On Measuring Causal Contributions via do-interventions

Yonghan Jung [1] **Shiva Prasad Kasiviswanathan** [2] **Jin Tian** [3]
Dominik Janzing [2] **Patrick Blöbaum** [2] **Elias Bareinboim** [4]

Abstract

Causal contributions measure the strengths of different causes to a target quantity. Understanding causal contributions is important in empirical sciences and data-driven disciplines since it allows to answer practical questions such as "what are the contributions of each cause to the effect?" In this paper, we develop a principled method for quantifying causal contributions. First, we provide desiderata in terms of axioms that causal contribution measures should satisfy and propose the *do-Shapley values* (inspired by do-interventions (Pearl, 2000)) as a method satisfying these properties. Next, we develop a criterion under which the do-Shapley values can be efficiently inferred from non-experimental data. Finally, we introduce do-Shapley estimators exhibiting consistency and statistical robustness. Simulation results corroborate with the proposed theory.

1. Introduction

Inferring causal effects is a fundamental problem throughout the data sciences since it can answer queries like "*what would be the expected outcome if inputs had been fixed to certain values?*". There is a growing literature that investigates the conditions under which causal conclusions can be drawn from observational and experimental data (causal effect identification) (Pearl, 1995; Tian & Pearl, 2003; Huang & Valtorta, 2006; Shpitser & Pearl, 2006; Bareinboim & Pearl, 2012; 2016; Jaber et al., 2018; Lee et al., 2019; 2020; Lee & Bareinboim, 2020), and in estimating the identified causal functions from data (causal effect estimation) (Jung et al., 2020; Bhattacharya et al., 2020; Jung et al., 2021a;b; Bhattacharyya et al., 2020; 2021; Xia et al., 2021). Beyond these tasks, interpreting the results of causal inference, including answering "*what is the most important cause of the effects?*", or more generally, "*what are the contributions of each cause to the effect?*" are also of practical importance. Answering these queries falls under the task of measuring *causal contributions*, which aims to quantify the degree of contribution of different causes to a target effect. As a motivational example, consider the following scenario described in (Lundberg, 2021):

Example 1. *A video streaming service company has collected data that contains various features including sales call* (S)*, product needs* (P)*, interaction with customers* (I)*, monthly usage* (M)*, discounts provided* (D)*, last upgrade* (L)*, economic factors* (E)*, ad spend* (A)*, and bugs reported* (B)*. These features are causally related and affect the outcome: customers retention* (Y) *(also, for further details, see Fig. 1). The company aims to measure the causal contributions of these features to the target effect – the expected customer retention if each feature had been fixed to a certain value (e.g., set to lower sales calls, higher product needs, etc.).*

Example 1 captures practical cases where the target quantity is related to the query '*what would be the output if inputs had been fixed to certain values?*' This includes cases where the target quantity is a machine learning (ML) model's output, which is derived by fixing inputs to specific values (see Remark 1).

In the area of explainable AI (XAI), there is a series of works concerned with measuring the contributions of features to an ML model output (Lundberg & Lee, 2017; Schwab & Karlen, 2019; Janzing et al., 2020b; Heskes et al., 2020; Covert et al., 2021). Most of these methods have focused on queries where the target quantity is induced from an *accessible* model – a model for target Y is said to be *accessible* if the model can be evaluated to obtain Y value for arbitrary input features – with less attention has been paid to settings where the target is induced by nature (i.e., the data-generating process is *inaccessible*, such as the customer retention in Example 1). Also, many existing techniques are based on the correlation between the features and the ML model output, including, (Lundberg & Lee, 2017; Frye et al., 2020) to cite a few. Another thread of this approach focused on measuring contributions based on causation (Schwab & Karlen, 2019; Janzing et al., 2020b; Heskes et al., 2020),

[1]Purdue University [2]Amazon [3]Iowa State Univerity [4]Columbia University. Correspondence to: Yonghan Jung <jung222@purdue.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

but often assumes that the data generating process for the target is known and accessible, allowing that an outcome corresponding to any arbitrary features can be generated. This rules out scenarios where the target quantity is induced from an inaccessible model (such as Example 1). A more detailed comparison with the existing literature is presented in Sec. 3.1.

In this paper, we generalize previous approaches to measure the causal contributions of each feature to a target effect induced by an inaccessible model and described by a joint interventional distribution. Our proposed method is applicable to the task of quantifying causal contributions of input features of an ML model prediction as well. More specifically, our contributions are as follows:

1. [Sec. 3] We *axiomatize* causal contribution measures. Specifically, we propose desiderata for causal contribution measures (a set of axioms), and introduce the *do*-Shapley, a Shapley value-based method (Shapley, 1953) specialized for quantifying the causal contributions described by *do*-interventions (Pearl, 2000).[1] Our axiomatic characterization provides a theoretical justification for using the *do*-Shapley for quantifying causal contributions.

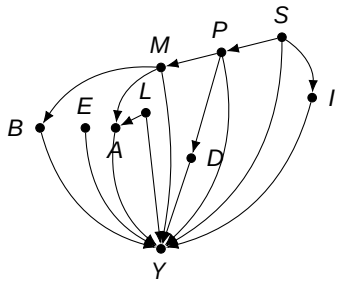
2. [Sec. 4] We provide conditions under which the *do*-Shapley values can be inferred from observational data (*identifiability*) in polynomial time. Even if verifying the identifiability can be done through existing causal-effect identification results, determining the identifiability of *do*-Shapley values is, in practice, not computationally feasible. In particular, we introduce sufficient conditions under which the identifiability of *do*-Shapley values is determined in polynomial time.

3. [Sec. 5] We develop estimators for the *do*-Shapley values, exhibiting consistency, identifiability, and statistical robustness. We developed three estimators based on the *inverse probability weighting* (IPW) (Rosenbaum & Rubin, 1983), *outcome regression* (REG) (Rubin, 1979), and *double/debiased machine learning* (DML) (Chernozhukov et al., 2018), respectively. We also show that the DML estimator displays statistical robustness to model misspecification and bias.

4. [Sec. 6]. Finally, we present simulation results on these estimators that corroborate with the theory.

Due to space constraints, the proofs and other omitted details are provided in the appendix.

[1]The *do*-Shapley is a generalization of the *causal Shapley* (Heskes et al., 2020), which also uses the *do*-interventions to the case where the target quantity is induced by an inaccessible model.



(a)

Figure 1: Causal graph for Example 1 (Lundberg, 2021).

2. Preliminaries

Notation. Each variable is represented with a capital letter (V) and its realized value with a small letter (v). We use bold letters $\mathbf{V}$ and $\mathbf{v}$ to denote a set of variables and their realized value, respectively. For any set S, we use $|S|$ to denote its cardinality. Given a topological order over the vertices $\mathbf{V} := \{V_1, \ldots, V_n\}$ of a graph G, we will use $\text{pre}(V_i)$ to denote the predecessors of V_i and use $\text{pre}(v_i)$ to denote a realization of $\text{pre}(V_i)$; i.e., $\text{pre}(v_i) = \mathbf{w}_i$ for $\text{pre}(V_i) = \mathbf{W}_i$. We use $\text{Ch}(V_i)$ to represent the children of a variable V_i in G. For an index set $[n] := \{1, \ldots, n\}$ and a subset $S \subseteq [n]$, we use $\mathbf{V}_S := \{V_k\}_{k \in S}$ and $\mathbf{V}_{\overline{S}} := \{V_k\}_{k \notin S}$. We use $\mathcal{D}$ to denote N samples from a distribution P over $\mathbf{V}$; i.e., $\mathcal{D} := \{\mathbf{V}_{(i)}\}_{i=1}^{N} \overset{\text{iid}}{\sim} P$, where $\mathbf{V}_{(i)}$ denotes the ith sample. For a function f, we use $\mathsf{E}[f(\mathbf{V})]$ as an expectation of $f(\mathbf{V})$ over P, and $\mathsf{E}_{\mathcal{D}}[f(\mathbf{V})] := (1/N)\sum_{i=1}^{N} f(\mathbf{V}_{(i)})$. We use $\|f(\mathbf{V})\| := \sqrt{\mathsf{E}[(f(\mathbf{V}))^2]}$ to denote the $L_2(P)$ norm of $f(\mathbf{V})$. $O_P(\,)$ and $o_P(\,)$ denotes the big O and little O in probability, respectively.

Structural Causal Models. We use the language of structural causal models (SCMs) as our basic semantical framework (Pearl, 2000). A structural causal model (SCM) is a tuple $\mathcal{M} := \langle \mathbf{V}, \mathbf{U}, \mathbf{F}, P(\mathbf{u}) \rangle$, where $\mathbf{V}, \mathbf{U}$ are sets of endogenous (observables) and exogenous variables (latents) respectively, $\mathbf{F}$ is a set of functions f_{V_i} one for each $V_i \in \mathbf{V}$ where $V_i \leftarrow f_{V_i}(PA_{V_i}, U_{V_i})$ for some $PA_{V_i} \subseteq \mathbf{V}$ and $U_{V_i} \subseteq \mathbf{U}$, and $P(\mathbf{u})$ is a strictly positive probability measure for $\mathbf{U}$. Each SCM $\mathcal{M}$ is associated to a causal diagram G over the node set $\mathbf{V}$ where $V_i \rightarrow V_j$ if V_i is an argument of f_{V_j}, and $V_i \dashleftarrow\dashrightarrow V_j$ if the corresponding U_{V_i} and U_{V_j} are not independent. Performing an intervention $\mathbf{X} = \mathbf{x}$ is represented through the *do*-operator, $do(\mathbf{X} = \mathbf{x})$ (shortly, $do(\mathbf{x})$), which encodes the operation of replacing the original equations of $\mathbf{X}$ by the constant $\mathbf{x}$ in the SCM $\mathcal{M}$, inducing a submodel $\mathcal{M}_{\mathbf{x}}$ and an interventional distribution $P(\mathbf{V} = \mathbf{v} | do(\mathbf{x}))$ (shortly, $P(\mathbf{v} | do(\mathbf{x}))$) (Bareinboim et al., 2020).

Causal Effect Identification. Given a causal graph G

over $\mathbf{V}$, an effect $P(\mathbf{y}|do(\mathbf{x}))$ where $\mathbf{X}, \mathbf{Y}$ $\mathbf{V}$ is *identifiable* if $P(\mathbf{y}|do(\mathbf{x}))$ is computable from the distribution $P(\mathbf{v})$ in any SCM M that induces G (Pearl, 2000, p. 77). One key notion for identification is *confounded components (in short, C-components)*: a set of nodes connected with a path composed solely of bi-directed edges $V_i \leftrightarrow V_j$ (Tian & Pearl, 2003). For any $\mathbf{C}$ $\mathbf{V}$, the quantity $Q[\mathbf{C}] := P(\mathbf{c}|do(\mathbf{v} \cap \mathbf{c}))$, called a *C-factor* (Tian & Pearl, 2003), is defined as an interventional distribution of $\mathbf{C}$ under an intervention on $\mathbf{V} \cap \mathbf{C}$. We use $C(V_i)_G$ (shortly, $C(V_i)$) to denote the C-component of V_i in G, the set of variables belonging to the same C-component as V_i. We use $C(\mathbf{W}) := \bigcup_{V_i \in \mathbf{W}} C(V_i)$ to denote the C-component of a set $\mathbf{W}$ $\mathbf{V}$.

Shapley Value. The Shapley value (Shapley, 1953) seeks to allocate the contribution of each player $i \in [n]$ on some function value $f([n])$ given a value function $\nu(S)$ that measures the value of coalition of players $i \in S$ $[n]$. The Shapley value is given by

$$\phi_i(\nu) := \sum_{S \subseteq [n] \setminus \{i\}} \omega(S) \{\nu(S \cup \{i\}) - \nu(S)\}, \quad (1)$$

where $\omega(S) := \frac{1}{n}\binom{n-1}{|S|}^{-1}$. The Shapley value is uniquely satisfying a set of some desiderata for fair allocation (Young, 1985) (See Appendix A for more details).

2.1. Problem Definition

We are given samples $\mathcal{D}$ drawn from a distribution $P := P_M$ and a compatible causal diagram $G := G_M$, induced by the SCM M, on topologically ordered variables $(\mathbf{V}, Y)$, with Y the final variable in the order. We assume that Y is bounded, and $\mathbf{V}$ is a set of discrete variables.

Given $(G, \mathcal{D}, \mathbf{v})$ where $\mathbf{v}$ is a realization of $\mathbf{V}$, the task is to measure the contribution of each $v_i \in \mathbf{v}$ to the target causal effect[2] $\mathsf{E}[Y|do(\mathbf{v})]$ based on their impact on Y if the SCM M has fixed the value of the variable as $V_i = v_i$. We set $\mathsf{E}[Y] = 0$ without loss of generality. We make no assumptions regarding the data generating process of Y for generality. With the following additional assumption on f, our problem is reduced to the problem of attributing the importance of features in an ML prediction model:

Remark 1 (Reduction to ML models). *If Y is generated by a deterministic function, e.g., Y is an output of an ML prediction model f s.t. $Y := f(\mathbf{V})$, then our task reduces to measuring the causal contribution of each features $v_i \in \mathbf{v}$ on the ML prediction $f(\mathbf{v})$, since $\mathsf{E}[Y|do(\mathbf{v})] = f(\mathbf{v})$, the deterministic function.*

[2]We focus on the average causal effect $\mathsf{E}[Y|do(\mathbf{v})]$, one of the most commonly used quantities in practice. Our method is applicable for any function of the causal distribution $P(y|do(\mathbf{v}))$. A condition whether the target quantity $\mathsf{E}[Y|do(\mathbf{v})]$ (or $P(y|do(\mathbf{v}))$) can be determined using non-experimental data is discussed in Sec. 4.

3. Axioms for Causal Contribution

We start by asking the question: "*What would be a good measure for causal contribution?*" To answer this question, we propose the following desiderata inspired by previous works (Young, 1985; Friedman & Moulin, 1999; Sundararajan et al., 2017; Sundararajan & Najmi, 2020):

Axiom 1 (Desiderata for Causal Contribution). Causal contributions $\{\phi_{v_i}\}_{i=1}^n$ is considered desirable if the following properties are satisfied:

1. **Perfect assignment:** The contributions are perfectly assigned; formally, $\mathsf{E}[Y|do(\mathbf{v})] = \sum_{v_i \in \mathbf{v}} \phi_{v_i}$.

2. **Causal irrelevance:** If V_i is causally irrelevant to Y for all witness $\mathbf{w}$ $\mathbf{v} \cap \{v_i\}$; formally, $\forall y, P(y|do(v_i, \mathbf{w})) = P(y|do(\mathbf{w}))$[3], then $\phi_{v_i} = 0$.

3. **Causal symmetry:** If $(v_i, v_j) \in \mathbf{v}$ have the same causal explanatory power[4] to Y, $\forall \mathbf{w}$ $\mathbf{v} \cap \{v_i, v_j\}$ ($\forall y$, $P(y|do(v_i, \mathbf{w})) = P(y|do(v_j, \mathbf{w}))$), then $\phi_{v_i} = \phi_{v_j}$.

4. **Causal approximation:** For any S $[n]$ and $\mathbf{v}_S := \{v_i\}_{i \in S}$, $\sum_{i \in S} \phi_{v_i}$ well approximates $\mathsf{E}[Y|do(\mathbf{v}_S)]$. Formally, $\{\phi_{v_i}\}_{i=1}^n$ is a solution to the following weighted least square; i.e., $\{\phi_{v_i}\}_{i=1}^n = \arg\min_{\{\phi_{v_i}\}_{i=1}^n} \sum_{S \subseteq [n]} (\mathsf{E}[Y|do(\mathbf{v}_S)] - \sum_{i \in S} \phi_{v_i}^{\emptyset})^2 \omega(S)$ for some positive and bounded function $\omega(S)$.

The rationale behind Axioms 1 is the following: **(1) Perfect assignment** is a natural requirement since we aim to attribute the degree of contributions of each feature $v_i \in \mathbf{v}$ to the target causal effect. **(2) Causal irrelevance** reflects a desire to understand the cause of the outcome by forcing zero contributions for variables not causing the outcome. **(3) Causal symmetry** enforces the equal contribution for a pair of features if they have the same causal explanatory power. **(4) Causal approximation** allows ϕ_{v_i} to be interpreted as a proxy for the causal effect s.t. $\sum_{i \in S} \phi_{v_i}$ $\mathsf{E}[Y|do(\mathbf{v}_S)]$ for any S $[n]$.

Perhaps surprisingly, there is a unique causal contribution measure $\{\phi_{v_i}\}$ satisfying the above four properties.

Definition 1 (*do*-**Shapley**). The *do*-Shapley[5] is a causal contribution measure $\{\phi_{v_i}\}_{i=1}^n$ of $\mathbf{v}$ on $\mathsf{E}[Y|do(\mathbf{v})]$ w.r.t. G

[3]V_i is *causally irrelevant* to Y given $\mathbf{V}_S$ if $P(y|do(v_i; \mathbf{v}_S)) = P(y|do(\mathbf{v}_S))$ (Galles & Pearl, 1997, Def. 7).

[4]A causal explanatory power of $\mathbf{X} = \mathbf{x}$ to $Y = y$ is a measure of making $Y = y$ 'more likely' if $\mathbf{X}$ had been fixed to $\mathbf{x}$; i.e., $P(y|do(\mathbf{x}))$ $P(y)$ (Eva & Stern, 2019).

[5]Heskes et al. (2020) proposed the same equation for measuring contributions in the accessible model setting and referred to as the *causal Shapley*. In this paper, we use the term *do-Shapley* to make it clearer that the definition is based on the *do*-intervention.

defined as:

$$\phi_{v_i} := \sum_{S \subseteq [n] \setminus \{i\}} \omega(S) \{ \mathsf{E}[Y | do(\mathbf{v}_{S;i})] - \mathsf{E}[Y | do(\mathbf{v}_S)] \}, \quad (2)$$

where $\omega(S) := (1/n) \binom{n-1}{|S|}^{-1}$.

Theorem 1 (Uniqueness of the *do*-Shapley). *The do-Shapley is a unique causal contribution measure satisfying all the properties in Axiom 1.*

Remark 2. *Thm. 1 is significant because Axiom 1 do not restrict the value function to any fixed form. Thm. 1 instead characterizes the do-Shapley as the unique causal contribution measure satisfying Axiom 1 among any arbitrary value functions and corresponding contribution measures, as in (Sundararajan & Najmi, 2020).*

The *do*-Shapley, as the name suggests, is a specialization of the Shapley value in Eq. (1) for $\nu(S) = \mathsf{E}[Y | do(\mathbf{v}_S)]$. The *do*-Shapley can be alternatively viewed as a marginal causal effect of $v_i \in \mathbf{v}$ (i.e., $\mathsf{E}[Y | do(\mathbf{v}_{S;i})] - \mathsf{E}[Y | do(\mathbf{v}_S)]$) weighted-averaging over a set S. The significance of Thm. 1 stems from that it codifies the guarantees of the *do*-Shapley, and provides a tool to compare and contrast with alternative contribution metrics.

Remark 3 (Attribution of contributions for a subset of variables). *It is worth noting that the do-Shapley allocates contributions to all $v_i \in \mathbf{v}$ from a joint interventional distribution. In practice, assigning contributions exclusively to a subset $\mathbf{x} \subseteq \mathbf{v}$ may lead to more interpretable results. For example, when $\mathbf{X} := Pa(Y) \setminus \mathbf{V}$, assigning contributions only to the features in $\mathbf{x} \subseteq \mathbf{v}$ might be more interpretable if it is needed that features indirectly affecting the outcome should be assigned zero contributions. Enforcing the do-Shapley to assign contributions only for the subset $\mathbf{x}$ can be simply done (without loss of generality) by the following procedure: (1) Derive a causal graph $G[\mathbf{X}]$ compatible with $P(\mathbf{x})$ by applying the projection of a graph[6]; and (2) Compute the do-Shapley w.r.t. $G[\mathbf{X}]$. See Appendix B for more details.*

3.1. Relation with Other Work

In this section, we compare the *do*-Shapley in Def. 1 with other known methods aiming to measure contributions of features on the outcome. Table 1 summarizes the comparison.

Conditional Shapley. The *conditional Shapley* ($\phi_{v_i}^{cond}$) is a specialization of the Shapley value with $\nu(S) = \mathsf{E}[Y | \mathbf{v}_S]$ (Lundberg & Lee, 2017; Frye et al., 2020). The conditional Shapley measures contributions based on associations rather than causation. In general, the conditional Shapley doesn't match with the *do*-Shapley; The causal irrelevance property doesn't hold in the conditional Shapley (see Example C.1).

[6] $G[\mathbf{X}]$ is constructed as follow: For any $V_i, V_j \in \mathbf{X}$, (1) add a directed edge $V_i \to V_j$ in $G[\mathbf{X}]$ if there exists a directed path from V_i to V_j in G such that every vertex on the path is not in $\mathbf{X}$; (2) add a bidirected edge $V_i \dashleftarrow\dashrightarrow V_j$ in $G[\mathbf{X}]$ if there exists a divergent path between V_i and V_j in G such that every vertex on the path is not in $\mathbf{X}$ (Tian & Pearl, 2003).

	Causality	*Inaccessibility*	*Axioms*
Conditional	✗	X	✗
Marginal	X	✗	X
Causal	X	✗	X
ICC	X	X	✗
***do*-Shapley**	X	X	X

Table 1: Summary of comparisons of the **conditional**, **marginal**, **causal** Shapley values, and the **ICC** with our method (***do*-Shapley**) w.r.t. consideration of *causality*, capability in handling outcomes induced by an *inaccessible model* (e.g., Example 1), and having justification from *axioms*.

Marginal Shapley. The *marginal Shapley* is another widely used contribution measure in the XAI in which the target variable is a model prediction $Y = f(\mathbf{V})$, where f is a deterministic (refer Remark 1) and *accessible* prediction model. The marginal Shapley is a specialization of the Shapley value with $\nu(S) = \mathsf{E}[f(\mathbf{v}_S, \mathbf{V}_{\overline{S}})]$ (Janzing et al., 2020b). The marginal Shapley is known to satisfy certain desiderata in attributing the feature importance (Sundararajan & Najmi, 2020). With access to the model f, and a particular graphical assumption that features are not causally affecting each other, the marginal Shapley matches with the *do*-Shapley (Janzing et al., 2019, Eq. (14)). In general settings where features are causally related as in Example 1, the marginal Shapley doesn't match with the *do*-Shapley.

Causal Shapley. The *causal Shapley* (Heskes et al., 2020) is most closely related to the *do*-Shapley. Specifically, (Heskes et al., 2020) proposed the same equation as the *do*-Shapley proposed here for measuring the contributions when the outputs are generated by the *accessible* models, and the graph is unknown (only a partial topological ordering of the graph is known). While the *do*-Shapley doesn't have a restriction that the output is induced by the accessible models and is defined specifically on causal with bidirected edges induced by SCM M (See Sec. 2) for which rich theories on causal effect identification and estimation are available.

Intrinsic Causal Contribution (ICC). Janzing et al. (2020b) proposed a new method called *Intrinsic Causal Contribution* (ICC) ($\phi_{v_i}^{icc}$) to measure the causal contribution under the setting where the causal graph is Markovian, and the structural functions are invertible in the sense that the noise values can be reconstructed from the observations. The ICC relies on so-called a structure-based intervention,

which intervenes to features while keeping a causal structure and a joint distribution unaffected, to measure the contribution of V_i on Y. By doing so, the ICC can measure the contribution of V_i on Y that is not via upstream variables. However, there is no axiomatic characterization of the ICC to the best of our knowledge. It is easy to show that ICC does not satisfy the causal symmetry property (see Example C.2).

Other Contribution Measures. Wang et al. (2021a) focused on measuring the relevance of paths in a causal graph to a target node, whereas Singal et al. (2021) provided a recursive approach to capture the flow of importance through the graph. The causal influence defined in Janzing et al. (2013) is based on an operation called 'deletion of edges' and measures the relevance of edges with respect to the joint distribution, but not the relevance of edges for a certain target node. Schamberg et al. (2020) describes a generalization of the information-theoretic approach of Janzing et al. (2013) which quantifies relevance of *paths or edges* for a target node, based on operations on edges. Under some particular graphical assumptions, e.g., *flat graphs*, (Singal et al., 2021, Def. 8), the path/edge-based Shapley values (Wang et al., 2021a; Singal et al., 2021) match with the *do*-Shapley. In general, however, the link between these lines of work is yet to be fully established.

4. Identification of the *do*-Shapley

In this section, we investigate the question of evaluating the *do*-Shapley values. To evaluate the *do*-Shapley, expressing $\mathsf{E}[Y|do(\mathbf{v}_S)]$ as a functional of an observational distribution P using G is essential because we are only given non-experimental dataset $\mathcal{D}$ drawn from the observational distribution P. For each $S \subseteq [n]$, complete causal effect identification algorithms $\mathsf{E}[Y|do(\mathbf{v}_S)]$ are already available (Tian & Pearl, 2003; Huang & Valtorta, 2006; Shpitser & Pearl, 2006). A major practical challenge still remains, however, in using them because determining the identifiability for all subsets $S \subseteq [n]$ takes *exponential* computation time. In this section, we address this challenge in determining the identifiability by presenting a graphical criterion where the identifiability can be determined in polynomial time, which makes this procedure feasible in practice. Formally,

Definition 2 (Identifiability & Feasibility). The *do*-Shapley values $\{\phi_{v_i}\}_{i=1}^n$ w.r.t. G are said to be *identifiable* if all elements in $\{\mathsf{E}[Y|do(\mathbf{v}_S)]\}_{S \subseteq [n]}$ are identifiable in the causal graph G. The identification of the *do*-Shapley values are said to be (computationally-) *feasible* if the identification can be done in $O(\text{poly}(n))$.

Since näively applying the existing causal effect identification algorithms to determine the identifiability of the *do*-Shapley values is not computationally feasible (requires $O(2^n)$ computations), we provide a simple sufficient graphical criterion under which determining the *do*-Shapley identifiability is feasible. We start with a definition (refer Sec. 2 for C-component, C-factor):

Definition 3 (C-partition). For a set of variables $\mathbf{X} \subseteq \mathbf{V}$, $\{\mathbf{X}_k\}_{k=1}^c$ is said to be the C-partition if $\mathbf{X} = \bigcup_{k=1}^c \mathbf{X}_k$ (where $\mathbf{X}_a \setminus \mathbf{X}_b = \emptyset$ for $a \neq b$) where $\forall k \in [c], \mathbf{X}_k$ is a set s.t. any two pairs $X_i, X_j \in \mathbf{X}_k$ are in the same C-component. in G.

Theorem 2 (Identifiability & Feasibility of *do*-Shapley). *The do-Shapley is identifiable if no variable in $V_i \in \{\mathbf{V}\}$ is connected to its child $\text{Ch}(V_i)$ by bidirected paths in G. Suppose Y is not connected by bidirected paths. In this case, for any $S \subseteq [n]$,*

$$\mathsf{E}[Y|do(\mathbf{v}_S)] = \sum\nolimits_{\mathbf{v}_{\overline{S}}} \mathsf{E}[Y|\mathbf{v}]\, Q[\mathbf{V} \cap \mathbf{V}_S](\mathbf{v}),$$

where $Q[\mathbf{V} \cap \mathbf{V}_S] := Q[\mathbf{V} \cap \mathbf{V}_S](\mathbf{v})$ is given as

$$Q[\mathbf{V} \cap \mathbf{V}_S] = \frac{P(\mathbf{v})}{Q[C(\mathbf{V}_S)]} \prod_{k=1}^{c} \sum_{\mathbf{s}_k} Q[C(\mathbf{S}_k)],$$

where $Q[C(\mathbf{V}_S)] = \prod_{V_a \in C(\mathbf{V}_S)} P(v_a|pre(v_a))$ is a C-factor of a C-component $\mathbf{V}_S$ ($C(\mathbf{V}_S)$); $\{\mathbf{S}_k\}_{k=1}^c$ is a C-partition of $\mathbf{V}_S$; and $Q[C(\mathbf{S}_k)] := \prod_{V_a \in C(\mathbf{S}_k)} P(v_a|pre(v_a))$ is a C-factor of a C-component $C(\mathbf{S}_k)$ for $\mathbf{S}_k$.

Fig. 2a is an example graph where a graph satisfies the conditions in Thm. 2. Specifically, for all computations $\mathbf{V}_S \subseteq \mathbf{V} := \{V_1, V_2, V_3\}$, the causal effects are identified through Thm. 2 as

$$\mathsf{E}[Y|do(\mathbf{v}_S)] = \begin{cases} \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}[Y|\mathbf{v}]\, P(v_2|v_1, v_3) P(\mathbf{v}_S), & \text{if } S \in \{1,3\}, \\ \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}[Y|\mathbf{v}]\, P(\mathbf{v}_{\overline{S}}), & \text{if } S \in \{\emptyset, 2, \{1,2\}, \{2,3\}\}, \\ \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}[Y|\mathbf{v}]\, P(\mathbf{v}_{\overline{S}}|\mathbf{v}_S), & \text{if } S \in \{\{1\}, \{1,3\}\}, \\ \mathsf{E}[Y|\mathbf{v}] & \text{if } S = \{1,2,3\}. \end{cases} \tag{3}$$

Thm. 2 suggests the feasibility of the *do*-Shapley values since the proposed graphical criteria (checking whether V_i and $\text{Ch}(V_i)$ are connected by bidirected paths) can be done in $O(n^3)$ by applying the breadth-first-search for each variable $V_i \in \mathbf{V}$.

To demonstrate the applicability of Theorem 2, we provide two special cases which are commonly considered in the literature:

1. **Markovian case**: All latent variables in the SCM is independent; i.e., G is given as a DAG (Janzing et al., 2013; 2019; Heskes et al., 2020; Basu, 2020; Wang et al., 2021b; Singal et al., 2021).

2. **Direct-cause case**: No pair of variables $(V_i, V_j) \in \mathbf{V}$ $(i \neq j)$ is connected by a directed path, no V_i are connected to Y via bidirected edges, and no directed edge from Y to V_i exists (i.e., only $V_i \to Y$ is allowed) (Janzing et al., 2020a;b).

For each of these cases, the identification result in Theorem 2 can be simplified as follows.

Corollary 1 (Identification – Markovian). *In the Markovian case,* $\mathsf{E}[Y|do(\mathbf{v}_S)]$ *is given as*

$$\mathsf{E}[Y|do(\mathbf{v}_S)] = \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}\left[Y|\mathbf{v}_S, \mathbf{v}_{\overline{S}}\right] \prod_{i \in S} P(v_i|pre(v_i)).$$

Corollary 2 (Identification – Direct-cause). *In the Direct-cause case,* $\mathsf{E}[Y|do(\mathbf{v}_S)]$ *is given as*

$$\mathsf{E}[Y|do(\mathbf{v}_S)] = \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}\left[Y|\mathbf{v}_S, \mathbf{v}_{\overline{S}}\right] P(\mathbf{v}_{\overline{S}}).$$

Figs. (2b,2c) provide example graphs for Markovian and Direct-cause cases. For Fig. 2b, Coro. 1 gives

$$\mathsf{E}\left[Y|do(\mathbf{v}_S)\right] = \begin{cases} \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}\left[Y|\mathbf{v}\right] P(\mathbf{v}_{\overline{S}}|\mathbf{v}_S), & \text{if } S \in \{1,3\}, \{1,3\}\}, \\ \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}\left[Y|\mathbf{v}\right] P(\mathbf{v}_{\overline{S}}), & \text{if } S \in \{\emptyset, 2\}, \{2,3\}, \{1,2\}\}, \\ \mathsf{E}\left[Y|\mathbf{v}\right] & \text{if } S = \{1,2,3\}. \end{cases} \tag{4}$$

For Fig. 2c, Coro. 2 gives

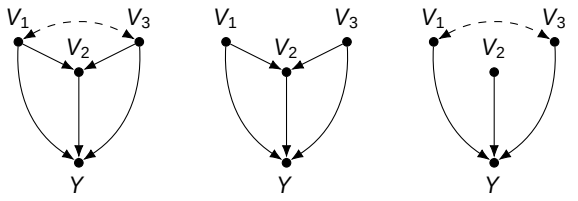
$$\mathsf{E}\left[Y|do(\mathbf{v}_S)\right] = \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}\left[Y|\mathbf{v}\right] P(\mathbf{v}_{\overline{S}}) \tag{5}$$

for all $\mathbf{v}_S$ $\{v_1, v_2, v_3\}$.

5. Estimation of the *do*-Shapley

Estimating the *do*-Shapley values in Eq. (2) is computationally and statistically challenging because (1) Iterating over all S $[n]$ takes time exponential in n, and (2) Estimating $\mathsf{E}[Y|do(\mathbf{v}_S)]$ might be vulnerable to bias due to finiteness of the sample dataset. In this section, we design computationally efficient and statistically robust estimators for the *do*-Shapley values to overcome these challenges, using three different techniques. For ease of presentation, we focus only on the Markovian & Direct-cause cases discussed in Sec. 4.

We first introduce estimators leveraging the idea of the inverse probability weighting (IPW) (Rosenbaum & Rubin, 1983). Our construction of the IPW estimator is based on the following result.



(a) Thm. 2 (b) Markovian (c) Direct-cause

Figure 2: Example graphs for Thm. 2 and two special cases: Markovian and Direct-cause.

Lemma 1 (Representation using IPW). *Let* $S = \{m_1, \ldots, m_s\}$ $[n]$ *denote an index set for* $\mathbf{V}_S$. *Let*

$$\omega_k^S := \prod_{r=1}^{k} \mathbb{1}_{v_{m_r}}(V_{m_r})/h_r^S, \text{ for } k = s, \ldots, 1;$$
$$\omega^S := \mathbb{1}_{\mathbf{v}_S}(\mathbf{V}_S)/h^S,$$

where $h_r^S := P(V_{m_r}|pre(V_{m_r}))$ *and* $h^S := P(\mathbf{V}_S|\mathbf{V}_{\overline{S}})$. *Then,* $\mathsf{E}[Y|do(\mathbf{v}_S)] = \mathsf{E}[Y\omega]$ *where* $\omega = \omega_k^S$ *for the Markovian case, and* $\omega = \omega^S$ *for the Direct-cause case.*

Using Lemma 1, we construct the IPW estimators.

Definition 4 (IPW for $\mathsf{E}[Y|do(\mathbf{v}_S)]$). The IPW estimator $T^{\text{ipw}}(S)$ for $\mathsf{E}[Y|do(\mathbf{v}_S)]$ is constructed as:

1. Split D randomly into two halves: D_0 and D_1;
2. Let $\widehat{\omega}_{s;p}^S, \widehat{\omega}_p^S$ denote estimators for ω_s^S, ω^S from $D_p \in \{D_0, D_1\}$, respectively.
3. For each $p \in \{0, 1\}$, set

$$T_p^{\text{ipw}}(S) := \begin{cases} \mathsf{E}_{D_{1-p}}\left[Y\widehat{\omega}_{s;p}^S\right] & \text{(Markovian)} \\ \mathsf{E}_{D_{1-p}}\left[Y\widehat{\omega}_p^S\right] & \text{(Direct-cause)} \end{cases}$$

4. $T^{\text{ipw}}(S) := \{T_0^{\text{ipw}}(S) + T_1^{\text{ipw}}(S)\}/2$.

The data-splitting (also known as sample-splitting) technique (Klaassen, 1987; Robins & Ritov, 1997; Robins et al., 2008; Zheng & van der Laan, 2011; Chernozhukov et al., 2018) will be employed in constructing all *do*-Shapley estimators discussed in this section. Without data-splitting, some restriction on the complexity of the estimator function class must be imposed to guarantee statistical consistency.

We introduce estimators leveraging the idea of outcome regression (REG) (Rubin, 1979). Our REG estimator is based on the following result.

Lemma 2 (Representation using REG). *Let* $S := \{m_1, \ldots, m_s\}$ $[n]$ *denote an index set for* $\mathbf{V}_S$. *Let*

$\theta^S_{s;1} := Y$. *For* $k = s, s \quad 1, \quad , 1,$

$$\theta^S_{k;2} := \mathsf{E}\left[\theta^S_{k;1} j V_{m_k}, pre(V_{m_k})\right]$$
$$\theta^S_{k\ 1;1} := \mathsf{E}\left[\theta^S_{k;1} j v_{m_k}, pre(V_{m_k})\right],$$
$$\theta^S_a := \mathsf{E}\left[Y j \mathbf{v}_S, \mathbf{V}_{\overline{S}}\right],$$
$$\theta^S_b := \mathsf{E}\left[Y j \mathbf{V}_S, \mathbf{V}_{\overline{S}}\right].$$

Then, $\mathsf{E}[Y j do(\mathbf{v}_S)] = \mathsf{E}[\theta]$ *where* $\theta = \theta^S_{0;1}$ *for the Markovian case, and* $\theta = \theta^S_a$ *for the Direct-cause case.*

We construct the REG estimator based on Lemma 2.

Definition 5 (REG for $\mathsf{E}[Y j do(\mathbf{v}_S)]$). The REG estimator $T^{\mathsf{reg}}(S)$ for $\mathsf{E}[Y j do(\mathbf{v}_S)]$ is constructed as:

1. Split D randomly into two halves: D_0 and D_1.
2. Let $\hat{\theta}^S_{k;2;p}, \hat{\theta}^S_{k\ 1;1;p}, \hat{\theta}^S_{a;p}$ denote an estimator for $\theta^S_{k;2}, \theta^S_{k\ 1;1}, \theta^S_a$ from $D_p \ 2 \ f D_0, D_1 g$, respectively.
3. For each $p \ 2 \ f0, 1g$,

$$T^{\mathsf{reg}}_p(S) := \begin{cases} \mathsf{E}_{D_1\ p}\left[\hat{\theta}^S_{0;1;p}\right] & \text{(Markovian)} \\ \mathsf{E}_{D_1\ p}\left[\hat{\theta}^S_{a;p}\right] & \text{(Direct-cause).} \end{cases}$$

4. $T^{\mathsf{reg}}(S) := f T^{\mathsf{reg}}_0(S) + T^{\mathsf{reg}}_1(S) g/2.$

For IPW and REG estimators to be consistent, one needs to estimate each individual functional (called *nuisances*) including $\mathsf{E}[Y j \mathbf{v}_S, \mathbf{v}_{\overline{S}}]$ or $P(v_i j \text{pre}(v_i))$ consistently. A desirable robust estimator is one that converges to the ground-truth at a fast rate even when estimates for nuisances are misspecified (i.e., wrongly specified) or converging relatively slowly. *Double/Debiased Machine Learning (DML)* (Chernozhukov et al., 2017) is a recently introduced technique to construct such estimators.

Lemma 3 (Representation using DML). *Let*

$$\eta^S := \begin{cases} f\theta^S_{0;1} g \ [\ f\theta^S_{k;1}, \theta^S_{k;2} g^s_{k=1} \ [\ f h^S_r g^s_{r=1} & \textit{(Markovian)} \\ f\theta^S_a, \theta^S_b, h^S g & \textit{(Direct-cause)}, \end{cases}$$

defined in Defs. (4, 5) above, and

$$V_S(\mathbf{V}^0; \eta^S) := \begin{cases} \theta^S_{0;1} + \sum_{k=1}^{s} \omega^S_k (\theta^S_{k;1} \quad \theta^S_{k;2}) & \textit{(Markovian)} \\ \theta^S_a + \omega^S \left(Y \quad \theta^S_b\right) & \textit{(Direct-cause)}, \end{cases}$$

where $\omega^S_k := \prod_{r=1}^{k} \mathbb{1}_{v_{m_r}}(V_{m_r})/h^S_r$ *and* $\omega^S := \mathbb{1}_{\mathbf{v}_S}(\mathbf{V}_S)/h^S$. *Then,* $\mathsf{E}[Y j do(\mathbf{v}_S)] = \mathsf{E}\left[V_S(\mathbf{V}^0; \eta^S)\right]$.

We construct the DML estimators based on Lemma 3:

Definition 6 (DML for $\mathsf{E}[Y j do(\mathbf{v}_S)]$). The DML estimator $T^{\mathsf{dml}}(S)$ is constructed as:

1. Split D randomly into two halves: D_0 and D_1;
2. Construct $\widehat{\eta^S_p}$, estimates of η^S from D_p, $p \ 2 \ f0, 1g$.
3. $T^{\mathsf{dml}}_p(S) := \mathsf{E}_{D_1\ p}\left[V_S(\mathbf{V}; \widehat{\eta^S_p})\right]$ for $p \ 2 \ f0, 1g$.
4. $T^{\mathsf{dml}}(S) := f T^{\mathsf{dml}}_0(S) + T^{\mathsf{dml}}_1(S))/2.$

Algorithm 1 `do-Shapley`$(M, T^{\mathsf{est}}(\ ))$

Input: M, Estimators $T^{\mathsf{est}}(\ )$ in Defs. (4,5,6).
Output: Estimates $f\widehat{\phi}_{v_i} g^n_{i=1}$.
Initialize $\widehat{\phi}_{v_i} = 0$ for all $V_i \ 2 \ \mathbf{V}$.
for $j = 1$ **to** M **do**
  Generate the random permutation π over $[n]$.
  for $i = 1$ **to** n **do**
    $\widehat{\phi}_{v_i} \quad \widehat{\phi}_{v_i} + T^{est}(f i, \text{pre}\ \ (i) g) \quad T^{est}(\text{pre}\ \ (i))$
  end for
end for
return $f\widehat{\phi}_{v_i}/M g^n_{i=1}$

Based on estimators in Defs. (4,5,6), we now propose a computationally efficient estimator for the *do*-Shapley values based on random permutations:

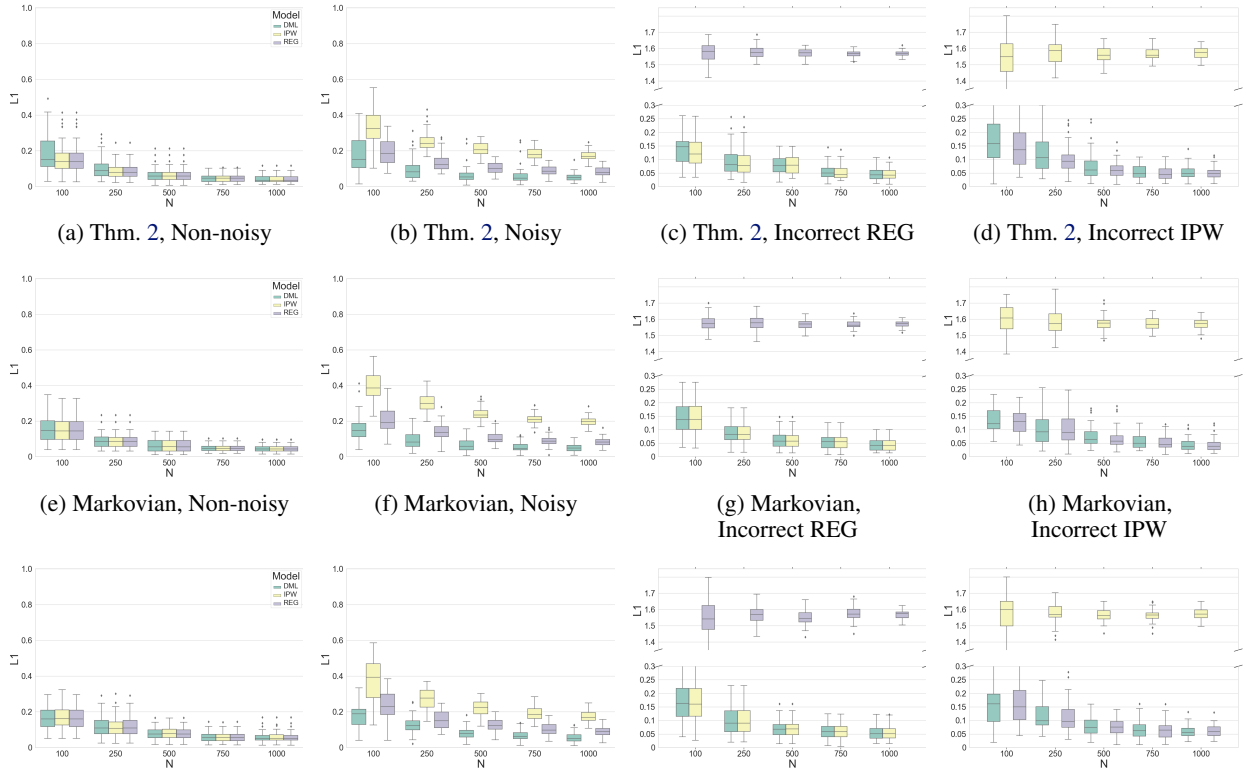
Definition 7 (*do*-Shapley estimators – Two cases). Let $T^{\mathsf{est}}(S) \ 2 \ f T^{\mathsf{ipw}}(S), T^{\mathsf{reg}}(S), T^{\mathsf{dml}}(S) g$ denote an estimator for $\mathsf{E}[Y j do(\mathbf{v}_S)]$ defined in Defs. (4,5,6), respectively. The *do*-Shapley estimator is given as

$$\phi^{\mathsf{est}}_{v_i} := \frac{1}{M} \sum_{j=1}^{M} (T^{\mathsf{est}}(f i, \text{pre}\ _j(i) g) \quad T^{\mathsf{est}}(\text{pre}\ _j(i))),$$

where M is the number of randomly generated permutation of $[n]$, π_j denotes the jth permutation, and pre $_j(i)$ is the set of elements that precedes i in π_j.

A systematic procedure for constructing *do*-Shapley estimators is provided in Algorithm 1. The following theorem summarizes the error analyses of all the three *do*-Shapley estimators.

Theorem 3 (Bias Analysis). *Let* $f\pi_j g^M_{j=1}$ *denote* M *randomly generated permutations of* $[n]$. *For the fixed index* i, *let* $S_{j;0} := pre\ _j(i)$ *and* $S_{j;1} := f i g \ [\ S_{j;a}$. *Let* $f\widehat{\eta}^{S_{j,0}}, \widehat{\eta}^{S_{j,1}} g^M_{j=1}$ *denote* L_2*-consistent estimates for all nuisances* $f\eta^{S_{j,0}}, \eta^{S_{j,1}} g^M_{j=1}$ *defined in Def. 6. Let* $R_{M;N} := O_P(M^{\ 1=2} + N^{\ 1=2})$. *Let* $e(\hat{g}) := k\hat{g} \quad g k$ *denote an error for a nuisance estimates for any* $\hat{g} \ 2 \ \hat{\eta}$ *and* $g \ 2 \ \eta$. *For the do-Shapley estimators defined in Def. 7, suppose the estimators* $T^{\mathsf{est}}(S)$ *are bounded. Let* $\epsilon^{\mathsf{est}}_{v_i} := \phi^{\mathsf{est}}_{v_i} \quad \phi_{v_i}$ (*where* est $\ 2 \ f$ipw, reg, dml$g$).



(a) Thm. 2, Non-noisy (b) Thm. 2, Noisy (c) Thm. 2, Incorrect REG (d) Thm. 2, Incorrect IPW

(e) Markovian, Non-noisy (f) Markovian, Noisy (g) Markovian, Incorrect REG (h) Markovian, Incorrect IPW

(i) Direct, Non-noisy (j) Direct, Noisy (k) Direct, Incorrect REG (l) Direct, Incorrect IPW

Figure 3: The L1-error plots. Plots are rendered in high resolution and can be zoomed in.

Under the Markovian case,

$$\epsilon^{\text{ipw}}_{V_i} = R_{M;N} + O_P\Big(\sum_{p\in\{0,1\}}\sum_{j=1}^{M} e(\hat{\omega}^{S_{j,p}}_{S_j})\Big),$$

$$\epsilon^{\text{reg}}_{V_i} = R_{M;N} + O_P\Big(\sum_{p\in\{0,1\}}\sum_{j=1}^{M} e(\hat{\theta}^{S_{j,p}}_{0;1})\Big),$$

$$\epsilon^{\text{dml}}_{V_i} = R_{M;N} + O_P\Big(\sum_{p\in\{0,1\}}\sum_{j=1}^{M}\sum_{k=1}^{s_j} e(\hat{h}^{S_{j,p}}_{k})e(\hat{\theta}^{S_{j,p}}_{k;2})\Big).$$

Under the Direct-cause case,

$$\epsilon^{\text{ipw}}_{V_i} = R_{M;N} + O_P\Big(\sum_{p\in\{0,1\}}\sum_{j=1}^{M} e(\hat{\omega}^{S_{j,p}})\Big),$$

$$\epsilon^{\text{reg}}_{V_i} = R_{M;N} + O_P\Big(\sum_{p\in\{0,1\}}\sum_{j=1}^{M} e(\hat{\theta}^{S_{j,p}}_{2})\Big),$$

$$\epsilon^{\text{dml}}_{V_i} = R_{M;N} + O_P\Big(\sum_{p\in\{0,1\}}\sum_{j=1}^{M} e(\hat{h}^{S_{j,p}})e(\hat{\theta}^{S_{j,p}}_{b})\Big).$$

Remark 4 (Properties of the Proposed Estimators). *Error analyses in Thm. 3 exhibit consistency of IPW, REG, DML estimators. Specifically, if nuisances are consistently estimated,* $\epsilon^{\text{ipw}}_{V_i} = \epsilon^{\text{reg}}_{V_i} = \epsilon^{\text{dml}}_{V_i} = O_P(1)$*, indicating that the estimators converge to the true quantity. Furthermore, the result presents the statistical robustness property of the DML. In particular, the DML estimates* $\hat{\phi}^{\text{dml}}_{V_i}$ *converges to the true value if either* $e(\hat{h}^{S_{j,p}}_{k})$ *or* $e(\hat{\theta}^{S_{j,p}}_{k;2})$ *under Markovian, and either* $e(\hat{h}^{S_{j,p}})$ *or* $e(\hat{\theta}^{S_{j,p}}_{b})$ *under Direct-cause are accurate (doubly robustness). Also,* $\hat{\phi}^{\text{dml}}_{V_i}$ *converges at the root-*N *rate if all nuisances* $\hat{h}^{S_{j,p}}_{k}, \hat{\theta}^{S_{j,p}}_{k;2}$ *under Markovian, and all nuisances* $\hat{h}^{S_{j,p}}, \hat{\theta}^{S_{j,p}}_{b}$ *under Direct-cause converge at least at* $N^{1=4}$ *rate (debiasedness).*

6. Experiments

In this section, we empirically compare the performance of the proposed *do*-Shapley estimators from the previous section. Details of the experiments and a different simulation example are provided in Appendices E and F.

Experimental Setup. We use synthetic datasets based on Figs. (2a, 2b, 2c) where each figures matches with Thm. 2, Markovian, and Direct-cause cases. We note that causal effects are identified as in Eqs. (3, 4, 5), respectively. Even if no known estimators for Thm. 2 exist generally, we note

that Eq. (3) is in an amenable form for which results in Sec. 5 are applicable. Throughout the simulation, we denote $f\phi_{V_i}g_{i=1}^{n}$ as the ground-truth *do*-Shapley values.

Comparison Between Estimators. We compare the three estimators (IPW, REG, DML), denoted by $f\phi_{V_i}^{\text{ipw}}, \phi_{V_i}^{\text{reg}}, \phi_{V_i}^{\text{dml}}g$ respectively, for scenarios depicted in graphs in Figs. (2a, 2b, 2c). Nuisances are estimated using gradient boosting model (Friedman, 2001).

Let $\phi_{V_{i,k}}^{\text{est}}$ 2 $f\phi_{V_{i,k}}^{\text{dml}}, \phi_{V_{i,k}}^{\text{ipw}}, \phi_{V_{i,k}}^{\text{reg}}g$ denote an estimated importance of the ith feature of jth samples (i.e., $V_{i;k}$ 2 $\mathbf{V}_{(k)}$ 2 D). We assess the quality of the estimator by computing the L_1 error as $L_1(\text{est}, k) := (1/n)\sum_{i=1}^{n}\left|\phi_{V_i;k}^{\text{est}} \quad \phi_{V_i;k}\right|$ (where n is the number of features). We ran the simulation for 50 randomly generated sets of samples; i.e., k 2 $f1, 2, \quad , 50g$, and with sample size $N := jDj$ 2 $f100, 250, 500, 750, 1000g$ to observe convergence behaviors of estimators. We fix $M = 20$. We refer the box-plot for $L_1(\text{est}, k)$ as the 'L1-error plot'.

For all fThm. 2, Markovian, Direct-cause$g$ cases, we compare the performances of the three *do*-Shapley estimators for (1) 'Non-noisy' where no noises are introduced in the model; (2) 'Noisy' where a 'converging noise' ϵ, decaying at a N rate (i.e., ϵ Normal(N $, N^{2}$)) for $\alpha = 1/4$, is added to the estimated nuisance to control the convergence rate, following the technique in (Kennedy, 2020); (3) 'Incorrect REG' where the model for the REG estimator in Def. 5 is wrongly specified; and (4) 'Incorrect IPW' the model for the IPW estimator in Def. 4 is wrongly specified.

Experimental Results. The L1-error plots for all cases are presented in Fig. 3. For the non-noisy setting, performances of all three models {DML, REG, IPW} are similar. In the noisy setting where the estimated nuisances are controlled to converge at $N^{1=4}$ rate, the DML estimators outperform the other two estimators by achieving a fast convergence with the smallest variance. This result corroborates the robustness property of the DML (Remark 4). Also, the DML estimator exhibits the doubly robustness property; the estimator converges in both the 'Incorrect IPW' and 'Incorrect REG' settings where each corresponding nuisance is wrongly specified.

Contrasting with Conditional Shapley. We contrast the *do*-Shapley and conditional Shapley in the non-noisy setting. We compare the importance ranking measured by the true *do*-Shapley with the ranks from the *do*-DML and conditional Shapley through the Spearman's rank correlation. The correlation is close to 1 if two ranks are similar and -1 if the ranks are opposite. The true data generating function is $Y = 3V_1 + 0.4V_2 + V_3 + U_Y$ and the true-*do*-Shapley identifies V_1 having the largest coefficient as the most important. As shown in Table 2, the *do*-DML-Shapley ranks the feature importance closer to the true rank. As noted, *do*-DML-Shapley identifies V_1 as the most important.

	Thm. 2	Markovian	Direct
DML	1.0	0.8	0.93
Conditional	-0.28	-0.74	0.52

Table 2: Comparison of the rank correlation.

7. Conclusion

We proposed the *do*-Shapley for measuring causal contribution and provided a theoretical justification through the axiomatic characterization (Thm. 1). Next, we developed conditions under which *do*-Shapley values can be inferred from non-experimental data in polynomial time (Thm. 2). We then proposed three *do*-Shapley estimators (IPW, REG, DML) that are consistent. We showed that the DML estimator has additional robustness property called doubly robustness and debiasedness (Thm. 3). We expect the proposed contribution measure will help empirical scientists to answer "*what are the contributions of each cause to the effect?*"

Acknowledgements

We thank the reviewers for their feedback and help in improving this manuscript. This work was done in part while Jin Tian was visiting the Simons Institute for the Theory of Computing. Elias Bareinboim and Yonghan Jung were partially supported by funding from the NSF, ONR, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

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

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

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 507–556. 2020.

Basu, D. On shapley credit allocation for interpretability. *arXiv preprint arXiv:2012.05506*, 2020.

Bhattacharya, R., Nabi, R., and Shpitser, I. Semiparametric inference for causal effects in graphical models with hidden variables. *arXiv preprint arXiv:2003.12659*, 2020.

Bhattacharyya, A., Gayen, S., Kandasamy, S., Maran, A., and Variyam, V. N. Learning and sampling of atomic interventions from observations. In *International Conference on Machine Learning*, pp. 842–853. PMLR, 2020.

Bhattacharyya, A., Gayen, S., Kandasamy, S., Raval, V., and Vinodchandran, N. Efficient inference of interventional distributions. *arXiv preprint arXiv:2107.11712*, 2021.

Charnes, A., Golany, B., Keane, M., and Rousseau, J. Extremal principle solutions of games in characteristic function form: Core, chebychev and shapley value generalizations. 1988.

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

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65, 2017.

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

Covert, I., Lundberg, S., and Lee, S.-I. Explaining by removing: A unified framework for model explanation. *Journal of Machine Learning Research*, 22(209):1–90, 2021.

Eva, B. and Stern, R. Causal explanatory power. *The British Journal for the Philosophy of Science*, 70(4):1029–1050, 2019.

Friedman, E. and Moulin, H. Three methods to share joint costs or surplus. *Journal of economic Theory*, 87(2): 275–312, 1999.

Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pp. 1189–1232, 2001.

Frye, C., Rowat, C., and Feige, I. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33, 2020.

Galles, D. and Pearl, J. Axioms of causal relevance. *Artificial Intelligence*, 97(1-2):9–43, 1997.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *Advances in Neural Information Processing Systems*, 33, 2020.

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

Jaber, A., Zhang, J., and Bareinboim, E. Causal identification under markov equivalence. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence*, 2018.

Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Schölkopf, B. Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358, 2013.

Janzing, D., Budhathoki, K., Minorics, L., and Blöbaum, P. Causal structure based root cause analysis of outliers. *arXiv preprint arXiv:1912.02724*, 2019.

Janzing, D., Blöbaum, P., Minorics, L., and Faller, P. Quantifying causal contributions via structure preserving interventions. *arXiv preprint arXiv:2007.00714*, 2020a.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916. PMLR, 2020b.

Jung, Y., Tian, J., and Bareinboim, E. Learning causal effects via weighted empirical risk minimization. *Advances in Neural Information Processing Systems*, 33, 2020.

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

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

Kennedy, E. H. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497*, 2020.

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

Klaassen, C. A. Consistent estimation of the influence function of locally asymptotically linear estimators. *The Annals of Statistics*, pp. 1548–1562, 1987.

Koster, J. T. et al. Marginalizing and conditioning in graphical models. *Bernoulli*, 8(6):817–840, 2002.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lee, S. and Bareinboim, E. Causal effect identifiability under partial-observability. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

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

Lee, S., Correa, J., and Bareinboim, E. Generalized transportability: Synthesis of experiments from heterogeneous domains. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2020.

Lundberg, S. Be careful when interpreting predictive models in search of causal insights. 2021. URL https://bit.ly/3gcZmgl.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777, 2017.

Molnar, C. *Interpretable machine learning*. Lulu. com, 2020.

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

Pearl, J. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, NY, 2000. 2nd edition, 2009.

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

Robins, J., Li, L., Tchetgen, E., van der Vaart, A., et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, pp. 335–421. Institute of Mathematical Statistics, 2008.

Robins, J. M. and Ritov, Y. Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine*, 16(3):285–319, 1997.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rubin, D. B. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366a):318–328, 1979.

Schamberg, G., Chapman, W., Xie, S.-P., and Coleman, T. P. Direct and indirect effects—an information theoretic perspective. *Entropy*, 22(8):854, 2020.

Schwab, P. and Karlen, W. Cxplain: Causal explanations for model interpretation under uncertainty. *Advances in Neural Information Processing Systems 32*, pp. 10220–10230, 2019.

Shapley, L. Annals of mathematics study no. 28. 1953.

Shapley, L. S. and Shubik, M. A method for evaluating the distribution of power in a committee system. *American political science review*, 48(3):787–792, 1954.

Shpitser, I. and Pearl, J. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the 21st AAAI Conference on Artificial Intelligence*, pp. 1219, 2006.

Singal, R., Michailidis, G., and Ng, H. Flow-based attribution in graphical models: A recursive shapley approach. *Available at SSRN 3845526*, 2021.

Štrumbelj, E. and Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International Conference on Machine Learning*, pp. 9269–9278. PMLR, 2020.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

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

Wang, J., Wiens, J., and Lundberg, S. Shapley flow: A graph-based approach to interpreting model predictions. In Banerjee, A. and Fukumizu, K. (eds.), *The 24th International Conference on Artificial Intelligence and Statistics,*, volume 130 of *Proceedings of Machine Learning Research*, pp. 721–729. PMLR, 2021a.

Wang, L., Zhang, Y., Richardson, T. S., and Robins, J. M. Estimation of local treatment effects under the binary instrumental variable model. *Biometrika*, 2021b.

Winter, E. The shapley value. *Handbook of game theory with economic applications*, 3:2025–2054, 2002.

Xia, K., Lee, K.-Z., Bengio, Y., and Bareinboim, E. The causal-neural connection: Expressiveness, learnability, and inference. *Advances in Neural Information Processing Systems*, 34:10823–10836, 2021.

Young, H. P. Monotonic solutions of cooperative games. *International Journal of Game Theory*, 14(2):65–72, 1985.

Zheng, W. and van der Laan, M. J. Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning*, pp. 459–474. Springer, 2011.

Appendix – On Measuring Causal Contributions via do-interventions

A. Fundamentals of the Shapley Value

The Shapley value (Shapley, 1953) in Eq. (1) seeks to allocate the contribution of each $i \in [n]$ on some function value $f([n])$ given a coalition function $\nu(S)$ that measures the value of coalition of values of players $i \in S$ (where $\nu([n]) = f([n])$). The Shapley value uniquely satisfying the following desiderata:

Theorem A.1 (**Axiomatization of the Shapley Value** (Shapley, 1953; Shapley & Shubik, 1954; Young, 1985))**.** *For any subset S of the players indexed $[n] = \{1, 2, \ldots, n\}$ and the value function of S, denoted $\nu(S)$, the Shapley value of the player i, denoted $\phi_i = \phi_i(\nu)$, equals*

$$\phi_i(\nu) := \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \binom{n-1}{|S|}^{-1} \{\nu(S \cup \{i\}) - \nu(S)\}, \tag{A.1}$$

is the unique attribution methods satisfying the following axioms (properties):

1. ***Efficiency**: $\sum_{i \in [n]} \phi_i = \nu([n])$;*
2. ***Dummy**: For some $i \in [n]$, if $\nu(S \cup \{i\}) = \nu(S)$ for all $S \subseteq [n] \setminus \{i\}$, then $\phi_i = 0$;*
3. ***Symmetry**: For some distinct $(i, j) \in [n]$, if $\nu(S \cup \{i\}) = \nu(S \cup \{j\})$ for all $S \subseteq [n] \setminus \{i, j\}$, then $\phi_i = \phi_j$;*
4. ***Linearity**: For all $i \in [n]$, for any two coalition functions ν_1 and ν_2, $\phi_i(\nu_1 + \nu_2) = \phi_i(\nu_1) + \phi_i(\nu_2)$.*

B. Details on Remark 3

Given a semi-Markovian causal graph G, a realized vector $(\mathbf{v}, y)$ corresponding to a set of variables $(\mathbf{V}, Y)$ and its subset $\mathbf{x} \subseteq \mathbf{v}$ corresponding to a set of variables $\mathbf{X} \subseteq \mathbf{V}$, a procedure for assigning contributions only to $x_i \in \mathbf{x}$ is the following:

1. Construct a graph $G[\mathbf{X}]$ composed of nodes in $\mathbf{X}$ and edges added as follows (Tian & Pearl, 2003).
   (a) add a directed edge $V_i \dashrightarrow V_j$ in $G[\mathbf{C}]$ if there exists a directed path from V_i to V_j in G such that every vertex on the path is not in $\mathbf{C}$;
   (b) add a bidirected edge $V_i \dashleftarrow\dashrightarrow V_j$ in $G[\mathbf{C}]$ if there exists a divergent path between V_i and V_j in G such that every vertex on the path is not in $\mathbf{C}$.
2. Construct the do-Shapley w.r.t. $\{y, \mathbf{x}\}$ on $G[\mathbf{X}]$. Specifically, for all $x_i \in \mathbf{x}$

$$\phi_{x_i} := \sum_{\mathbf{x}_S \subseteq \mathbf{x} \setminus x_i} \omega^{\mathbf{x}}(S) \{\mathsf{E}[Y | do(\mathbf{x}_{S \cup i})] - \mathsf{E}[Y | do(\mathbf{x}_S)]\}, \tag{B.1}$$

   where $\omega^{\mathbf{x}}(S) := (1/|\mathbf{x}|)\binom{|\mathbf{x}|-1}{|\mathbf{x}_S|}^{-1}$.

Then, $\{\phi_{x_i}\}_{x_i \in \mathbf{x}}$ is a unique causal contribution measure:

Proposition S.1. *$\{\phi_{x_i}\}_{x_i \in \mathbf{x}}$ is a unique causal contribution measure w.r.t. $\{y, \mathbf{x}\}$ on G.*

Proof. It suffices to show that $G[\mathbf{X}]$ is a graph corresponding to $P(\mathbf{X})$, because of ϕ_{x_i} is the do-Shapley value defined on a graph corresponding to $P(\mathbf{X})$. By (Koster et al., 2002), $G[\mathbf{X}]$ is a graph corresponding to $P(\mathbf{X})$. □

C. Relation with Other Work - Examples

In this section, we provide examples to demonstrate that other types of Shapley values doesn't satisfy the Axiom 1. We first note that the conditional Shapley doesn't satisfy the causal irrelevance property in Axiom 1.

Example C.1 (**Causal Irrelevance Property doesn't hold for the conditional Shapley** (Janzing et al., 2020b))**.** *Consider* $G = \{V_1 \to V_2 \to Y\}$ *where* $V_1, V_2 \in \{0,1\}$*, and the bidirected edge means the existence of hidden confounders. Suppose* $P(v_1, v_2) = 1/2$ *whenever* $v_1 = v_2$*. Note* V_1 *and* Y *is causally irrelevant. Causal irrelevance property doesn't hold in the conditional Shapley. Specifically, for any* v_1, v_2*,* $\mathbb{E}[Y|v_1] - \mathbb{E}[Y] = v_1 - 1/2 \neq 0$*, which leads that* $\phi_{v_1}^{cond} \neq 0$*. In contrast,* $\mathbb{E}[Y|do(v_1)] - \mathbb{E}[Y] = \mathbb{E}[Y|do(v_1, v_2)] - \mathbb{E}[Y|do(v_2)] = 0$*. Therefore,* $\phi_{v_1} = 0$*, implying that do-Shapley satisfies the causal irrelevance property, unlike to the conditional Shapley.*

The ICC doesn't satisfy the causal symmetry property in Axiom 1.

Example C.2 (**Causal Symmetry Property doesn't hold for the ICC Approach**)**.** *Consider a following SCM* $\mathcal{M}$*: For all binary variables* $U_{V_1}, U_{V_2}, U_Y, V_1, V_2, Y \in \{0,1\}$*,* $P(U_1 = 1) = 0.5$*,* $P(U_2 = 1) = 0.2$*, and* $P(U_Y = 1) = 0.8$*. Also,* $V_1 \leftarrow f_{V_1}(U_{V_1}) = U_{V_1}$*;* $V_2 \leftarrow f_{V_2}(V_1, U_2) = V_1 \cdot U_{V_2}$*; and* $Y \leftarrow f_Y(V_2, U_Y) = V_2 \cdot U_Y$*. A corresponding causal diagram is* $G = \{V_1 \to V_2 \to Y, V_1 \dashrightarrow Y\}$*. Let* $y = v_1 = v_2 = 1$*. Then,* $P(y|do(v_1)) = P(y|v_1) = 0.8$*,* $P(y|do(v_2)) = P(y|v_2) = 0.8$*,* $P(y|do(v_1, v_2)) = P(y|v_2) = 0.8$*, and* $P(y) = 0.65$*. We first note that* v_1 *and* v_2 *have the same causal explanatory power to* Y *since* $P(y|do(v_1)) = P(y|do(v_2)) = 0.8$*. Also, the do-Shapley values for* v_1, v_2 *are the same as* $\phi_{v_1} = \phi_{v_2} = 0.075$*, which exhibits the causal symmetry. To compute the ICC of the features* $v_1 = v_2 = 1$*, we fix* $u_1 = 1$ *and* $u_2 = 0$*, which makes* $v_1 = v_2 = 1$*. Let* $\phi_{v_i}^{icc}$ *denote the ICC of* v_i*. Then,* $\phi_{v_1}^{icc} = 0.225$ *and* $\phi_{v_2}^{icc} = 0.075$ *even if* v_1, v_2 *have the same causal explanatory power. This implies that the causal symmetry doesn't hold.*

D. Proofs

We provide complete proofs and additional missing details here.

D.1. Proofs from Section 3

We use

$$\nu_{do}(S) \coloneqq \mathbb{E}[Y|do(\mathbf{v}_S)]$$

in the proof.

Theorem D.1 (Restated Theorem 1)**.** *The do-Shapley is a unique causal contribution measure satisfying all the properties in Axiom 1.*

Proof. We first prove that *do*-Shapley satisfies all the properties in Axiom 1.

Lemma S.1 (**Soundness of** *do*-**Shapley**)**.** *The do-Shapley satisfies all properties in Axiom 1.*

Proof. First, consider the **perfect assignment** property. By the result of (Štrumbelj & Kononenko, 2014), we can represent the *do*-Shapley as

$$\phi_{v_i}(\nu_{do}) = \frac{1}{n!} \sum_{\pi \in \Pi([n])} \{\nu_{do}(\{i\} \cup \text{pre}_\pi(i)) - \nu_{do}(\text{pre}_\pi(i))\},$$

where $\Pi([n])$ is a set of all possible permutations of $[n]$, π is an individual permutation in $\Pi([n])$, and $\text{pre}_\pi(i) \coloneqq \{k \in [n]$ such that $k < i$ in $\pi([n])\}$. Then,

$$\begin{aligned}
\sum_{i=1}^{n} \phi_{v_i}(\nu_{do}) &= \frac{1}{n!} \sum_{\pi \in \Pi([n])} \sum_{i=1}^{n} \{\nu_{do}(\{i\} \cup \text{pre}_\pi(i)) - \nu_{do}(\text{pre}_\pi(i))\} \\
&= \frac{1}{n!} \sum_{\pi \in \Pi([n])} \{\nu_{do}([n]) - \nu_{do}(\emptyset)\} \\
&= \nu_{do}([n]) - \nu_{do}(\emptyset) = \mathbb{E}[Y|do(\mathbf{v})] - \mathbb{E}[Y] = \mathbb{E}[Y|do(\mathbf{v})].
\end{aligned}$$

Now we consider the **causal irrelevance** property. Suppose V_i is causally irrelevant to Y in expectation for all witness $\mathbf{w}$ $\mathbf{v}nfv_ig$. Then, the equality $\nu_{do}(S \,[\, i) \quad \nu_{do}(S) = 0$ holds immediately for all S $[n]nfig$.

Next we consider the **causal symmetry** property. Suppose v_i, v_j has the same causal explanatory power w.r.t. any witnesses $\mathbf{w}$ $\mathbf{v}nfv_i, v_jg$. This leads $\nu_{do}(fig \,[\, S) = \nu_{do}(fjg \,[\, S)$ for any S $[n]nfi, jg$. Then,

$$
\begin{aligned}
{}_{v_i}(\ {}_{do}) &= \frac{1}{n} \sum_{S \subseteq [n]\setminus\{i\}} \binom{n\ \ 1}{jSj}^{-1} f\ {}_{do}(S \,[\, i) \quad {}_{do}(S)g \\
&= \frac{1}{n} \sum_{S \subseteq [n]\setminus\{i,j\}} \binom{n\ \ 1}{jSj}^{-1} f\ {}_{do}(S \,[\, i) \quad {}_{do}(S)g + \frac{1}{n} \sum_{S \subseteq [n]\setminus\{i,j\}} \binom{n\ \ 1}{jSj+1}^{-1} f\ {}_{do}(S \,[\, fi;jg) \quad {}_{do}(S \,[\, fjg)g \\
&= \frac{1}{n} \sum_{S \subseteq [n]\setminus\{i,j\}} \binom{n\ \ 1}{jSj}^{-1} f\ {}_{do}(S \,[\, j) \quad {}_{do}(S)g + \frac{1}{n} \sum_{S \subseteq [n]\setminus\{i,j\}} \binom{n\ \ 1}{jSj+1}^{-1} f\ {}_{do}(S \,[\, fi;jg) \quad {}_{do}(S \,[\, fig)g \\
&= \frac{1}{n} \sum_{S \subseteq [n]\setminus\{j\}} \binom{n\ \ 1}{jSj}^{-1} f\ {}_{do}(S \,[\, j) \quad {}_{do}(S)g = \ {}_{v_j}(\ {}_{do});
\end{aligned}
$$

where the third equality holds since (v_i, v_j) has the same causal explanatory power.

Now, we prove that the do-Shapley satisfies the **causal approximation** property by showing that there exists $\omega(S)$ that makes the do-Shapley as the solution of the weighted least square problem defined in Axiom 1. For a coalition function $\nu(S)$ (see the "Shapley value" paragraph in Sec. 2), it's known that there exists a specific weight function $\omega(S)$ that makes the Shapley value in Eq. (1) as the solution of the following WLS problem: $\arg\min_{f\ {}^{0}_{v_i}g^{n}_{i=1}} \sum_{S\ \ [n]} (\nu(S) \quad \sum_{i2S} \phi^{0}_{v_i})^2 \omega(S)$ (by (Charnes et al., 1988, Thm. 4) and (Lundberg & Lee, 2017, Theorem 2)). This implies that such an $\omega(S)$ is the weight function that makes the do-Shapley as the solution of the weighted least square problem defined in Axiom 1.

□

We now show the other direction that a measure satisfying all properties in Axiom 1 is the do-Shapley.

Lemma S.2 (Completeness of do-Shapley). *A vector $f\phi_{v_i}g_{v_i 2\mathbf{v}}$ satisfying Axiom 1 is the do-Shapley value.*

Proof. Throughout the proof, we will define a *canonical SCM* as follow: Let $T \quad [n]$ denote any fixed index set. A SCM is called *canonical* for T if $\mathsf{E}[Yjdo(\mathbf{V}_S = 1)] = 1$ iff $T \quad S$, and 0 otherwise. We use $\nu^T_{do}(S)$ denote the causal coalition function induced by the canonical SCM. Note $\nu^T_{do}(S) = 1$ iff $T \quad S$, and 0 otherwise, by the definition of the canonical SCM.

We first note that a vector ϕ_{v_i} that satisfies the *causal approximation* property can be represented as a linear function of $\nu_{do}(S)$, because ϕ_{v_i} is a solution of the weighted least square linear regression problem. Therefore,

$$\phi_{v_i} = \sum_{S \quad [n]} a^i_S \nu_{do}(S). \tag{D.1}$$

for some constants fa^i_Sg.

Now we focus on the **causal irrelevance** property. Suppose $T \quad [n]nfig$. For any $S \quad [n]$, $(T \quad S) \ =) \ (T \quad S \,[\, fig)$. With $i \,\mathbf{2}\, T$, $(T \,\mathbf{6}\, S) \ =) \ (T \,\mathbf{6}\, S \,[\, fig)$. Therefore, $\nu^T_{do}(S) = \nu^T_{do}(S \,[\, i)$ for all $S \quad [n]$. Then, by the causal irrelevance property, $\phi_{v_i}(\nu^T_{do}) = 0$ if $T \quad [n]ni$. Then, $\phi_{v_i}(\nu^{[n]ni}_{do}) = a^i_{[n]} + a^i_{[n]ni} = 0$.

Suppose it has been shown that $a^i_{T\,[\,i} + a^i_T = 0$ for $T \quad [n]ni$ such that $jTj \quad k$ for some k. Then, for any $S \quad [n]ni$ such that $jSj = k \quad 1$,

$$
\begin{aligned}
\phi_{v_i}(\nu^S_{do}) &= \sum_{T \quad [n]} a^i_T \nu^S_{do}(T) = \sum_{\substack{T \\ T}} \sum_{\substack{[n] \\ S}} a^i_T = \sum_{T \ \ \substack{[n]nfig \\ T \quad S}} \left(a^i_{T\,[\,i} + a^i_T\right) \\
&= \left\{ \sum_{\substack{T \\ T}} \sum_{\substack{[n]ni \\ S \text{ but } T \not\in S}} \left(a^i_{T\,[\,i} + a^i_T\right) \right\} + \left(a^i_{S\,[\,i} + a^i_S\right) = a^i_{S\,[\,i} + a^i_S,
\end{aligned}
$$

where the first equality by Eq. (D.1), the second by the property of the canonical SCM, the third and fourth by the standard algebra, and the fifth by the inductive hypothesis. Since $S \subseteq [n] \setminus \{i\}$, by causal irrelevance property, $\phi_{v_i}(\nu^S_{do}) = 0$. This implies that $a^i_{S \cup \{i\}} + a^i_S = 0$. Therefore, for any $T \subseteq [n] \setminus i$, $a^i_{T \cup \{i\}} + a^i_T = 0$.

Fix $p^i_T := a^i_{T \cup \{i\}} = -a^i_T$. Then,

$$\phi_{v_i}(\nu_{do}) = \sum_{T \subseteq [n]} a^i_T \nu_{do}(T) = \sum_{T \subseteq [n] \setminus i} \left(a^i_{T \cup \{i\}} \nu_{do}(T \cup \{i\}) + a^i_T \nu_{do}(T) \right) = \sum_{T \subseteq [n] \setminus i} p^i_T \left(\nu_{do}(T \cup \{i\}) - \nu_{do}(T) \right).$$

Now we focus on the **causal symmetry** property. Suppose v_i and v_j have the same causal explanatory power with any given witness $\mathbf{w} \subseteq \mathbf{V} \setminus \{v_i, v_j\}$ in the canonical SCM for $[n]$; i.e., $\nu^{[n]}_{do}(S \cup \{i\}) = \nu^{[n]}_{do}(S \cup \{j\})$ for $S \subseteq [n] \setminus \{i, j\}$. We note that $\phi_{v_i}(\nu^{[n]}_{do}) = p^i_{[n] \setminus i} = \phi_{v_j}(\nu^{[n]}_{do}) = p^j_{[n] \setminus j}$. This implies that there exists $p_{n-1} := p^i_{[n] \setminus i} = p^j_{[n] \setminus j}$.

Again, suppose v_i, v_j have the same causal explanatory power with any given witness $\mathbf{w} \subseteq \mathbf{V} \setminus \{v_i, v_j\}$ in the canonical SCM for $[n] \setminus k$ for any fixed $k \notin \{i, j\}$. Then,

$$\phi_{v_i}(\nu^{[n] \setminus k}_{do}) = p^i_{[n] \setminus \{i,k\}} + p_{n-1} = \phi_{v_j}(\nu^{[n] \setminus k}_{do}) = p^j_{[n] \setminus \{j,k\}} + p_{n-1}.$$

This implies that there exists a constant $p_{n-2} := p^i_{[n] \setminus \{i,k\}} = p^j_{[n] \setminus \{j,k\}}$. By repeating this, we can have a $p_1, \ldots, p_{n-1}$ where $p_{|T|}$ is a constant applying to all p^j_T for any $T \subseteq [n] \setminus i$. Therefore, there are constants $\{p_{|T|}\}_{T \subseteq [n] \setminus i}$ such that

$$\phi_{v_i} := \sum_{T \subseteq [n] \setminus i} p_{|T|} \left(\nu_{do}(T \cup \{i\}) - \nu_{do}(T) \right).$$

Finally, we focus on the **perfect assignment** property. An attribution ϕ_{v_i} satisfies the perfect assignment property if and only if $\sum_{i \in N} p_{n-1} = 1$, and for any nonempty $T \subseteq [n]$, $\sum_{i \in T} p_{|T|-1} = \sum_{j \notin T} p_{|T|}$ (Winter, 2002, Chap. 7, Theorem 11). This gives $p_{n-1} = 1/n$, and for any nonempty $T \subseteq [n]$, $|T| p_{|T|-1} = (n - |T|) p_{|T|}$. Then, a closed form for $p_{|T|}$ is given as

$$p_{|T|} = \frac{(n - |T| - 1)! |T|!}{n!} = \frac{1}{n} \binom{n-1}{|T|}.$$

□

Taking a conjunction of Lemmas (S.1,S.2) completes the proof of the Theorem D.1. □

D.2. Proofs from Section 4

Theorem D.2 (Restated Theorem 2)**.** *The do-Shapley is identifiable if no variable in $V_i \in \{\mathbf{V} \setminus \mathrm{Ch}(V_i)\}$ is connected to its child $\mathrm{Ch}(V_i)$ by bidirected paths in G. Suppose Y is not connected by bidirected paths. In this case, for any $S \subseteq [n]$,*

$$\mathsf{E}[Y|do(\mathbf{v}_S)] = \sum\nolimits_{\mathbf{v}_{\overline{S}}} \mathsf{E}\left[Y|\mathbf{v}\right] Q\left[\mathbf{V} \setminus \mathbf{V}_S\right],$$

where $Q\left[\mathbf{V} \setminus \mathbf{V}_S\right] := Q\left[\mathbf{V} \setminus \mathbf{V}_S\right](\mathbf{v})$ is given as

$$Q\left[\mathbf{V} \setminus \mathbf{V}_S\right] = \frac{P(\mathbf{v})}{Q\left[C(\mathbf{V}_S)\right]} \prod_{k=1}^{c} \sum_{\mathbf{S}_k} Q\left[C(\mathbf{S}_k)\right],$$

where $Q\left[C(\mathbf{V}_S)\right] = \prod_{V_a \in C(\mathbf{V}_S)} P(v_a|pre(v_a))$ is a C-factor of a C-component $\mathbf{V}_S$ ($C(\mathbf{V}_S)$); $\{\mathbf{S}_k\}_{k=1}^c$ is a C-partition of $\mathbf{V}_S$; and $Q\left[C(\mathbf{S}_k)\right] := \prod_{V_a \in C(\mathbf{S}_k)} P(v_a|pre(v_a))$ is a C-factor of a C-component $C(\mathbf{S}_k)$ for $\mathbf{S}_k$.

Proof. We prove the following, which would imply the above theorem.

Proposition S.1 (Generalized Tian's Adjustment – Complete identification criteria for $P(\mathbf{V}|do(\mathbf{X}))$). *$P(\mathbf{V}|do(\mathbf{X}))$ is identifiable if $\forall X_a \in \mathbf{X}$ and $\mathrm{Ch}(X_a)$ is not connected by bidirected paths. If identifiable, it's given as*

$$P(\mathbf{V}|do(\mathbf{X})) = \frac{P(\mathbf{V})}{Q\left[C(\mathbf{X})\right]} \prod_{k=1}^{c} \sum_{\mathbf{x}_k} Q\left[C(\mathbf{X}_k)\right], \tag{D.2}$$

where $\{\mathbf{X}_k\}_{k=1}^{c}$ is a C-partition of $\mathbf{X}$ in G, and $Q\left[C(\mathbf{X})\right] \coloneqq \prod_{V_i \in C(\mathbf{X})} P(V_i|\mathrm{pre}(V_i))$ is a C-factor of a C-component of $\mathbf{X}$ ($C(\mathbf{X})$), and $Q\left[C(\mathbf{X}_k)\right] \coloneqq \prod_{V_i \in C(\mathbf{X}_k)} P(V_i|\mathrm{pre}(V_i))$ is a C-factor of a C-component of $\mathbf{X}_k$ ($C(\mathbf{X}_k)$).

Proof. In the proof, for a vector $\mathbf{W}$, we will use $\mathrm{De}(\mathbf{W})$ to denote a set of descendants of $W_i \in \mathbf{W}$ in G.

Suppose $\forall X_a \in \mathbf{X}$ is not connected with $\mathrm{Ch}(X_a)$ by bidirected paths. We first show that $P(\mathbf{V}|do(\mathbf{X}_1))$ (for any $\mathbf{X}_1 \in \mathbf{X}$ a C-component in $G(\mathbf{X})$) is identifiable and given as

$$P(\mathbf{V}|do(\mathbf{X}_1)) = \frac{P(\mathbf{V})}{Q\left[C(\mathbf{X}_1)\right]} \sum_{\mathbf{x}_1} Q\left[C(\mathbf{X}_1)\right]. \tag{D.3}$$

By the result of (Jaber et al., 2018, Lemma 1), it suffices to show that $\mathbf{X}_1 = \mathrm{De}(\mathbf{X}_1)_{G(C(\mathbf{X}_1))}$. We show this by contradiction. Suppose $V_a \in \mathrm{De}(\mathbf{X}_1)_{G(C(\mathbf{X}_1))}$ such that $V_a \notin \mathbf{X}_1$. Since $V_a \in G(C(\mathbf{X}_1))$, V_a is connected with $\mathbf{X}_1$ by bidirected paths. Since V_a is a descendent of some $X_a \in \mathbf{X}_1$ in $G(C(\mathbf{X}_1))$, this means that $V_b \in \mathrm{Ch}(X_1)$ is also in the $G(C(\mathbf{X}_1))$. This means that V_b and X_a is connected by a bidirected path, which is a contradiction of the given condition. Therefore, $\mathbf{X}_1 = \mathrm{De}(\mathbf{X}_1)_{G(C(\mathbf{X}_1))}$, and Eq. (D.3) holds.

Now, consider a following inductive hypothesis for $i = 1, 2, \ldots, c$:

$$Q\left[\mathbf{V} \cap \mathbf{X}^{(i)}\right] = \frac{Q\left[\mathbf{V} \cap \mathbf{X}^{(i-1)}\right]}{Q\left[C(\mathbf{X}_i)\right]} \sum_{\mathbf{x}_i} Q\left[C(\mathbf{X}_i)\right]. \tag{D.4}$$

As shown in the above, it holds for $i = 1$. Suppose it holds for some $i - 1 \geq 1$ for $i \geq 2$. Then, we first note that $\mathbf{X}_i = \mathrm{De}(\mathbf{X}_i)_{G(C(\mathbf{X}_i))_{G(\mathbf{V} \cap \mathbf{X}^{(i-1)})}}$. To witness, consider the contradiction – for some $X_a \in \mathbf{X}_i$ there exists $V_a \in \mathrm{De}(\mathbf{X}_i)_{G(C(\mathbf{X}_i))_{G(\mathbf{V} \cap \mathbf{X}^{(i-1)})}}$ s.t. $V_a \notin \mathbf{X}_i$. First, V_a is connected with X_a by bidirected paths since $V_a \in G(C(\mathbf{X}_i))_{G(\mathbf{V} \cap \mathbf{X}^{(i-1)})}$. Also, V_a is a descendent of X_a, this means that a child of X_a is also in $G(C(\mathbf{X}_i))$, connected by bidirected paths. This is a contradiction. Therefore, $\mathbf{X}_i = \mathrm{De}(\mathbf{X}_i)_{G(C(\mathbf{X}_i))_{G(\mathbf{V} \cap \mathbf{X}^{(i-1)})}}$.

Now, we show that $C(\mathbf{X}_i)_{G(\mathbf{V} \cap \mathbf{X}^{(i-1)})} = C(\mathbf{X}_i)_G$. We start from an obvious observation – $C(\mathbf{X}_i)_{G(\mathbf{V} \cap \mathbf{X}^{(i-1)})} \subseteq C(\mathbf{X}_i)_G$. We now prove $C(\mathbf{X}_i)_G \subseteq C(\mathbf{X}_i)_{G(\mathbf{V} \cap \mathbf{X}^{(i-1)})}$. For some $V_a \in C(\mathbf{X}_i)_G$, suppose $V_a \notin C(\mathbf{X}_i)_{G(\mathbf{V} \cap \mathbf{X}^{(i-1)})}$. This means that bidirected paths connecting V_a to some nodes in $X_1 \in \mathbf{X}_i$ must be via other nodes in $X_2 \in \mathbf{X}^{(i-1)}$. This means that V_a, X_2, X_1 are connected by bidirected paths. However, given that $X_2 \in \mathbf{X}^{(i-1)}$ and $X_1 \in \mathbf{X}_i$, this is a contradiction, because they are in distinct C-partitions. Therefore, $C(\mathbf{X}_i)_{G(\mathbf{V} \cap \mathbf{X}^{(i-1)})} = C(\mathbf{X}_i)_G$.

Then, Eq. (D.4) holds. By unfolding it,

$$Q\left[\mathbf{V} \cap \mathbf{X}^{(i)}\right] = \frac{P(\mathbf{V})}{\prod_{k=1}^{i} Q\left[C(\mathbf{X}_i)\right]} \prod_{k=1}^{c} \sum_{\mathbf{x}_i} Q\left[C(\mathbf{X}_i)\right].$$

We note $Q\left[C(\mathbf{X}^{(i)})\right] = \prod_{k=1}^{i} Q\left[C(\mathbf{X}_i)\right]$, since

$$Q\left[C(\mathbf{X}^{(i)})\right] = \prod_{V_i \in C(\mathbf{X}^{(i)})} P(v_i|\mathrm{pre}(v_i)) = \prod_{k=1}^{c} \prod_{V_i \in C(\mathbf{X}_k)} P(v_i|\mathrm{pre}(v_i)) = \prod_{k=1}^{c} Q\left[C(\mathbf{X}_k)\right].$$

This completes the proof. □

Now back to witness Thm. D.2. Under the given condition that Y is not connected via bidirected paths to any nodes, the following holds: for any $S \subseteq [n]$,

$$(Y \perp\!\!\!\perp \mathbf{V}_S | \mathbf{V}_{\overline{S}})_{G_{\underline{\mathbf{V}_S}}}.$$

Therefore,

$$P(Y, \mathbf{V} j do(\mathbf{V}_S)) = P(Y j do(\mathbf{V}_S), \mathbf{V}_{\overline{S}}) Q[\mathbf{V}_S] = P(Y j \mathbf{V}) Q[\mathbf{V}_S],$$

which implies that

$$\mathsf{E}[Y j do(\mathbf{V}_S)] = \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}[Y j \mathbf{V}] \frac{Q[\mathbf{V}]}{Q[C(\mathbf{V}_S)]} \prod_{k=1}^{c} \sum_{\mathbf{x}_k} Q[C(\mathbf{X}_k)].$$

This completes the proof. □

Corollary D.2 (Restated Corollary 1)**.** *In the Markovian case,* $\mathsf{E}[Y j do(\mathbf{v}_S)]$ *is given as*

$$\mathsf{E}[Y j do(\mathbf{v}_S)] = \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}[Y j \mathbf{v}_S, \mathbf{v}_{\overline{S}}] \prod_{i \in S} P(v_i j pre(v_i)).$$

Proof. In the Markovian case, $C(\mathbf{W}) = \mathbf{W}$ for all $\mathbf{W} \quad \mathbf{V}$. Then,

$$P(Y, \mathbf{V}_{\overline{S}} j do(\mathbf{V}_S)) = \frac{P(\mathbf{V}, Y)}{Q[C(\mathbf{V}_S)]} \prod_{k=1}^{c} \sum_{\mathbf{x}_k} Q[C(\mathbf{X}_k)] = \frac{P(\mathbf{V}, Y)}{Q[\mathbf{V}_S]} = P(Y j \mathbf{V}) \prod_{V_i \in \mathbf{V}_{\overline{S}}} P(V_i j \text{pre}(V_i)).$$

This completes the proof. □

Corollary D.2 (Restated Corollary 2)**.** *In the Direct-cause case,* $\mathsf{E}[Y j do(\mathbf{v}_S)]$ *is given as*

$$\mathsf{E}[Y j do(\mathbf{v}_S)] = \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}[Y j \mathbf{v}_S, \mathbf{v}_{\overline{S}}] P(\mathbf{v}_{\overline{S}}).$$

Proof. In the Direct-cause case, $Q[\mathbf{W}] = P(\mathbf{W})$ for all $\mathbf{W} \quad \mathbf{V}$ since there are no causal paths between a pair of variables in $\mathbf{V}$. Therefore, $Q[\mathbf{V} n \mathbf{V}_S] = P(\mathbf{V} n \mathbf{V}_S) = P(\mathbf{V}_{\overline{S}})$, which completes the proof. □

D.3. Proofs from Section 5

Lemma D.1 (Restated Lemma 1)**.** *Let* $S = \{m_1, \quad , m_s\} \quad [n]$ *denote an index set for* $\mathbf{V}_S$*. Let*

$$\omega_k^S := \prod_{r=1}^{k} \mathbb{1}_{v_{m_r}}(V_{m_r}) / h_r^S, \text{ for } k = s, \quad , 1;$$
$$\omega^S := \mathbb{1}_{\mathbf{v}_S}(\mathbf{V}_S) / h^S,$$

where $h_r^S := P(V_{m_r} j pre(V_{m_r}))$ *and* $h^S := P(\mathbf{V}_S j \mathbf{V}_{\overline{S}})$*. Then,* $\mathsf{E}[Y j do(\mathbf{v}_S)] = \mathsf{E}[Y \omega]$ *where* $\omega = \omega_k^S$ *for the Markovian case, and* $\omega = \omega^S$ *for the Direct-cause case.*

Proof. For the Markovian case,

$$\mathsf{E}[Y j do(\mathbf{v}_S)] = \sum_{\mathbf{v}_{\overline{S}}} \frac{P(\mathbf{v})}{\prod_{r=1}^{s} P(v_{m_r} j \text{pre}(v_{m_r})}) = \mathsf{E}\left[\frac{\mathbb{1}_{v_{m_r}}(V_{m_r})}{h_r^S} \right].$$

For the Direct-cause case,

$$\mathsf{E}[Y j do(\mathbf{v}_S)] = \sum_{\mathbf{v}_{\overline{S}}} \frac{P(\mathbf{v})}{P(\mathbf{v}_S j \mathbf{v}_{\overline{S}})} = \mathsf{E}\left[\frac{\mathbb{1}_{\mathbf{v}_S}(\mathbf{V}_S)}{P(\mathbf{V}_S j \mathbf{V}_{\overline{S}})} \right].$$

□

Lemma D.2 (Restated Lemma 2)**.** *Let* $S \coloneqq \{m_1, \ldots, m_s\} \subseteq [n]$ *denote an index set for* $\mathbf{V}_S$*. Let* $\theta^S_{s;1} \coloneqq Y$*. For* $k = s, s-1, \ldots, 1$*,*

$$\begin{aligned}
\theta^S_{k;2} &\coloneqq \mathsf{E}\left[\theta^S_{k;1} | V_{m_k}, pre(V_{m_k})\right] \\
\theta^S_{k-1;1} &\coloneqq \mathsf{E}\left[\theta^S_{k;1} | v_{m_k}, pre(V_{m_k})\right], \\
\theta^S_a &\coloneqq \mathsf{E}\left[Y | \mathbf{v}_S, \mathbf{V}_{\overline{S}}\right], \\
\theta^S_b &\coloneqq \mathsf{E}\left[Y | \mathbf{V}_S, \mathbf{V}_{\overline{S}}\right].
\end{aligned}$$

Then, $\mathsf{E}[Y | do(\mathbf{v}_S)] = \mathsf{E}[\theta]$ *where* $\theta = \theta^S_{0;1}$ *for the Markovian case, and* $\theta = \theta^S_a$ *for the Direct-cause case.*

Proof. For the Markovian case, we will prove the following, which implies the result.

Lemma S.3. *Suppose* $\mathbf{V}^0 = \{Y\} \cup \mathbf{V}$ *where* $\mathbf{V}^0$ *is an ordered set. Assume that* Y *is the last variable in the given order. Let* $\mathbf{V}_S \coloneqq \{V_{m_1}, \ldots, V_{m_s}\} \subseteq \mathbf{V}$ *(where* $\{m_1, \ldots, m_s\} \subseteq [n]$*) denote a set of discrete variables. Let* $\mathbf{V}_{\overline{S}} \coloneqq \mathbf{V} \setminus \mathbf{V}_S$*. For each* $k = 2, \ldots, s$*, let* $\mathbf{V}_{\setminus_k} \coloneqq \{V_j \in \mathbf{V}_{\overline{S}} : V_{m_{k-1}} \prec V_j \prec V_{m_k}\}$*. Let* $\mathbf{V}_{\setminus_1} \coloneqq \{V_j \in \mathbf{V}_{\overline{S}} : V_j \prec V_{m_k}\}$ *and* $\mathbf{V}_{\setminus_{s+1}} \coloneqq \{V_j \in \mathbf{V}_{\overline{S}} : V_{m_s} \prec V_j\}$*.*

Let $g_S(P)$ *denote a following functional (a.k.a. g-formula (Robins, 1986)).*

$$g_S(P) \coloneqq \int_{V_{\overline{S}}} \mathsf{E}[Y | \mathbf{v}] \prod_{i \in S} P(v_i | pre(v_i))\, d[\mathbf{v}_{\overline{S}}].$$

Let $\theta_{s;1} \coloneqq Y$*. For* $k = s, \ldots, 1$*, and*

$$\begin{aligned}
\theta_{k;2} &\coloneqq \mathsf{E}[\theta_{k;1} | V_{m_k}, pre(V_{m_k})] \\
\theta_{k-1;1} &\coloneqq \mathsf{E}[\theta_{k;1} | v_{m_k}, pre(V_{m_k})].
\end{aligned}$$

Then, the following holds:

$$g_S(P) = \mathsf{E}[\theta_{0;1}].$$

Proof. Let

$$\begin{aligned}
\mathbf{A}_k &\coloneqq \{\text{pre}(V_{m_k})\} \\
\mathbf{B}_k &\coloneqq \{\mathbf{V}_{\setminus_{k+1}}, \mathbf{V}_{\setminus_{k+2}}, \ldots, \mathbf{V}_{\setminus_{s+1}}\} \\
\mathbf{C}_k &\coloneqq \{V_{m_{k+1}}, V_{m_{k+2}}, \ldots, V_{m_s}\}.
\end{aligned}$$

For $\mathbf{W} \subseteq \mathbf{V}$,

$$q(\mathbf{W}) \coloneqq \begin{cases} \prod_{V_i \in \mathbf{W}} P(v_i | \text{pre}(v_i)) & \text{If } \mathbf{W} \neq \emptyset; \\ 1 & \text{If } \mathbf{W} = \emptyset. \end{cases}$$

Then, it suffices to show that

$$\begin{aligned}
\theta_{k;2} &= \int_{B_k; C_k} \mathsf{E}[Y | V_{m_k}, \mathbf{A}_k, \mathbf{b}_k, \mathbf{c}_k]\, q(\mathbf{b}_k) \mathbb{1}_{\mathbf{c}_k}(\mathbf{C}_k)\, d[\mathbf{b}_k, \mathbf{c}_k] \\
\theta_{k-1;1} &= \int_{B_k; C_{k-1}} \mathsf{E}[Y | \mathbf{A}_k, \mathbf{b}_k, \mathbf{c}_{k-1}]\, q(\mathbf{b}_k) \mathbb{1}_{\mathbf{c}_{k-1}}(\mathbf{C}_{k-1})\, d[\mathbf{b}_k, \mathbf{c}_{k-1}],
\end{aligned}$$

because witnessing $\mathsf{E}[\theta_{0;1}] = g_S(P)$ becomes trivial. Let $\theta_{s;1} \coloneqq Y$. Then, it's easy to check that the above holds for $\theta_{s;2}$ and $\theta_{s-1;1}$.

Suppose the above equation holds for $k, k+1, \ldots, s$. Then, consider $k-1$. By the given definition,

$$\begin{aligned}
\theta_{k-1;2} &\coloneqq \mathsf{E}\left[\theta_{k-1;1} | V_{m_{k-1}}, \text{pre}(V_{m_{k-1}})\right] \\
\theta_{k-2;1} &\coloneqq \mathsf{E}\left[\theta_{k-1;1} | v_{m_{k-1}}, \text{pre}(V_{m_{k-1}})\right].
\end{aligned}$$

Then,

$$\theta_{k\ 1;2} = \mathsf{E}\left[\int_{B_k;C_{k\ 1}} \mathsf{E}\left[Y \middle| \mathbf{A}_k, \mathbf{b}_k, \mathbf{c}_{k\ 1}\right] q(\mathbf{b}_k)\mathbb{1}_{\mathbf{c}_{k\ 1}}(\mathbf{C}_{k\ 1})\, d[\mathbf{b}_k, \mathbf{c}_{k\ 1}] \middle| V_{m_{k\ 1}}, \mathbf{A}_{k\ 1}\right]$$
$$= \int_{B_{k\ 1};C_{k\ 1}} \mathsf{E}\left[Y \middle| V_{m_{k\ 1}}, \mathbf{A}_{k\ 1}, \mathbf{b}_{k\ 1}, \mathbf{c}_{k\ 1}\right] q(\mathbf{b}_{k\ 1})\mathbb{1}_{\mathbf{c}_{k\ 1}}(\mathbf{C}_{k\ 1})\, d[\mathbf{b}_{k\ 1}, \mathbf{c}_{k\ 1}],$$

and

$$\theta_{k\ 2;1} = \mathsf{E}\left[\int_{B_k;C_{k\ 1}} \mathsf{E}\left[Y \middle| \mathbf{A}_k, \mathbf{b}_k, \mathbf{c}_{k\ 1}\right] q(\mathbf{b}_k)\mathbb{1}_{\mathbf{c}_{k\ 1}}(\mathbf{C}_{k\ 1})\, d[\mathbf{b}_k, \mathbf{c}_{k\ 1}] \middle| v_{m_{k\ 1}}, \mathbf{A}_{k\ 1}\right]$$
$$= \int_{B_{k\ 1};C_{k\ 2}} \mathsf{E}\left[Y \middle| \mathbf{A}_{k\ 1}, \mathbf{b}_{k\ 1}, \mathbf{c}_{k\ 2}\right] q(\mathbf{b}_{k\ 1})\mathbb{1}_{\mathbf{c}_{k\ 2}}(\mathbf{C}_{k\ 2})\, d[\mathbf{b}_{k\ 1}, \mathbf{c}_{k\ 2}].$$

Therefore,

$$\theta_{0;1} = \int_{B_1;C_0} \mathsf{E}\left[Y \middle| \mathbf{A}_1, \mathbf{b}_1, \mathbf{c}_0\right] q(\mathbf{b}_1)\mathbb{1}_{\mathbf{c}_0}(\mathbf{C}_0)\, d[\mathbf{b}_1, \mathbf{c}_0],$$

which gives the equality $\mathsf{E}[\theta_{0;1}] = g_S(P)$. □

For the Direct-cause case,

$$\mathsf{E}\left[\theta_a^S\right] = \sum_{\mathbf{v}_{\overline{S}}} \mathsf{E}\left[Y \middle| \mathbf{v}\right] P(\mathbf{v}_{\overline{S}}) = \mathsf{E}[Y|do(\mathbf{v}_S)],$$

which completes the proof. □

Lemma D.3 (Restated Lemma 3)**.** *Let*

$$\eta^S := \begin{cases} \{\theta_{0;1}^S\} \cup \{\theta_{k;1}^S, \theta_{k;2}^S\}_{k=1}^s \cup \{h_r^S\}_{r=1}^s & \textit{(Markovian)} \\ \{\theta_a^S, \theta_b^S, h^S\} & \textit{(Direct-cause)}, \end{cases}$$

defined in Defs. (4, 5) above, and

$$V_S(\mathbf{V}^0; \eta^S) := \begin{cases} \theta_{0;1}^S + \sum_{k=1}^s \omega_k^S(\theta_{k;1}^S \quad \theta_{k;2}^S) & \textit{(Markovian)} \\ \theta_a^S + \omega^S\left(Y \quad \theta_b^S\right) & \textit{(Direct-cause)}, \end{cases}$$

where $\omega_k^S := \prod_{r=1}^k \mathbb{1}_{v_{m_r}}(V_{m_r})/h_r^S$ *and* $\omega^S := \mathbb{1}_{\mathbf{v}_S}(\mathbf{V}_S)/h^S$*. Then,* $\mathsf{E}[Y|do(\mathbf{v}_S)] = \mathsf{E}\left[V_S(\mathbf{V}^0; \eta^S)\right]$.

Proof. For the Markovian case, it suffices to show that $\mathsf{E}\left[\theta_{k;1}^S \quad \theta_{k;2}^S\right]$ for any $k = 1, 2, \quad , s$. This holds since

$$\mathsf{E}\left[\theta_{k;1}^S \quad \theta_{k;2}^S\right] = \mathsf{E}\left[\mathsf{E}\left[\theta_{k;1}^S \quad \theta_{k;2}^S \middle| V_{m_k}, \mathrm{pre}(V_{m_k})\right]\right] = \mathsf{E}\left[\theta_{k;2}^S \quad \theta_{k;2}^S\right] = 0.$$

Therefore, $\mathsf{E}\left[V(\mathbf{V}^0; \eta^S)\right] = \mathsf{E}\left[\theta_{0;1}^S\right] = \mathsf{E}[Y|do(\mathbf{v}_S)]$, where the 2nd equality holds by Lemma 2.

For the Direct-case case, $\mathsf{E}\left[Y \quad \theta_2^S\right] = 0$ by the definition of θ_2^S. Therefore, $\mathsf{E}\left[V(\mathbf{V}^0; \eta^S)\right] = \mathsf{E}\left[\theta_1^S\right] = \mathsf{E}[Y|do(\mathbf{v}_S)]$. □

Theorem D.3 (Restated Theorem 3)**.** *Let* $\{\pi_j\}_{j=1}^M$ *denote* M *randomly generated permutations of* $[n]$*. For the fixed index* i*, let* $S_{j;0} := pre_{\ j}(i)$ *and* $S_{j;1} := \{i\} \cup S_{j;a}$*. Let* $\{\widehat{\eta}^{S_{j,0}}, \widehat{\eta}^{S_{j,1}}\}_{j=1}^M$ *denote* L_2*-consistent estimates for all nuisances* $\{\eta^{S_{j,0}}, \eta^{S_{j,1}}\}_{j=1}^M$ *defined in Def. 6. Let* $R_{M;N} := O_P(M^{-1/2} + N^{-1/2})$*. Let* $e(\hat{g}) := \|\hat{g} - g\|$ *denote an error for a nuisance estimates for any* $\hat{g} \in \hat{\eta}$ *and* $g \in \eta$*. For the do-Shapley estimators defined in Def. 7, suppose the estimators* $T^{\mathsf{est}}(S)$ *are bounded. Let* $\epsilon_{v_i}^{\mathsf{est}} := \phi_{v_i}^{\mathsf{est}} - \phi_{v_i}$ *(where* $\mathrm{est} \in \{\mathrm{ipw}, \mathrm{reg}, \mathrm{dml}\}$*).*

Under the Markovian case,

$$\epsilon_{v_i}^{\text{ipw}} = R_{M;N} + O_P\Big\{\sum_{p\in\{0;1\}^M}\sum_{j=1}^{M} e(\hat{\omega}_{s_j}^{s_{j,p}})\Big\},$$

$$\epsilon_{v_i}^{\text{reg}} = R_{M;N} + O_P\Big\{\sum_{p\in\{0;1\}^M}\sum_{j=1}^{M} e(\hat{\theta}_{0;1}^{s_{j,p}})\Big\},$$

$$\epsilon_{v_i}^{\text{dml}} = R_{M;N} + O_P\Big\{\sum_{p\in\{0;1\}^M}\sum_{j=1}^{M}\sum_{k=1}^{s_j} e(\hat{h}_k^{s_{j,p}})e(\hat{\theta}_{k;2}^{s_{j,p}})\Big\}.$$

Under the Direct-cause case,

$$\epsilon_{v_i}^{\text{ipw}} = R_{M;N} + O_P\Big\{\sum_{p\in\{0;1\}^M}\sum_{j=1}^{M} e(\hat{\omega}^{s_{j,p}})\Big\},$$

$$\epsilon_{v_i}^{\text{reg}} = R_{M;N} + O_P\Big\{\sum_{p\in\{0;1\}^M}\sum_{j=1}^{M} e(\hat{\theta}_2^{s_{j,p}})\Big\},$$

$$\epsilon_{v_i}^{\text{dml}} = R_{M;N} + O_P\Big\{\sum_{p\in\{0;1\}^M}\sum_{j=1}^{M} e(\hat{h}^{s_{j,p}})e(\hat{\theta}_b^{s_{j,p}})\Big\}.$$

Proof. In the proof, we will use a notation $\mathsf{E}_{D\ \ P}[f(\mathbf{V})]$ for $f(\mathbf{V}) \coloneqq \mathsf{E}_D[f(\mathbf{V})] \quad \mathsf{E}[f(\mathbf{V})]$. We use $N \coloneqq |D|$. Also, for any quantity A, B, $A \ . \ B$ if there is a constant c s.t. $A \quad cB$. We first introduce a useful tool for analyzing errors of the proposed estimator.

Lemma S.4. *Let η_0 denote some nuisance and $\hat{\eta}$ denote its L_2 consistent estimate. Let $f(\mathbf{V};\eta)$ denote an arbitrary function having a bounded second moment for any fixed η. Suppose samples used for constructing $\hat{\eta}$ and for evaluating $f(\mathbf{V};\hat{\eta})$ are independent.*

$$\mathsf{E}_D[f(\mathbf{V};\hat{\eta})] \quad \mathsf{E}[f(\mathbf{V};\eta_0)] = O_P(N^{\ 1=2}) + \mathsf{E}[f(\mathbf{V};\hat{\eta}) \quad f(\mathbf{V};\eta_0)].$$

Proof. We first note that

$$\mathsf{E}_D[f(\mathbf{V};\hat{\eta})] \quad \mathsf{E}[f(\mathbf{V};\eta_0)] = \mathsf{E}_{D\ \ P}[f(\mathbf{V};\eta_0)] \quad \mathsf{E}_{D\ \ P}[f(\mathbf{V};\hat{\eta}) \quad f(\mathbf{V};\eta_0)] + \mathsf{E}[f(\mathbf{V};\hat{\eta}) \quad f(\mathbf{V};\eta_0)].$$

First, $\mathsf{E}_{D\ \ P}[f(\mathbf{V};\eta_0)] = O_P(N^{\ 1=2})$ by the classical central limit theorem. Second, $\mathsf{E}_{D\ \ P}[f(\mathbf{V};\hat{\eta}) \quad f(\mathbf{V};\eta_0)] = O_P(N^{\ 1=2})$ under given conditions by (Kennedy et al., 2020, Lemma 2). □

Now, we introduce an equivalent representation of the *do*-Shapley:

Proposition S.2 ((Štrumbelj & Kononenko, 2014, Eq. (10))). *An equivalent representation of the do-Shapley in Eq. (2) is given as*

$$\tilde{\phi}_{v_i} \coloneqq \frac{1}{n!}\sum_{2\ ([n])}\left\{\mathsf{E}[Y|do(\mathbf{v}_{pre_\pi(i);i})] \quad \mathsf{E}[Y|do(\mathbf{v}_{pre_\pi(i)})]\right\},$$

where $\Pi([n])$ is a set of all possible permutations of $[n]$, π is an individual permutation in $\Pi([n])$, pre $(i) \coloneqq \{k \in [n]$ such that $k < i$ in $\pi([n])\}$.

This representation motivates a following Monte-Carlo-based approximation:

$$\tilde{\phi}_{v_i} \coloneqq \frac{1}{M}\sum_{j=1}^{M}\left\{\mathsf{E}[Y|do(\mathbf{v}_{\text{pre}_{\pi_j}(i);i})] \quad \mathsf{E}[Y|do(\mathbf{v}_{\text{pre}_{\pi_j}(i)})]\right\}, \tag{D.5}$$

where M is the number of randomly generated permutation of $[n]$ and π_j denotes kth permutation. Convergence of $\tilde{\phi}_{v_i}$ is guaranteed by the following result:

Lemma S.5.

$$\tilde{\phi}^{v_i} \quad \phi_{v_i} = O_P(M \quad ^{1=2}). \tag{D.6}$$

Proof. Let $Z(\sigma) \coloneqq \mathsf{E}[Yjdo(\mathbf{v}_{i;\mathrm{pre}_{\sigma(k)}(i)})] \quad \mathsf{E}[Yjdo(\mathbf{v}_{\mathrm{pre}_{\sigma(k)}(i)})]$ denote a random variable where the randomness is over the permutation σ, where $P(\sigma) = \frac{1}{n!}$. Then, $\mathsf{E}_P[Z(\sigma)] = \phi_{v_i}$. By the given assumption, $Z(\sigma)$ and

$$\tilde{\phi}_{v_i} \coloneqq \frac{1}{M}\sum_{k=1}^{M} Z(\sigma(k))$$

are bounded random variables. Let B denote such bound. Then, by (Lattimore & Szepesvári, 2020, Corollary 5.5),

$$\tilde{\phi}_{v_i} > \phi_{v_i} \quad \sqrt{\frac{2B^2\log(1/\delta)}{M}} \text{ and } \tilde{\phi}_{v_i} < \phi_{v_i} + \sqrt{\frac{2B^2\log(1/\delta)}{M}}$$

in probability $(1 \quad \delta)$, which implies that $\tilde{\phi}_{v_i}$ converges in $^{p}\overline{M}$ rate. This completes the proof. □

Let $S_{j;a} \coloneqq \mathrm{pre}_{\ j}(i)$ and $S_{j;b} \coloneqq \mathit{fig}\ [\ \mathrm{pre}_{\ j}(i)$. By Def. 7, Eqs. (D.5,D.6),

$$\phi^{\mathsf{est}}_{v_i} \quad \phi_{v_i} = \phi^{\mathsf{est}}_{v_i} + \tilde{\phi}_{v_i} \quad \tilde{\phi}_{v_i} + \phi_{v_i} \tag{D.7}$$

$$= \frac{1}{M}\sum_{i=1}^{j}\left(\left\{T^{\mathsf{est}}(S_{j;b}) \quad \mathsf{E}[Yjdo(\mathbf{v}_{S_{j,b}})]\right\} + \left\{T^{\mathsf{est}}(S_{j;a}) \quad \mathsf{E}[Yjdo(\mathbf{v}_{S_{j,a}})]\right\}\right) + O_P(M \quad ^{1=2}). \tag{D.8}$$

Now, we analyze each of IPW, REG, DML estimators in Defs. (4,5,6).

Lemma S.6 (Error analysis for IPW). *For any nonempty* $S \quad [n]$,

$$T^{\mathsf{ipw}}(S) \quad \mathsf{E}[Yjdo(\mathbf{v}_S)] = \begin{cases} O_P(N \quad ^{1=2}) + O_P\left(\left\|\hat{\omega}^S_s \quad \omega^S_s\right\|\right) & \textit{(Markovian)} \\ O_P(N \quad ^{1=2}) + O_P\left(\left\|\hat{\omega}^S \quad \omega^S\right\|\right) & \textit{(Direct-cause)}, \end{cases} \tag{D.9}$$

Proof. We will prove only for the Markovian case, since the exactly same proof is applied for the Direct-cause case. First, $\mathsf{E}[Yjdo(\mathbf{v}_S)] = \mathsf{E}\left[Y\omega^S_s\right]$.

From Lemma S.4, it suffices to show that $\mathsf{E}\left[Y\hat{\omega}^S_s \quad Y\omega^S_s\right] = O_P\left(\left\|\hat{\omega}^S_s \quad \omega^S_s\right\|\right)$. It can be shown by

$$\mathsf{E}\left[Y\hat{\omega}^S_s \quad Y\omega^S_s\right] \quad \mathit{kYk}\left\|\hat{\omega}^S_s \quad \omega^S_s\right\| \ . \ \left\|\hat{\omega}^S_s \quad \omega^S_s\right\|,$$

where the first inequality by Cauchy-Schwarz inequality and the second by the boundness of Y. □

Lemma S.7 (Error analysis for REG). *For any nonempty* $S \quad [n]$,

$$T^{\mathsf{reg}}(S) \quad \mathsf{E}[Yjdo(\mathbf{v}_S)] = \begin{cases} O_P(N \quad ^{1=2}) + O_P\left(\left\|\hat{\theta}^S_{0;1} \quad \theta^S_{0;1}\right\|\right) & \textit{(Markovian)} \\ O_P(N \quad ^{1=2}) + O_P\left(\left\|\hat{\theta}^S_a \quad \theta^S\right\|\right) & \textit{(Direct-cause)}\ . \end{cases}$$

Proof. We will prove only for the Markovian case, since the exactly same proof is applied for the Direct-cause case. We note that $\mathsf{E}\left[\theta^S_{0;1}\right] = \mathsf{E}[Yjdo(\mathbf{v}_S)]$ by Lemma 2. From Lemma S.4, it suffices to show that $\mathsf{E}\left[\hat{\theta}^S_{0;1} \quad \theta^S_{0;1}\right] = O_P(\left\|\hat{\theta} \quad \theta\right\|)$. It holds by Cauchy-Schwarz inequality. □

Lemma S.8 (Error analysis for DML). *For any nonempty* $S \quad [n]$,

$$T^{\mathsf{dml}}(S) \quad \mathsf{E}[Yjdo(\mathbf{v}_S)] = \begin{cases} O_P(N \quad ^{1=2}) + \sum_{j=1}^{s} O_P\left(\left\|\hat{\theta}^S_{j;2} \quad \theta^S_{j;2}\right\| \left\|\hat{h}^S_j \quad h^S_j\right\|\right) & \textit{(Markovian)} \\ O_P(N \quad ^{1=2}) + O_P\left(\left\|\hat{\theta}^S_a \quad \theta^S_a\right\| \left\|\hat{h}_S \quad h_S\right\|\right) & \textit{(Direct-cause)}. \end{cases}$$

Proof. We note that $\mathsf{E}\left[V(\mathbf{V}^0, \eta^S)\right] = \mathsf{E}[Yjdo(\mathbf{v}_S)]$ by Lemma 3. From Lemma S.4, it suffices to show that

$$\mathsf{E}\left[V(\mathbf{V}^0, \hat{\eta}^S) \quad V(\mathbf{V}^0, \eta^S)\right] = \begin{cases} \sum_{j=1}^{s} O_P\left(\left\|\hat{\theta}^S_{j;2} \quad \theta^S_{j;2}\right\| \left\|\hat{h}^S_j \quad h^S_j\right\|\right) & \text{(Markovian)} \\ O_P\left(\left\|\hat{\theta}^S_a \quad \theta^S_a\right\| \left\|\hat{h}_S \quad h_S\right\|\right) & \text{(Direct-cause).} \end{cases}$$

First, consider the Markovian case. We omit the superscript S. Consider a following quantity: For $j = 1, 2, \quad , s$

$$Q_j \coloneqq \theta_{j \;\; 1;1} + \sum_{k=j}^{s} \omega_{j:k}(\theta_{k;1} \quad \theta_{k;2}),$$

where $\omega_{j:k} \coloneqq \prod_{r=j}^{k} \frac{\mathbb{1}_{v_{m_r}}(V_{m_r})}{h^S_r}$. Let $Q_{s+1} \coloneqq Y$ and $\omega_{j+1:} = 0$. We note that $Q_1 = V(\mathbf{V}^0; \eta^S)$, and $\mathsf{E}[Q_1] = \mathsf{E}[Yjdo(\mathbf{v}_S)]$. Also, the following holds, by the definition of $\theta_{k \;\; 1;1}, \theta_{k;2}$:

$$\mathsf{E}\left[\theta_{k \;\; 1;1}\right] = \mathsf{E}\left[\mathbb{1}_{v_{m_k}}(V_{m_k})\theta_{k;2}\right],$$
$$\mathsf{E}\left[\hat{\theta}_{k \;\; 1;1}\right] = \mathsf{E}\left[\mathbb{1}_{v_{m_k}}(V_{m_k})\hat{\theta}_{k;2}\right].$$

First, we note that Q_j can be written in a recursion as follow: For $j = 1, 2, \quad , s$

$$Q_j = \theta_{j \;\; 1;1} + \omega_{j:j}\left(Q_{j+1} \quad \theta_{j;2}\right).$$

To witness, consider the followings:

$$\begin{aligned} Q_j &= \theta_{j \;\; 1;1} + \omega_{j:j}(\theta_{j;1} \quad \theta_{j;2}) + \omega_{j:j+1}(\theta_{j+1;1} \quad \theta_{j+1;2}) + \omega_{j:j+2}(\theta_{j+2;1} \quad \theta_{j+2;2}) + \\ Q_{j+1} &= \theta_{j;1} + \omega_{j+1:j+1}(\theta_{j+1;1} \quad \theta_{j+1;2}) + \omega_{j+1:j+2}(\theta_{j+2;1} \quad \theta_{j+2;2}) + \\ \omega_{j:j}Q_{j+1} &= \omega_{j:j}\theta_{j;1} + \omega_{j:j+1}(\theta_{j+1;1} \quad \theta_{j+1;2}) + \omega_{j:j+2}(\theta_{j+2;1} \quad \theta_{j+2;2}) + \quad . \end{aligned}$$

Then,

$$\begin{aligned} Q_j &= \omega_{j:j}Q_{j+1} \quad \omega_{j:j}\theta_{j;1} + \theta_{j \;\; 1;1} + \omega_{j:j}(\theta_{j;1} \quad \theta_{j;2}) \\ &= \theta_{j \;\; 1;1} + \omega_{j:j}\left(Q_{j+1} \quad \theta_{j;2}\right). \end{aligned}$$

Finally, we will witness the following holds:

$$\mathsf{E}\left[\hat{Q}_j \quad Q_j\right] = \mathsf{E}\left[\hat{Q}_j \quad \theta_{j \;\; 1;1}\right] = \sum_{k=j}^{s} O_P\left(\left\|\theta_{k;2} \quad \hat{\theta}_{k;2}\right\| \left\|\hat{h}_k \quad h_k\right\|\right).$$

We will prove this by using an inductive hypothesis. First, at $j = s$,

$$\begin{aligned} \mathsf{E}\left[\hat{Q}_S \quad Q_S\right] = \mathsf{E}\left[\hat{Q}_S \quad \theta_{S \;\; 1;1}\right] &= \mathsf{E}\left[\hat{\theta}_{S \;\; 1;1} + \hat{\omega}_{S:S}(Y \quad \hat{\theta}_{S;2}) \quad \theta_{S \;\; 1;1}\right] \\ &= \mathsf{E}\left[\hat{\theta}_{S \;\; 1;1} + \frac{\mathbb{1}_{v_{m_S}}(V_{m_S})}{\hat{\pi}_S}(Y \quad \hat{\theta}_{S;2}) \quad \theta_{S \;\; 1;1}\right] \\ &= \mathsf{E}\left[\mathbb{1}_{v_{m_S}}(V_{m_S})(\hat{\theta}_{S;2} \quad \theta_{S;2}) + \frac{\mathbb{1}_{v_{m_S}}(V_{m_S})}{\hat{\pi}_S}(\theta_{S;2} \quad \hat{\theta}_{S;2})\right] \\ &= O_P\left(\left\|\theta_{S;2} \quad \hat{\theta}_{S;2}\right\| k\hat{\pi}_S \quad \pi_S k\right). \end{aligned}$$

For any $j = s \quad 1, \quad , 1,$

$$\begin{aligned}
\mathsf{E}\left[\hat{Q}_j \quad Q_j\right] = \mathsf{E}\left[\hat{Q}_j \quad \theta_j \;_{1:1}\right] &= \mathsf{E}\left[\hat{\theta}_j \;_{1:1} \quad \theta_j \;_{1:1} + \hat{\omega}_{j:j}\left(\hat{Q}_{j+1} \quad \hat{\theta}_{j;2}\right)\right] \\
&= \mathsf{E}\left[\hat{\theta}_j \;_{1:1} \quad \theta_j \;_{1:1} + \hat{\omega}_{j:j}\left(\hat{Q}_{j+1} \quad \theta_{j;1}\right) + \hat{\omega}_{j:j}\left(\theta_{j;1} \quad \hat{\theta}_{j;2}\right)\right] \\
&= \mathsf{E}\left[\hat{\omega}_{j:j}\left(\hat{Q}_{j+1} \quad \theta_{j;1}\right)\right] + \mathsf{E}\left[\mathbb{1}_{v_{m_j}}(V_{m_j})\left(\hat{\theta}_{j;2} \quad \theta_{j;2}\right) + \hat{\omega}_{j:j}\left(\theta_{j;1} \quad \hat{\theta}_{j;2}\right)\right] \\
&= \mathsf{E}\left[\hat{\omega}_{j:j}\left(\hat{Q}_{j+1} \quad \theta_{j;1}\right)\right] + \mathsf{E}\left[\frac{1}{\hat{P}(V_{m_j} \boldsymbol{j} \mathbf{W}_{m_j})}\left\{\theta_{j;2} \quad \hat{\theta}_{j;2}\right\}\left\{\hat{h}_j \quad h_j\right\}\right] \\
&\quad . \; \mathsf{E}\left[\left(\hat{Q}_{j+1} \quad \theta_{j;1}\right)\right] + \mathsf{E}\left[\left\{\theta_{j;2} \quad \hat{\theta}_{j;2}\right\}\left\{\hat{h}_j \quad h_j\right\}\right] \\
&\quad \mathsf{E}\left[\left(\hat{Q}_{j+1} \quad \theta_{j;1}\right)\right] + \left\|\theta_{j;2} \quad \hat{\theta}_{j;2}\right\|\left\|\hat{h}_j \quad h_j\right\|.
\end{aligned}$$

If we assume $\mathsf{E}\left[\hat{Q}_r \quad \theta_r \;_{1:1}\right] = \sum_{k=r}^{s} O_P\left(\left\|\theta_{k;2} \quad \hat{\theta}_{k;2}\right\|\left\|\hat{h}_k \quad h_k\right\|\right)$ for $r = j+1, \quad , s$, then it's easy to witness that it holds for $r = j$, too. Therefore, by an induction, the equality holds for all $r = 1, 2, \quad , 1$. This completes the proof for Markovian case.

For Direct-cause case,

$$\begin{aligned}
\mathsf{E}\left[V(\mathbf{V}^{\emptyset}; \eta^S) \quad V(\mathbf{V}; \hat{\eta}^S)\right] &= \mathsf{E}\left[\frac{\mathbb{1}_{\mathbf{v}_S}(\mathbf{V}_S)}{\hat{h}^S}(Y \quad \hat{\theta}_a) + \hat{\theta}_b \quad \theta_b\right] \\
&= \mathsf{E}\left[\frac{\mathbb{1}_{\mathbf{v}_S}(\mathbf{V}_S)}{\hat{h}^S}(\theta_a \quad \hat{\theta}_a) + \hat{\theta}_b \quad \theta_b\right] \\
&= \mathsf{E}\left[\frac{h^S}{\hat{h}^S}(\theta_b \quad \hat{\theta}_b) + \hat{\theta}_b \quad \theta_b\right] \\
&= \mathsf{E}\left[\frac{1}{\hat{h}^S}(\theta_b \quad \hat{\theta}_b)(h^S \quad \hat{h}^S)\right] \\
&\quad . \; \mathsf{E}\left[\frac{1}{\hat{h}^S}(\theta_b \quad \hat{\theta}_b)(h^S \quad \hat{h}^S)\right] \\
&\quad \left\|\theta_b \quad \hat{\theta}_b\right\|\left\|h^S \quad \hat{h}^S\right\|.
\end{aligned}$$

□

By combining Lemmas (S.4,S.5,S.6,S.7,S.8), we complete the proof of Theorem D.3. □

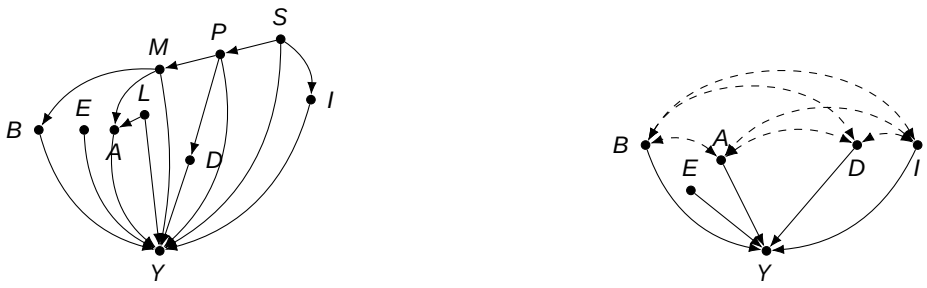
E. Additional Experimental Details From Section 6

E.1. Data Generating Processes

Here, we present the structural causal model for the data generating processes used for the data generating process used in Section 6.

We first note that $U \quad$ Bernoulli(0.4), $U_{V_1} \quad$ Bernoulli(0.8), $U_{V_3} \quad$ Bernoulli(0.4), $U_{V_2} \quad$ Bernoulli(0.3), and U_Y Normal$(0, 1)$. The SCM that induced the graph in Fig. 2a is

$$\begin{aligned}
V_1 &= U_{V_1} \quad U \\
V_3 &= U_{V_3} _ U \\
V_2 &= (V_1 \; \wedge \; V_3) _ U_{V_2} \\
Y &= 3V_1 + 0.4V_2 + V_3 + U_Y \,.
\end{aligned}$$



(a) (b)

Figure F.4: **(a)** Causal graph for Example 1, taken from Lundberg (2021) **(b)** Variables $\{S, P, M, L\}$ are hidden. These graphs are used for Appendix F.

The SCM that induced the graph in Fig. 2b is

$$\begin{aligned} V_1 &= U_{V_1} \\ V_3 &= U_{V_3} \\ V_2 &= (V_1 \wedge V_3) \_ U_{V_2} \\ Y &= 3V_1 + 0.4V_2 + V_3 + U_Y \,. \end{aligned}$$

The SCM that induced the graph in Fig. 2c is

$$\begin{aligned} V_1 &= U_{V_1} \quad U \\ V_3 &= U_{V_3} \_ U \\ V_2 &= U_{V_2} \\ Y &= 3V_1 + 0.4V_2 + V_3 + U_Y \,. \end{aligned}$$

F. Additional Experiments

In this section, we consider a different data generation process based on Example 1.

Experimental Setup. We use synthetic datasets based on: (a) Example 1 for which the corresponding causal graph Fig. F.4a is Markovian, and (b) the graph in Fig. F.4b which matches with Direct-cause case. These two graphs share the same data generating process since the graph in Fig. F.4b is generated from the graph in Fig. F.4a by omitting a set of variables. Details of the data generating process are provided in Appendix E. Throughout the simulation, we denote $\{\phi_{V_i}\}_{i=1}^{n}$ as the ground-truth *do*-Shapley values.

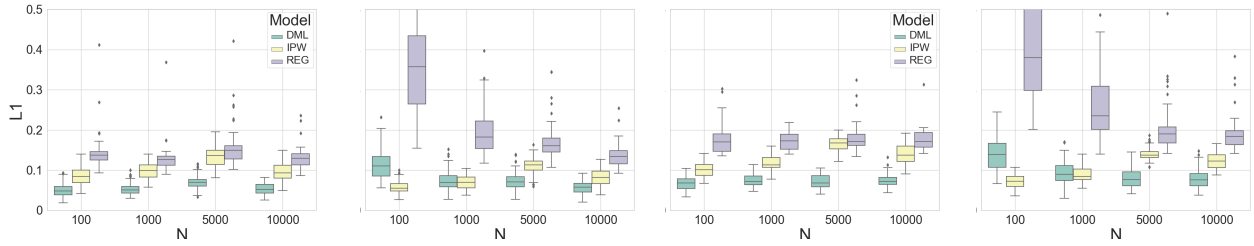
Comparison Between Estimators. We compare the three estimators (IPW, REG, DML), denoted by $\{\phi_{V_i}^{\text{ipw}}, \phi_{V_i}^{\text{reg}}, \phi_{V_i}^{\text{dml}}\}$ respectively, for scenarios depicted in graphs in Figs. (F.4a,F.4b). For all estimators, nuisances are estimated using gradient boosting model called XGBoost (Chen & Guestrin, 2016).

Let $\phi_{V_{i,k}}^{\text{est}} \in \{\phi_{V_{i,k}}^{\text{dml}}, \phi_{V_{i,k}}^{\text{ipw}}, \phi_{V_{i,k}}^{\text{reg}}\}$ denote an estimated importance of the ith feature of jth samples (i.e., $V_{i;k} \in \mathbf{V}_{(k)} \in D$). As in Section 6, we assess the quality of the estimator by computing the L_1 error as

$$L_1(\text{est}, k) \coloneqq (1/n) \sum_{i=1}^{n} \left| \phi_{V_i;k}^{\text{est}} \quad \phi_{V_i;k} \right| ,$$

(where n is the number of features). We ran the simulation for 100 randomly samples; i.e., $k \in \{1, 2, \ldots, 100\}$, and with sample size $N \coloneqq |D| \in \{100, 1000, 5000, 10000\}$ to observe convergence behaviors of estimators. We fix $M = 100$.

Data Generating Processes. Here, we present the structural causal model for the data generating processes used for the data generating process, where the qualitative graphical description is provided as causal graphs in Fig. F.4a. We will denote V_0 : sales calls , V_1 : interaction , V_2 : economic factors , V_3 : last upgrade , V_4 : product needs , V_5 : discounts provided , V_6 : monthly usage , V_7 : Ad spend , V_8 : bugs reported , Y : customer retention (target variable).



(a) Markovian, Non-noisy (b) Markovian, Noisy (c) Direct-cause, Non-noisy (d) Direct-cause, Noisy

Figure F.5: The L1-error plots for the scenario in Section F.

$V_0 \quad P(U_{V_0}), \quad V_2 \quad P(U_{V_2}) \quad V_3 \quad P(U_{V_3})$ where $P(U_{V_0})$ is a Uniform distribution ranging over 0 to 4; $P(U_{V_2})$ is a Uniform distribution ranging over 0 to 1; and $P(U_{V_3})$ is a Uniform distribution ranging over 0 to 20. For the rest of variables,

$$
\begin{aligned}
V_1 &= V_0 + U_{V_1} \\
V_4 &= 0.1 \quad V_0 + U_{V_4} \\
V_5 &= 0.5(1 \quad \text{logit}(V_4)) + 0.5 \quad U_{V_5} \\
V_6 &= \text{logit}\,(0.3 \quad V_4 + U_{V_6}) \\
V_7 &= V_6 \quad U_{V_6} + \mathbb{1}_{V_3<1}(V_3) + \mathbb{1}_{V_3<2}(V_3) \\
V_8 &= U_{V_8}(V_6),
\end{aligned}
$$

where U_{V_1} is a Poisson random variable with the parameter 0.2, $U_4 \quad \text{Normal}(0,1)$, U_{V_5} is a Uniform variable ranging $(0,1)$, $U_{V_6} \quad \text{normal}(0,1)$, U_{V_6} is an Uniform variable ranging $(0.9, 0.99)$, and $U_{V_8}(V_6)$ is a Poisson random variable such that its parameter follows $2V_6$, and $\mathbb{1}_{V_a<c}(V_a)$ is an indicator function for the variable V_a for the event $V_a < c$ for some constant c. Finally,

$$
Y^0 = 0.9V_4^0 + 0.8V_6^0 \quad 0.2V_2^0 + 0.05V_5^0 \quad 0.015(1 \quad V_8^0) + 0.2V_0^0 + 0.3V_1 + 0.5(V_3 + 0.25) + 0.6V_7 \quad U_Y \quad 0.45,
$$

where $\{V_4^0, V_6^0, V_2^0, V_5^0, V_8^0, V_0^0\}$ are random variables from the normal distribution where the variance is 1 and their means are $\{V_4, V_6, V_2, V_5, V_8, V_0\}$. Finally, $Y = \text{logit}(7Y)$.

For the Case 2, we drop the variable V_0, V_4, V_6, and we used

$$
Y^0 = \quad 0.2V_2^0 + 0.05V_5^0 \quad 0.015(1 \quad V_8^0) + 0.3V_1 + 0.5/(V_3/4 + 0.25) + 0.6V_7 \quad U_Y \quad 0.45.
$$

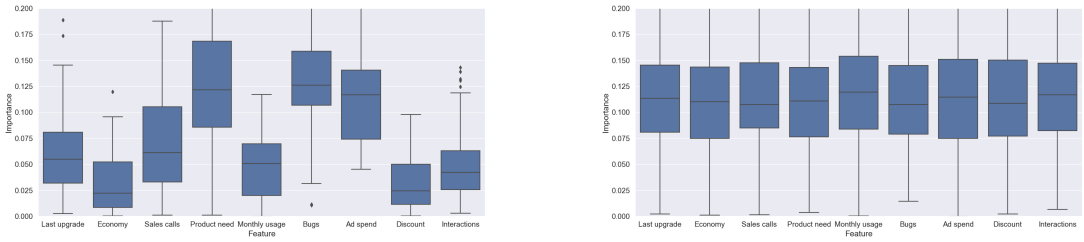
We also recommend checking the code `data_generator_1.py`, `data_generator_2.py` for the detailed configurations of the data generating processes.

Experimental Results. For the non-noisy setting, the L1-error plots for {Markovian, Direct-cause} cases are presented in Figs. (F.5a, F.5c) respectively. The DML-based estimator $\{\phi_{V_i}^{\text{dml}}\}_{i=1}^n$ outperforms $(\{\phi_{V_i}^{\text{ipw}}\}, \{\phi_{V_i}^{\text{reg}}\}_{i=1}^n)$ for all $N \in \{100, 1000, 5000, 10000\}$, and it achieves the smallest variance compared to other estimators. This result corroborates with the robustness property of the DML-based estimator (see Remark 4). The L1-error plots for the noisy setting for {Markovian, Direct-cause} cases are presented in Figs. (F.5b, F.5d) respectively. In this case, the DML-based estimator $\{\phi_{V_i}^{\text{dml}}\}_{i=1}^n$ exhibits the debiasedness property against the converging noise, while other estimators converge much slower.

Contrasting with the ICC Approach (Janzing et al., 2020a). We contrast the *do*-Shapley with the ICC approach (Janzing et al., 2020a). The *do*-Shapley measures the feature importance based on the total effect of variables, while the ICC measures based on their *intrinsic* effects. It is not possible to quantitatively compare these two contrasting definitions.

We compute the feature importance as proposed in (Molnar, 2020, Chap. 9.6.5), where the importance of the jth feature is defined as:

$$
I_j := \frac{1}{|D^0|} \sum_{\mathbf{V}_{(j)} \in D^0} \left| \phi_i(\mathbf{V}_{(j)}) \right|,
$$



(a) *do*-DML-Shapley (b) ICC

Figure F.6: Feature importance plots for the *do*-DML-Shapley and the ICC approaches.

	S	P	I	M	D	L	E	A	B
do-DML	0.07	0.12	0.05	0.05	0.03	0.06	0.03	0.12	0.13
ICC	0.12	0.13	0.12	0.13	0.13	0.12	0.12	0.12	0.12

Table 3: Average of feature importances produced by the *do*-DML-Shapley and ICC approaches.

where ϕ_i is the Shapley value, and D^0 D is a subset of samples.

In our experiments, we randomly selected 100 samples and compare the feature importance using the *do*-DML-Shapley ($\phi_{V_i}^{\text{dml}}$) and the ICC approach, denoted $\phi_{V_i}^{\text{icc}}$. The average of the estimated importance of each features described in Example 1 is presented in Table 3. In Fig. F.6, we present the bar-plot for both the *do*-DML-Shapley and the ICC approaches using the observations $f\left|\phi_i^{\text{dml}}(\mathbf{V}_{(j)})\right|g_{\mathbf{V}_{(j)}2D^0}$ and $f\left|\phi_i^{icc}(\mathbf{V}_{(j)})\right|g_{\mathbf{V}_{(j)}2D^0}$.

In our experiments, the *do*-Shapley approach gives that the production needs (P) has the largest total effect where P is in fact the variable with largest coefficient (0.9). in our data generating process, whereas ICC approach gives that all variables have similar intrinsic effects.