol{\mathcal{Z}}} \;= (\omega_1, \omega_2, \ldots \omega_d)^T \end{array} \tag{A.82}$$

where $f(y, \boldsymbol{W}, \boldsymbol{U}, \boldsymbol{X}, \mathcal{Z}) \in \widetilde{\mathcal{F}}_Z = \{\boldsymbol{\phi}_{\mathcal{Z}} \in IR_Z, IR_Z = IR_Z^1 \cap IR_Z^2\}$ leads to the following constraints that we really use:

$$IR_Z^1 = \{\boldsymbol{\phi}_Z : \begin{bmatrix} -\boldsymbol{I_{d*d}} \\ \boldsymbol{I_{d*d}} \end{bmatrix} \begin{bmatrix} f(y, \boldsymbol{W}, x, z \in \mathcal{Z})^T \\ f(\boldsymbol{W}, x, z \in \mathcal{Z})^T \\ f(\boldsymbol{W}, \neg x, z \in \mathcal{Z})^T \end{bmatrix}^T - \begin{bmatrix} -\overline{P(\boldsymbol{W} \mid \boldsymbol{U})} \\ \underline{P(\boldsymbol{W} \mid \boldsymbol{U})} \end{bmatrix} \boldsymbol{\phi}_{\mathcal{Z}} \geq \boldsymbol{0}\}. \tag{A.83}$$

Moreover, the set $IR_{\boldsymbol{\Phi}}^2$ indicates the natural constraints by default:

$$IR_Z^2 = \left\{ \boldsymbol{\phi}_Z : \begin{bmatrix} \boldsymbol{1}^T \boldsymbol{\theta} \\ \boldsymbol{1}^T \boldsymbol{\phi} \\ \boldsymbol{1}^T \boldsymbol{\omega} \end{bmatrix} = \begin{bmatrix} f(y, x, z \in \mathcal{Z}) \\ f(x, z \in \mathcal{Z}) \\ f(\neg x, z \in \mathcal{Z})\} \end{bmatrix}, \forall i, \begin{Bmatrix} \theta_i \in [0, f(y, x, z \in \mathcal{Z})] \\ \phi_i \in (0, f(x, z \in \mathcal{Z})] \\ \omega_i \in [0, f(\neg x, z \in \mathcal{Z})] \end{Bmatrix} \right\}. \tag{A.84}$$

A.7.4 Discussion 4: the continuous confoundings

Assumption 3 *(partial observability assumption for continuous confoundings) $P(\boldsymbol{W} \mid \boldsymbol{U}) \in \mathscr{P}$, where* $\mathscr{P} = \{P(\boldsymbol{W} \mid \boldsymbol{U}) : \begin{bmatrix} \overline{P(\boldsymbol{W} \mid \boldsymbol{U} \in [u_{i-1}, u_i])} - P(\boldsymbol{W} \mid \boldsymbol{U} \in [u_{i-1}, u_i]) \\ P(\boldsymbol{W} \mid \boldsymbol{U} \in [u_{i-1}, u_i]) - \underline{P(\boldsymbol{W} \mid \boldsymbol{U} \in [u_{i-1}, u_i])} \end{bmatrix}$ *is non-negative*$\}$.

Assumption 4 *(Lipschitz condition) $\forall y \in \boldsymbol{Y}, \forall \{u^{'}, u^{''}\} \in \boldsymbol{U}$, we have $\left|\frac{f(y,u^{'},x)-f(y,u^{''},x)}{f(u^{'},x)-f(u^{''},x)}\right| \leq C_1, \left|\frac{f(u^{'},x)-f(u^{''},x)}{u^{'}-u^{''}}\right| \leq C_2$, where C_1, C_2 are positive constants.*

lemma 8 *Suppose that Assumption. 3-4 hold. $\forall i \in \{0, 1, ...d-1\}, \forall u \in [u_i, u_{i+1}]$, we have*

$$\frac{\int_{u_i}^{u_{i+1}} f(y,u,x)du}{\int_{u_i}^{u_{i+1}} f(u,x)du} \leq \frac{f(y,u,x)}{f(u,x)} \frac{1}{1-\frac{1}{2}C_2\eta} + \frac{\frac{1}{2}C_1C_2\eta}{1-\frac{1}{2}C_2\eta}. \tag{A.85}$$

On the other hand,

$$\frac{\int_{u_i}^{u_{i+1}} f(y,u,x)du}{\int_{u_i}^{u_{i+1}} f(u,x)du} \geq \frac{f(y,u,x)}{f(u,x)} \frac{1}{1+\frac{1}{2}C_2\eta} - \frac{\frac{1}{2}C_1C_2\eta}{1+\frac{1}{2}C_2\eta}. \tag{A.86}$$

The proof of lemma. (8) We do partition on the confounding interval $[U^L, U^U]$ as $[u_0, u_1, u_2, ..., u_{d-1}, u_d]$, where $u_0 = U^L, u_d = U^U$. The independent variables is re-defined by

$$\theta_i = f(y, U \in [u_i, u_{i+1}], x), \psi_i = f(U \in [u_i, u_{i+1}], x), \omega_i = f(U \in [u_i, u_{i+1}], \neg x), \tag{A.87}$$

$$\begin{aligned} \forall u^{'} \in [u_i, u_{i+1}], \frac{\int_{u_i}^{u_{i+1}} f(y,u,x)du}{\int_{u_i}^{u_{i+1}} f(u,x)du} &\leq \frac{\int_{u_i}^{u_{i+1}} \left[f(y,u^{'},x) + C_1\left|f(u,x) - f(u^{'},x)\right|\right] du}{\int_{u_i}^{u_{i+1}} [f(u^{'},x) + (f(u,x) - f(u^{'},x))] du} \\ &\leq \frac{f(y,u^{'},x)(u_{i+1}-u_i) + C_1C_2\frac{1}{2}(u_{i+1}-u_i)^2}{f(u^{'},x)(u_{i+1}-u_i) - C_2\frac{1}{2}(u_{i+1}-u_i)^2} \\ &\leq \frac{\frac{f(y,u^{'},x)}{f(u^{'},x)} + \frac{\frac{1}{2}C_1C_2\eta\delta}{f(u^{'},x)}}{1-\frac{\frac{1}{2}C_2\eta\delta}{f(u^{'},x)}} \\ &\leq \frac{f(y,u^{'},x)}{f(u^{'},x)} \frac{1}{1-\frac{1}{2}C_2\eta} + \frac{\frac{1}{2}C_1C_2\eta}{1-\frac{1}{2}C_2\eta}. \end{aligned} \tag{A.88}$$

Here $i = 0, 1, ...d-1$. Analogously, we can prove the other direction. Thus we have proved the lemma.

Corollary 1 *(PI-SFP's error for continuous confoundings) Suppose that Assumption. 3-4 holds. When U is continuous, and $\max\limits_{i \in \{1,2,...d\}} |u_i - u_{i-1}| < \eta\delta$. Then $\underline{f(Y_x = y)} \leq \lim\limits_{n \to +\infty} \underline{\underline{f_{opt}^n(Y_x = y)}} \leq \frac{1}{1-\frac{1}{2}C_2\eta} \underline{f(Y_x = y)} + \frac{C_1 f(\neg x) - f(y,x)}{2-C_2\eta} C_2\eta$.*

Then we prove this corollary. **The proof of Corollary. 1** If we use $\underline{f(y, \boldsymbol{U}, x)}, \underline{f(\boldsymbol{U}, x)}, \underline{f(\boldsymbol{U}, \neg x)}$ to denote the optimal solution of the optimal value $\underline{f(Y_x = y)}$ in the continuous case, then we have

$$\begin{aligned}
&\underline{f(Y_x = y)} - f(y,x) \\
=&\int_{U^L}^{U^U} \frac{\underline{f(y,u,x)}}{\underline{f(u,x)}} \underline{f(u,\neg x)}du \\
=&\sum_{i=0}^{d-1} \int_{u_i}^{u_{i+1}} \frac{\underline{f(y,u,x)}}{\underline{f(u,x)}} \underline{f(u,\neg x)}du \\
\geq&\sum_{i=0}^{d-1} \int_{u_i}^{u_{i+1}} \left[\frac{\int_{u_i}^{u_{i+1}} \underline{f(y,u,x)}du}{\int_{u_i}^{u_{i+1}} \underline{f(u,x)}du} - \frac{\frac{1}{2}C_1C_2\eta}{1-\frac{1}{2}C_2\eta} \right] (1 - \frac{1}{2}C_2\eta)\underline{f(u,\neg x)}du \\
=&(1-\frac{1}{2}C_2\eta) \sum_{i=0}^{d-1} \frac{\int_{u_i}^{u_{i+1}} \underline{f(y,u,x)}du}{\int_{u_i}^{u_{i+1}} \underline{f(u,x)}du} \int_{u_i}^{u_{i+1}} \underline{f(u,\neg x)}du - \frac{1}{2}C_1C_2\eta f(\neg x).
\end{aligned} \tag{A.89}$$

Here $\{\int_{u_i}^{u_{i+1}} \underline{f(y,u,x)}du, \int_{u_i}^{u_{i+1}} \underline{f(u,x)}du, \int_{u_i}^{u_{i+1}} \underline{f(u,\neg x)}du, i = 0, 1, ...d-1\}$ is within the feasible region of PI-SFP in the discrete case. Then we have

$$\begin{aligned}
(\text{A.89}) \geq (1-\frac{1}{2}C_2\eta)\left(\lim_{n\to+\infty} \underline{\underline{f_{opt}^n(Y_x = y)}} - f(y,x) \right) - \frac{1}{2}C_1C_2\eta f(\neg x) \\
\underline{f(Y_x = y)} \geq (1-\frac{1}{2}C_2\eta) \lim_{n\to+\infty} \underline{\underline{f_{opt}^n(Y_x = y)}} + \frac{1}{2}C_2\eta f(y,x) - \frac{1}{2}C_1C_2\eta f(\neg x) \\
\lim_{n\to+\infty} \underline{\underline{f_{opt}^n(Y_x = y)}} \leq \frac{1}{1-\frac{1}{2}C_2\eta}\underline{f(Y_x = y)} + \frac{\frac{1}{2}C_1 f(\neg x) - \frac{1}{2}f(y,x)}{1-\frac{1}{2}C_2\eta}C_2\eta
\end{aligned} \tag{A.90}$$

On the other hand, each optimal solution by PI-SFP corresponds to a solution in the continuous case. Namely if the discrete PI-SFP's optimal solution is denoted as $\{\int_{u_i}^{u_{i+1}} f(y,u,x)du, \int_{u_i}^{u_{i+1}} f(u,x)du, \int_{u_i}^{u_{i+1}} f(y,u,\neg x)du, i = 0, 1, ...d-1\}$. Then we can construct

$$\begin{aligned}
f^{opt}(y,u,x) &= \frac{\int_{u_i}^{u_{i+1}} f(y,u,x)du}{u_{i+1}-u_i}, u \in [u_i, u_{i+1}). \\
f^{opt}(u,x) &= \frac{\int_{u_i}^{u_{i+1}} f(u,x)du}{u_{i+1}-u_i}, u \in [u_i, u_{i+1}). \\
f^{opt}(u,\neg x) &= \frac{\int_{u_i}^{u_{i+1}} f(u,\neg x)du}{u_{i+1}-u_i}, u \in [u_i, u_{i+1})
\end{aligned} \tag{A.91}$$

as one of the solution in the continuous case. Hence $\lim_{n\to+\infty} \underline{\underline{f_{opt}^n(Y_x = y)}} \geq \underline{f(Y_x = y)}$. We have finished the proof.

A.7.5 Comment on Theorem 1: the relationship between L_n and n

Remark 2 *(Hardness of establishing functional associations between L_n and n) The worst-case scenario for L_n is $L_n = \lfloor log(n) \rfloor + 1$. In this situation, PI-SFP is equivalent to the method of exhaustion, which exhibits slow polynomial convergence, as shown in Theorem 1, with a rate of $O(n^{-\alpha})$, where $\alpha = \frac{1}{4d}log(\frac{2}{\sqrt{3}})$. However, empirical evidence suggests that this scenario is rare, and in the simulation section, the convergence rate is faster than $O(n^{-\alpha})$. Additionally, pruning strategies can be employed to further improve the convergence rate, which is discussed in Section. 7.*

Despite this empirical observation, it is well beyond the scope of this paper to theoretically estimate L_n w.r.t n. During iteration, each optimal solution (converging point) may be covered by increasing number of nested sequences[11]. These sequences possess different lengths and are difficult to estimate. More seriously, the number of optimal solutions is not necessarily finite either, namely $|\mathbf{\Phi}_{opt}| < +\infty$ may not be guaranteed.

[11] Notice that it is equivalent to guarantee each converging point is covered by finite partitions. It resorts to the regularity condition of simplices (identified in [Ciarlet, 2002]). However, whether Longest-edge bisection can promise a family of regular partitions is still a conjecture [Korotov et al., 2016] to be solved.

Notice that $L_n = O(n)$ when each ϕ_{opt} is partitioned via finite number of simplices and $|\boldsymbol{\phi_{opt}}| < +\infty$ in the infinite process. By this motivation, we aim to prove that such finiteness is true under a fairly broad assumption mentioned in Ciarlet [2002], Korotov et al. [2016]. In fact, we address this conjecture and it is our by-contribution:

what is the maximum intersection number of regular simplicial partitions? In other words, for arbitrary point in $\mathcal{S}_0$, what is the maximum number of simplices it can be affiliated with during partitioning, under fairly broad assumption?

Assumption 5

$$vol(S) \geq \eta h(S)^d. \tag{A.92}$$

Here $vol(S)$ denotes the volumn of simplex S, $h(S)$ denotes the longest edge of S, and $d \geq 2$ is the dimension.

Roughly speaking, the regularity assumption guarantees the simplex would not degenerate to the hyper-plane, or else $\lim\limits_{vol(S)\to 0} \frac{vol(S)}{h(S)^d} = 0$. In other words, $\frac{vol(S)}{h(S)^d}$ can be higher when the simplex “seems to be regular"", namely each edge keeps the same length. In addition, the simplicial partitoins during partitioning are denoted as $\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2$.... Our problem can be summarized as follows:

Theorem 3 *Suppose that Ass. 5 holds. Each point in $\mathcal{S}_0$ is included within at most $\frac{1}{\eta}\left(\frac{2e\pi}{d}\right)^{\frac{d}{2}}$ simplices.*

proof For $\gamma_0 \in \mathcal{S}_0$, we construct a ball $\mathcal{B}(\gamma_0, r)$. We use $A(\cdot)$ to denote the surface area of sphere:

$$A\left(\mathcal{B}\left(\gamma_0, r\right) \cap \mathcal{S}_0\right) = \sum_{S \in \mathcal{S}_k, \gamma_0 \in S} A\left(\mathcal{B}\left(\gamma_0, r\right) \cap S\right). \tag{A.93}$$

On the one hand, the LHS of Eqn. A.93 can be upper bounded as:

$$A\left(\mathcal{B}\left(\gamma_0, r\right) \cap S_0\right) \leq A\left(\mathcal{B}\left(\gamma_0, r\right)\right) = \frac{2\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}\right)} r^{d-1} < +\infty. \tag{A.94}$$

On the other hand, we calculate the RHS of Eqn. A.93. To solve it, we take advantage of the following integral (we take γ_0 as the origin):

$$\int_{\mathcal{W}_S \gamma \geq 0} e^{-\|\gamma\|^2} d\gamma. \tag{A.95}$$

Here for each term in the right side, the sphere of the ball is cut by certain facets of each S, whose normal vector is denoted as a set $\mathcal{W}_S$, whose each row denotes a $1 * 4d$ normal vector. Without loss of generation, for any facets, we assume that its normal vector points to the remaining supporting vector, namely their inner product is positive.

Here $\|\cdot\|$ also denotes the Euclidean norm. If we use the polar coordinates, we can get a new expression by differential element method:

$$\int_{\mathcal{W}_S \gamma \geq 0} e^{-\|\gamma\|^2} d\gamma = \int_{\mathcal{W}_S \gamma \geq 0} d\Omega \int_0^{+\infty} e^{-l^2} l^{d-1} dl = \frac{A\left(\mathcal{B}\left(\gamma, r\right) \cap S\right)}{r^{d-1}} \int_0^{+\infty} e^{-l^2} l^{d-1} dl. \tag{A.96}$$

According to Eqn. A.94-A.96, we have

$$\begin{aligned} A\left(\mathcal{B}\left(\gamma_0, r\right) \cap S\right) &= \frac{2 \int_{\mathcal{W}_S \gamma \geq 0} e^{-\|\gamma\|^2} d\gamma}{\Gamma\left(\frac{d}{2}\right)} r^{d-1} \\ \forall k, \frac{A\left(\mathcal{B}\left(\gamma_0, r\right) \cap S\right)}{A\left(\mathcal{B}\left(\gamma_0, r\right) \cap \mathcal{S}_0\right)} &\geq \frac{\int_{\mathcal{W}_S \gamma \geq 0} e^{-\|\gamma\|^2} d\gamma}{\pi^{\frac{d}{2}}}. \end{aligned} \tag{A.97}$$

Hence we only need prove the integral is lower bounded by a constant above zero. We will do this by extracting a sub-space from $\mathcal{W}_S \gamma \geq 0$, which is easy to be integrated. Specifically, we consider a sub space which is an affine transformation on $\mathcal{W}_S \gamma \geq 0$. We introduce the diameter of simplex S as $\mathrm{dia}(S) = \max_{s_1, s_2 \in S} \|s_1 - s_2\|$.

$$\begin{aligned}\int_{\mathcal{W}_S\gamma\geq 0} e^{-\|\gamma\|^2} d\gamma &\geq \int_{\gamma=t\gamma',\gamma'\in S} e^{-\|\gamma\|^2} d\gamma \quad (\forall t>0, t \text{ is arbitrarily chosen })\\ &= \int_{\gamma'\in S} t^{4d} e^{-t^2\|\gamma'\|^2} d\gamma' \\ &\overset{*}{\geq} t^* \operatorname{Vol}(S) e^{-t^2 dia(S)^2} \\ &\overset{**}{\geq} \eta t^{4d} h(S)^{4d} e^{-t^2 dia(S)^2}\end{aligned} \tag{A.98}$$

$*$ is due to $\forall \gamma' \in S$, we have $\|\gamma'\| \leq \operatorname{dia}(S)$, since γ_l is chosen as the origin. $**$ is due to Ass. 5. Additionally, we further show that in the above Formulation, $h(S)=\operatorname{dia}(S)$. Namely the longest edge of simplex always serves as the diameter. It is equal to prove (S^i denotes the supporting vector):

$$\begin{aligned}\operatorname{dia}(S) &= \max_{s_1,s_2\in S} \|s_1-s_2\| \quad \left(s_2=\sum_i \lambda_i S^i, \lambda_i\in[0,1]\right)\\ &= \max_{s_1,s_2\in S} \left\|\left(\sum_i \lambda_i\right) s_1 - \sum_i \left(\lambda_i S^i\right)\right\| \\ &= \max_{s_1,s_2\in S} \left\|\sum_i \lambda_i \left(s_1 - S^i\right)\right\| \\ &\leq \max_{s_1\in S}\max_i \left\|s_1-S^i\right\| \\ &\overset{*}{\leq} \max_{i,j} \left\|S^j-S^i\right\| \leq h(S).\end{aligned} \tag{A.99}$$

$*$ is due to we also do the expansion on s_1, namely $s_1=\sum_i \lambda_i' S^i, \lambda_i' \in [0,1]$. On the other hand, we have $h(S) \leq \max_{s_1,s_2\in S}\|s_1-s_2\| = \operatorname{dia}(S)$ by definition. Thus $\operatorname{dia}(S)=h(S)$. Hence

$$\forall t, \int_{\mathcal{W}_S\boldsymbol{\gamma}\geq 0} e^{-\|\gamma\|^2} d\boldsymbol{\gamma} \geq \eta(tdia(S))^d e^{-\left(t\operatorname{dia}(S)^2\right)}. \tag{A.100}$$

Due to the arbitrary of t, we have

$$\int_{\mathcal{W}_S\boldsymbol{\gamma}\geq 0} e^{-\|\gamma\|^2} d\boldsymbol{\gamma} \geq \eta \max_x x^d e^{-x^2} = \eta x^d e^{-x^2}\Big|_{x=\sqrt{\frac{d}{2}}} = \eta(\frac{d}{2})^{\frac{d}{2}} e^{-\frac{d}{2}}. \tag{A.101}$$

Finally, we have

$$\forall k, \frac{A\left(\mathcal{B}\left(\gamma,r\right)\cap S\right)}{A\left(\mathcal{B}\left(\gamma,r\right)\cap S_0\right)} \geq \eta\left(\frac{d}{2e\pi}\right)^{\frac{d}{2}}, \text{the intersection number } N \leq \frac{1}{\eta}\left(\frac{2e\pi}{d}\right)^{\frac{d}{2}} < +\infty. \tag{A.102}$$

We have finished the proof.

A.8 SIMULATIONS AND AUXILIARY EXPERIMENTS

Simulations The visualization of our PI-SFP's performance in simulations are presented in Figure 2 and Figure 3.

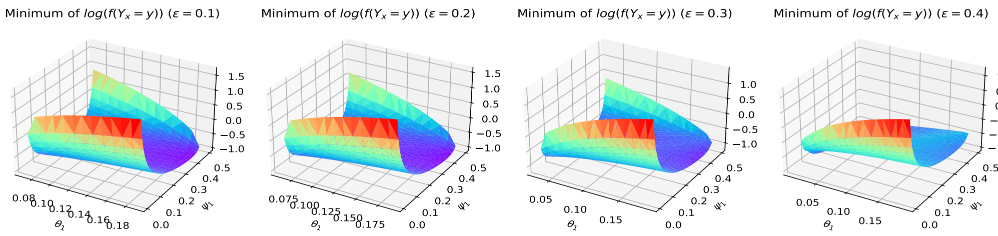


FIGURE 2: *We search the minimum of $f(Y_x = y)$ in the binary case, by conducting naive linear programming on each fixed $\theta_1(\theta_2)$ and $\psi_1(\psi_2)$ in Eqn* (3).

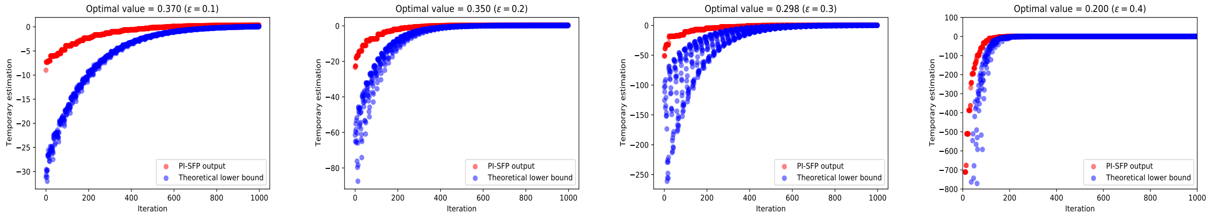


FIGURE 3: *Results of PI-SFP. PI-SFP (blue) converges to the optimal value of $\underline{f(Y_x = y)}$ with ε changing from* 0.1 *to* 0.4. *The red line denotes the theoretical convergence rate (ground truth).*

Real-world experiments We present the numerical result of real-world experiments as follows. Here semi-parametric COCA and Doubly-robust parametric COCA method is followed by Tchetgen et al. [2023], Park and Tchetgen [2023]; moreover, the standard difference-in-difference (DID) method is followed by Card and Krueger [1993], Angrist and Pischke [2009].

To facilitate fair comparison, we adopt the same experimental setting as in Tchetgen et al. [2023] and we refer readers to specific details in their experimental part. Specifically, in the dataset, we set 185 and 488 municipalities as samples from two areas Pernambuco (PE),Rio Grande do Sul (RS), as the treatment and control group, respectively. The covariate $\boldsymbol{U}$ contain three parts: (i) municipality-level population size, (ii) population density, and (iii) proportion of females measured in 2014. Moreover, we force "post-epidemic municipality-level birth rate" in 2016 as the outcome Y, and whether individuals are infected by the virus as treatment control X. Finally, we choose "preepidemic municipality-level birth rates in 2013 and 2014" as the outcome proxies W_1, W_2, respectively, which is so-called NCO in Table 3. Our transition matrix $P(\boldsymbol{W} \mid \boldsymbol{U})$ is also approximated from the observations in the public data.

In this process, we choose $P(\boldsymbol{X}, \boldsymbol{Y})$, $P(\boldsymbol{W})$ and partial observed transition matrix $P(\boldsymbol{W} \mid \boldsymbol{U})$ as observed data, and others as the protected feature. Our PI-SFP exhibits a narrower lower and upper bound compared with the previous literature in most cases.

Estimator	Statistic	NCO		
		W_1	W_2	(W_1, W_2)
Semi-parametric COCA	Estimate	-2.410	-2.182	-2.180
	SE	0.356	0.503	0.342
	95%CI	$(-3.107, -1.713)$	$(-3.168, -1.196)$	$(-2.850, -1.510)$
Doubly-robust parametric COCA	Estimate	-2.235	-1.833	-2.182
	SE	0.502	0.519	0.415
	95%CI	$(-3.220, -1.250)$	$(-2.850, -0.816)$	$(-2.996, -1.368)$
Standard DiD	Estimate	-1.156	-1.041	-1.041
	SE	0.199	0.195	0.195
	95% CI	$(-1.546, -0.767)$	$(-1.424, -0.658)$	$(-1.424, -0.658)$
PI-SFP (ours)	bound	$[-3.012, -1.201]$	$[-2.732, -1.232]$	$[-2.742, -1.203]$

TABLE 3: *Real-world experiment.*