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 queries like “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 of properties (axioms) that causal contribution measures should satisfy and propose the do-Shapley values (inspired by do-interventions (Pearl, 2000)) as a unique 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 provide do-Shapley estimators exhibiting consistency, computational feasibility, and statistical robustness. Simulation results corroborate with the theory.

1. Introduction

Inferring causal effects is a fundamental problem throughout the data sciences since it can answer queries like “what would be an expected outcome if inputs had been fixed to certain values?” There is a growing literature tackling this question in both understanding the conditions under which causal conclusions can be drawn from non-experimental data (causal effect identification) (Pearl, 1995; Tian & Pearl, 2003; Huang & Vaftorta, 2006; Shpitser & Pearl, 2006; Bareinboim & Pearl, 2016; Jaber et al., 2018; Lee et al., 2019; 2020; Lee & Bareinboim, 2020), and in estimating the identified causal functions using the data (causal effect estimation) (Jung et al., 2020; Bhattacharya et al., 2020; Jung et al., 2021a,b; Bhattacharyya et al., 2020; 2021). Beyond the task of causal inference, interpreting the result of causal inference, including “what is the most important cause of the effect?” or more generally, “what are the contributions of each cause to the effect?”, is also practically important. Answering these queries fall under the task of measuring causal contributions, which quantifies the degree of contribution of causes to a target effect. As a motivation, consider the following scenario:

Example 1. A video streaming service company has collected data that contains various features (as described in (Lundberg, 2021)) including sales call (S), product needs (P), interaction with customers (I), monthly usage (M), discounts provided (D), last upgrade (U), economic factors (E), Ad spend (A), bugs reported (B). These features are causally related and affect the target variable: customer retention (Y) (see Fig. 1a). The company aims to measure 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 output since the output is the quantity derived by fixing inputs to specific values, which details are described in Remark 1.

In the area of explainable AI (XAI), there has been a recent thrust on measuring the contributions of features to the ML output (e.g., (Lundberg & Lee, 2017; Schwab & Karlen, 2019; Janzing et al., 2020b; Heskes et al., 2020; Covert et al., 2021)). The majority of existing methods have focused on queries where the target quantity is induced from an accessible model (we say a model for target Y is accessible if the model can be evaluated to obtain Y value for arbitrary input features), with little attention paid to the cases where the target is induced by nature (i.e., the data-generating process is inaccessible; e.g., the customer retention in Example 1) or the ML models are inaccessible. Also, many existing techniques are based on correlation (rather than causation) between the features and the ML model output (e.g., (Lundberg & Lee, 2017; Frye et al., 2020)). Even if another thread of methods focused on measuring contributions based on causation (e.g., (Schwab & Karlen, 2019; Janzing et al., 2020b; Heskes et al., 2020)), these methods often assume
that the data generating process for the target is known and accessible, allowing that an outcome corresponding to any arbitrary features can be generated, ruling out scenarios where the target quantity is induced from an inaccessible model (e.g., Example 1). A detailed comparison with 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. Our proposed method is directly applicable to the task of quantifying causal contributions of input features of an ML model prediction (formalized in Remark 1). Our key contributions, in further detail, are as follows:

1. [Sec. 3] We axiomatize causal contribution measures. Specifically, we propose desiderata for causal contribution measures as axioms, and propose the do-Shapley, the Shapley value (Shapley, 1953)-based method specialized for quantifying the causal contribution by leveraging the do-intervention (Pearl, 2000). Our axiomatic characterization provides a theoretical advocation in using the do-Shapley for quantifying causal contributions.

2. [Sec. 4] We provide conditions under which the do-Shapley values can be inferred from the observational data (identifiability) in polynomial time (computational feasibility). Even if verifying the identifiability can be done through existing theories in causal-effect identification, they do not provide computational feasibility in determining the identifiability of the do-Shapley. To address this, we provide a sufficient condition under which the identifiability and computational feasibility of the do-Shapley can be efficiently determined.

3. [Sec. 5] We construct estimators for the do-Shapley, exhibiting consistency, computational feasibility, 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). We prove that all estimators manifest consistency and computational feasibility. In addition, we show that the DML estimator additionally displays statistical robustness to model misspecification and bias. Finally, we present simulation results on these estimators that corroborate with the theory [Sec. 6].

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

---

1The 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.

---

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

2. Preliminaries

**Notation.** Each variable is represented with a capital letter (V) and its realized value with the small letter (v). We use bold letters V and v to denote a set of variables and a realized value of it, respectively. For any set S, we use |S| to denote its cardinality. Given a topological order \( < \) over the vertices \( V := \{V_1, \ldots, V_n\} \) of a graph \( G \), we will use \( \text{pre}(V_i) \) to denote the predecessors of \( V_i \). We use \( \text{pre}(v_i) \) as a realization of a set of variables \( \text{pre}(V_i) \); i.e., \( \text{pre}(v_i) = w_i \) for \( \text{pre}(V_i) = W_i \). 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 \( V_S := \{V_k\}_{k \in S} \) and \( V_{\overline{S}} := \{V_k\}_{k \in \overline{S}} \). We use \( D \) for the \( N \) samples from a distribution \( P \) over \( V \); i.e., \( D := \{V_{(i)}\}_{i=1}^N \sim P \), where \( V_{(i)} \) denotes the \( i \)th sample. For a function \( f \), we use \( \mathbb{E}[f(V)] \) as an expectation of \( f(V) \) over \( P \), and \( \mathbb{E}_D[f(V)] := (1/N) \sum_{i=1}^N f(V_{(i)}) \). We use \( \|f(V)\| := \mathbb{E}[\|f(V)\|^2] \) to denote the \( L_2(P) \) norm of \( f(V) \). \( O_P(\cdot) \) and \( o_P(\cdot) \) 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 \( M := \langle V, U, F, P(u) \rangle \) where \( V, U \) are a sets of endogenous (observables) and exogenous variables (latents), \( F \) is a set of functions \( f_{V_i} \) one for each \( V_i \in V \) where \( V_i \leftarrow f_{V_i}(PA_{V_i}, U_{V_i}) \) for some \( PA_{V_i} \subseteq V \) and \( U_{V_i} \subseteq U \), and \( P(u) \) is a strictly positive probability measure for \( U \). Each SCM \( M \) induces a semi-Markovian causal graph \( G \) over the node set \( V \) here \( V_i \rightarrow V_j \) if \( V_i \) is an argument of \( f_{V_j} \), and \( V_i \leftarrow V_j \) if \( U_{V_i} \) and \( U_{V_j} \) are correlated. Performing an intervention \( X = x \) is represented through the do-operator, \( do(X = x) \) (shortly, \( do(x) \)), which encodes the operation of replacing the original equations of \( X \) by the constant \( x \) in the SCM \( M \), inducing a submodel \( M_x \) and an interventional distribution \( P(V = v|do(x)) \) (shortly, \( P(v|do(x)) \)).
Causal Effect Identification. Given a causal graph $G$ over $V$, an effect $P(y|do(x))$ where $X, Y \subseteq V$ is identifiable if $P(y|do(x))$ is computable from the distribution $P(v)$ in any SCM $\mathcal{M}$ that induces $G$ (Pearl, 2000, p. 77). One key notion is called confounded components (in short, C-component): a set of nodes connected with a path composed solely of bi-directed edge $V_i \leftrightarrow V_j$ (Tian & Pearl, 2003). For any $\mathcal{C} \subseteq V$, the quantity $\mathcal{Q}(\mathcal{C}) := P(v|do(c))$, called a C-factor (Tian & Pearl, 2003), is defined as an interventional distribution of $\mathcal{C}$ under an intervention on $V \setminus \mathcal{C}$. We use $C(V_i)_G$ (shortly, $C(V_i)$) to denote $C$-component of $V_i$ in $G$, a set of variables belonging to the same $C$-component as in $V_i$. We use $C(W) := \bigcup_{V_i \in W} C(V_i)$ to denote a $C$-component of a set $W \subseteq V$.

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 function value $\nu(S)$ that measures the value of coalition of players $i \in S \subseteq [n]$. The Shapley value, given as

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

where $\omega(S) := \frac{(|S|-1)!}{(|S|-1)!}$, is a unique value 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 $D$ drawn from a distribution $P := P_M$ and a compatible semi-Markovian causal graph $G := G_M$, induced by the SCM $\mathcal{M}$, on topologically ordered variables $(V, Y)$, with $Y$ supposed to be the final variable in the order. We assume that $Y$ is bounded, and $V$ is a set of discrete variables. Given $(G, D, v)$ where $v$ is a realized value for $V$, our goal is to measure the the contribution of $v_i \in v$ to the target causal effect $\mathbb{E}[Y|do(v)]$ based on the impact to $Y$ if the SCM $\mathcal{M}$ has fixed the value of the variable as $V_i = v_i; P_M(y|do(v_i))$ where $P_M$ denote the distribution induced by $\mathcal{M}$. We set $\mathbb{E}[Y] = 0$ without loss of generality. We make no assumptions regarding the data generating process on $Y$ for generality. With the following additional assumption on $f$, our problem is straightforwardly reduced to the problem of attributing the importance of features in the XAI:

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

3. Axioms for Causal Contribution

We start by asking the question: “What makes a good causal contribution measure?” To answer, we propose the following desiderata:

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

1. Perfect assignment: Contributions are perfectly assigned; i.e., $\mathbb{E}[Y|do(v)] = \sum_{v_i \in v} \phi_{v_i}$.

2. Causal irrelevance: If $V_i$ is causally irrelevant to $Y$ for all witness $w \subseteq v \setminus \{v_i\}$ (i.e., $\forall y \in P(y|do(v_i, w)) = P(y|do(w))$), then $\phi_{v_i} = 0$.

3. Causal symmetry: If $\{v_i, v_j\} \in v$ have the same causal explanatory power to $Y$ for all witness $w \subseteq v \setminus \{v_i, v_j\}$ (i.e., $\forall y \in P(y|do(v_i, w), P(y|do(v_j, w)))$, then $\phi_{v_i} = \phi_{v_j}$.

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

The rationale behind Axiom 1 is the following: (1) Perfect assignment is a natural requirement since we aim to attribute the degree of contributions of each feature $v_i \in v$ 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 $\phi_{v_i}$ to be interpreted as a proxy for the causal effect s.t. $\sum_{v_S \in v} \phi_{v_i} \approx \mathbb{E}[Y|do(v_S)]$ for any $S \subseteq [n]$.

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

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

---

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

\[ \phi_{vi} := \sum_{S \subseteq [n] \setminus \{i\}} \omega(S) \{E[Y|do(v_{S,i})] - E[Y|do(v_{S})]\}, \quad (2) \]

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

**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 the axiom doesn’t restrict the value function to any fixed form. Thm. 1 instead characterizes the do-Shapley as the unique causal contribution measures satisfying Axiom 1 among any arbitrary value functions and corresponding contribution measures, as in (Sundararajan & Najmi, 2020).

The do-Shapley, as the name implies, is a specialization of the Shapley value in Eq. (1) for \( \nu(S) = E[Y|do(v_{S})] \). The do-Shapley can be alternatively viewed as a marginal causal effect of \( v_i \in \nu \) (i.e., \( E[Y|do(v_{S,i})] - E[Y|do(v_{S})] \)) 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 \( v_i \in \nu \). In practice, assigning contributions exclusively to a subset \( \nu \subseteq \nu \) may engender more interpretable result. For example, when \( \nu := \nu(Y) \subseteq \nu \), assigning contributions only to the features in \( \nu \subseteq \nu \) might be more interpretable if it is needed that features indirectly affecting to the outcome should be assigned zero contributions. Enforcing the do-Shapley to assign contributions only for the subset \( \nu \) can be simply done (without loss of generality) by the following procedure: (1) Derive a causal graph \( G[\nu] \) compatible with \( P(\nu) \) by applying the projection of a graph\(^6\); and (2) Compute the do-Shapley w.r.t. \( G[\nu] \). 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Causality</th>
<th>Inaccessibility</th>
<th>Axioms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>Marginal</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>Causal</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
</tr>
<tr>
<td>ICC</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>do-Shapley</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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 models (e.g., Example 1), and characterization by axioms among measures based on arbitrary value functions.

Shapley measures contributions based on association 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).

**Marginal Shapley.** The marginal Shapley is another widely used contribution measure in the XAI in which the target variable is a model prediction \( Y = f(V) \), where \( f \) is a deterministic (refer Remark 1) and accessible prediction model. The marginal Shapley is a specialization of the Shapley value with \( \nu(S) = E[f(v_{S}, V_{\overline{S}})] \) (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 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 semi-Markovian causal graphs (DAGs with bidirected edges. 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) \( \phi_{iv}^{ICC} \) 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 contri-
bution 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 $\mathbb{E}[Y|\text{do}(v_S)]$ as a functional of an observational distribution $P$ using $G$ is essential because we are only given non-experimental dataset $D$. For each $S \subseteq [n]$, complete causal effect identification algorithms for identifying $\mathbb{E}[Y|\text{do}(v_S)]$ 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 computational challenges 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 $\{\phi_v\}_{v \in V}$ w.r.t. $G$ are said to be **identifiable** if all elements in $\{\mathbb{E}[Y|\text{do}(v_S)]\}_{S \subseteq [n]}$ are identifiable in the causal graph $G$. The identifiability of the do-Shapley values are said to be (computationally-)**feasible** if the identification can be done in $O(\text{poly}(n))$.

Since naively 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 $X \subseteq V$, $\{X_k\}_{k=1}^c$ is said to be the $C$-partition if $X = \bigcup_{k=1}^c X_k$ (where $X_a \cap X_b = \emptyset$ for $a \neq b$) where $\forall k \in [c]$, $X_k$ is a set s.t. any two pairs $X_i, X_j \in X_k$ are in the same $C$-component in $G$.

**Theorem 2 (Identifiability & Feasibility of do-Shapley).**

The do-Shapley is identifiable if no variable in $V_i \in \{V\}$ 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 $S \subseteq [n]$, $\mathbb{E}[Y|\text{do}(v_S)] = \sum_{v_S} \mathbb{E}[Y|v] Q[V\setminus V_S]$, where $Q[V\setminus V_S] := Q[V\setminus V_S](v)$ is given as $Q[V\setminus V_S] = \frac{P(v)}{Q[C(V_S)]} \prod_{k=1}^c Q[C(S_k)]$, where $Q[C(V_S)] = \prod_{v \in C(S_k)} P(v|\text{pre}(v))$ is a $C$-factor of a $C$-component $V_S$ ($C(V_S)$): $\{S_k\}_{k=1}^c$ is a $C$-partition of $V_S$; and $Q[C(S_k)] := \prod_{v \in C(S_k)} P(v|\text{pre}(v))$ is a $C$-factor of a $C$-component $C(S_k)$ for $S_k$.

Fig. 2a provides an example graph satisfying the conditions in Thm. 2. Specifically, for all computations $V_S \subseteq V := \{V_1, V_2, V_3\}$, the causal effects are identified through Thm. 2 as $\mathbb{E}[Y|\text{do}(v_S)]$ $= \sum_{v_S} \mathbb{E}[Y|v] P(v_2|v_1, v_3)P(v_S)$, if $S \in \{1, 3\}$, $\sum_{v_S} \mathbb{E}[Y|v] P(v_2), \text{if } S \in \{0, 2, \{1, 2\}, \{2, 3\}\}$, $\sum_{v_S} \mathbb{E}[Y|v] P(v_S|V_S)$, if $S \in \{\{1\}, 3\}$, $\mathbb{E}[Y|v]$ if $S = \{1, 2, 3\}$.

Thm. 2 entails 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 $V_i \in V$. To demonstrate the wide applicability of Theorem 2, we provide two special cases which are commonly considered in the literature:

1. **Markovian case:** No latent confounders exist in the system; 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 $(V_i, V_j) \in V$ ($i \neq j$) is connected by a directed path, no $V_i$ are connected
to Y via bidirected edges, and no directed edge from Y to V; exists (i.e, only V; → 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, \( \mathbb{E}[Y|do(v_S)] \) is given as

\[
\mathbb{E}[Y|do(v_S)] = \sum_{v_S} \mathbb{E}[Y|\mathcal{S}, v_S] \prod_{i \in S} P(v_i|\text{pre}(v_i)).
\]

**Corollary 2 (Identification – Direct-cause).** In the Direct-cause case, \( \mathbb{E}[Y|do(v_S)] \) is given as

\[
\mathbb{E}[Y|do(v_S)] = \sum_{v_S} \mathbb{E}[Y|\mathcal{S}, v_S] P(v_S).
\]

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

\[
\mathbb{E}[Y|do(v_S)] = \sum_{v_S} \mathbb{E}[Y|\mathcal{S}, v_S] \prod_{i \in S} P(v_i|\text{pre}(v_i)),
\]

for all \( v_S \subseteq \{v_1, v_2, v_3\} \).

**5. Estimation of the do-Shapley**

Estimating the do-Shapley values in Eq. (2) is computationally and statistically challenging because (1) iterating over all \( S \subseteq [n] \) takes computation time exponential in \( n \), and (2) estimating \( \mathbb{E}[Y|do(v_S)] \) 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, because we are not aware of any general causal effect estimators suitable for estimating the expression in Thm. 2.\(^7\) Throughout this section, we assume all variables are discrete.

\(^7\)General causal effect estimators proposed by (Jung et al., 2021b) assumed that the expression is given by the identification algorithm. Therefore, the results in (Jung et al., 2021b) are not directly applicable to estimate the expression in Thm. 2.

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

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.

**Lemma 1 (Representation using IPW).** Let \( S = \{m_1, \cdots, m_s\} \subseteq [n] \) denote an index set for \( \mathcal{S}_S \). Let

\[
\omega_k^S := \sum_{r=1}^k \mathbb{I}_{v_{m_r}}(V_{m_r}) / h_r^S, \quad \text{for } k = s, \cdots, 1; \\
\omega^S := \mathbb{I}_{v_S}(V_S) / h^S,
\]

where \( h_r^S := P(V_{m_r}, v_{m_r}) \) and \( h^S := P(V_S|V_S) \).

Then, \( \mathbb{E}[Y|do(v_S)] = \mathbb{E}[Y|\omega] \) where \( \omega = \omega^S \) 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 \( \mathbb{E}[Y|do(v_S)] \)).** The IPW estimator \( T_{IPW}^S(S) \) for \( \mathbb{E}[Y|do(v_S)] \) is constructed as:

1. Split \( D \) randomly into two halves: \( D_0 \) and \( D_1 \);
2. Let \( \hat{\omega}_{S,p}^S \) denote estimators for \( \omega_S, \omega^S \) from \( D_p \subseteq \{D_0, D_1\} \), respectively.
3. For each \( p \in \{0, 1\} \), set

\[
T_{IPW}^S(S) := \left\{ \begin{array}{ll}
\mathbb{E}_{D_{1-p}}[Y|\hat{\omega}_{S,p}^S] & \text{(Markovian)} \\
\mathbb{E}_{D_{1-p}}[Y|\hat{\omega}_{p}^S] & \text{(Direct-cause)}
\end{array} \right.
\]

4. \( T_{IPW}^S(S) := \{T_{0 IPW}^S(S) + T_{1 IPW}^S(S)\} / 2 \).

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 $S := \{m_1, \ldots, m_s\} \subseteq [n]$ denote an index set for $V_S$. Let $\theta^S_{k,1} := Y$. For $k = s, s-1, \ldots, 1$,
\[
\begin{align*}
\theta^S_{k,2} & := \mathbb{E}[\theta^S_{k,1}(V_{m_k}, \text{pre}(V_{m_k}))] \\
\theta^S_{k-1,1} & := \mathbb{E}[\theta^S_{k,1}(V_{m_k}, \text{pre}(V_{m_k}))], \\
\theta^S_n & := \mathbb{E}[Y|V_S, V_{S}], \\
\theta^S_p & := \mathbb{E}[Y|V_S, \mathbb{V}_S].
\end{align*}
\]
Then, $\mathbb{E}[Y|\text{do}(v_S)] = \mathbb{E}[\theta]$ where $\theta = \theta^S_{n,1}$ for the Markovian case, and $\theta = \theta^S_n$ for the Direct-cause case.

We construct the REG estimator based on Lemma 2.

Definition 5 (REG for $\mathbb{E}[Y|\text{do}(v_S)]$). The REG estimator $T^{\text{reg}}(S)$ for $\mathbb{E}[Y|\text{do}(v_S)]$ 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 $D_p \in \{D_0, D_1\}$, respectively.
3. For each $p \in \{0, 1\}$,
\[
T^{\text{reg}}(S) := \begin{cases}
\mathbb{E}_{D_{1-p}}[\hat{\theta}^S_{0,1,p}] & \text{(Markovian)} \\
\mathbb{E}_{D_{1-p}}[\hat{\theta}^S_{a,p}] & \text{(Direct-cause)}.
\end{cases}
\]
4. $T^{\text{reg}}(S) := (T_0^{\text{reg}}(S) + T_1^{\text{reg}}(S))/2$.

For IPW and REG estimators to be consistent, one needs to estimate each individual functional (called nuisances) including $\mathbb{E}[Y|V_S, V_{S}]$ or $P(v_i|\text{pre}(v_i))$ consistently. A desirable robust estimator is one that converges to the ground-truth at a fast rate even when estimates for nuisances are mis-specified (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
\[
\eta^S := \left\{\theta^S_{0,1} \cup \{\theta^S_{k,1} \cup \{\theta^S_{a,1}, \mathbb{h}^S_a\} \cup \{h^S_r\}_{r=1}^{s}\} \right\} \quad \text{(Markovian)}
\]
\[
\{\theta^S_{a,1}, \mathbb{h}^S_a\} \quad \text{(Direct-cause),}
\]
defined in Defs. (4, 5) above, and
\[
V_S(V'; \eta^S) := \begin{cases}
\theta^S_{0,1} + \sum_{k=1}^{s} \omega^S_k (\theta^S_{k,1} - \theta^S_{k,2}) & \text{(Markovian)} \\
\theta^S_a + \omega^S (Y - \theta^S) & \text{(Direct-cause),}
\end{cases}
\]
where $\omega^S_k := \prod_{l=1}^{k} \mathbb{1}_{V_{m_l}}(V_{m_l})/h^S$ and $\omega^S := \mathbb{1}_{V_S}(V_S)/h^S$. Then, $\mathbb{E}[Y|\text{do}(v_S)] = \mathbb{E}[V_S(V'; \eta^S)].$

We construct the DML estimators based on Lemma 3.

Definition 6 (DML for $\mathbb{E}[Y|\text{do}(v_S)]$). The DML estimator $T^{\text{dml}}(S)$ is constructed as:

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

1. Input: $M$, Estimators $T^{\text{est}}(\cdot)$ in Defs. (4,5,6).
2. Output: Estimates $\{\hat{\phi}_{v_i}\}_{i=1}^{n}$.
3. Initialize $\hat{\phi}_{v_i} = 0$ for all $V_i \in V$.
4. for $j = 1$ to $M$ do
5. Generate the random permutation $\pi$ over $[n]$.
6. for $i = 1$ to $n$ do
7. $\hat{\phi}_{v_i} \leftarrow \hat{\phi}_{v_i} + T^{\text{est}}(\{i, \text{pre}_\pi(i)) - T^{\text{est}}(\text{pre}_\pi(i))$
8. end for
9. end for
10. return $\{\hat{\phi}_{v_i}/M\}_{i=1}^{n}$.

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 $T^{\text{est}}(S) \in \{T^{\text{ipw}}(S), T^{\text{reg}}(S), T^{\text{dml}}(S)\}$ denote an estimator for $\mathbb{E}[Y|\text{do}(v_S)]$ defined in Defs. (4,5,6), respectively. The do-Shapley estimator is given as
\[
\hat{\phi}^{\text{est}}_{v_i} := \frac{1}{M} \sum_{j=1}^{M} (T^{\text{est}}(\{i, \text{pre}_\pi(j)) - T^{\text{est}}(\text{pre}_\pi(i))$
\]
where $M$ is the number of randomly generated permutations of $[n], \pi_j$ denotes the $j$th permutation, and $\text{pre}_\pi(i)$ is the set of elements that precedes $i$ in $\pi_j$.

Equipped with the above results, 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 $\{v_i\}_{j=1}^{M}$ denote $M$ randomly generated permutations of $[n]$. For the fixed index $i$, let $S_{j,0} := \text{pre}_{\pi_{j}}(i)$ and $S_{j,1} := \{i\} \cup S_{j,0}$. Let $\{\hat{\eta}_{S_{j,0}}, \hat{\eta}_{S_{j,1}}\}_{j=1}^{M}$ 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 $\epsilon(\hat{g}) := \|\hat{g} - g\|$ denote an error for a nuisance estimates for any $\hat{g} \in \hat{g}$ and $g \in \eta$. For the do-Shapley estimators defined in Def. 7, suppose the estimators $T^{\text{est}}(S)$ are bounded. Let $\epsilon^{\text{est}}_{v_i} := \hat{\phi}^{\text{est}}_{v_i} - \phi_{v_i}$ (where est $\in \{\text{ipw, reg, dml}\}$).
Under the Markovian case,

\[ \epsilon_{\text{ipw}}^{\text{vi}} = R_{M,N} + O_P\left\{ \sum_{p \in \{0,1\}} \sum_{j=1}^{M} e(\hat{\omega}_{S,j}^{i,p}) \right\}, \]
\[ \epsilon_{\text{reg}}^{\text{vi}} = R_{M,N} + O_P\left\{ \sum_{p \in \{0,1\}} \sum_{j=1}^{M} e(\hat{\theta}_{S,j}^{i,p}) \right\}, \]
\[ \epsilon_{\text{dml}}^{\text{vi}} = R_{M,N} + O_P\left\{ \sum_{p \in \{0,1\}} \sum_{j=1}^{M} e(\hat{h}_{S,j}^{i,p}) e(\hat{\theta}_{S,j}^{i,p}) \right\}. \]

Under the Direct-cause case,

\[ \epsilon_{\text{ipw}}^{\text{vi}} = R_{M,N} + O_P\left\{ \sum_{p \in \{0,1\}} \sum_{j=1}^{M} e(\hat{\omega}_{S,j}^{i,p}) \right\}, \]
\[ \epsilon_{\text{reg}}^{\text{vi}} = R_{M,N} + O_P\left\{ \sum_{p \in \{0,1\}} \sum_{j=1}^{M} e(\hat{\theta}_{S,j}^{i,p}) \right\}, \]
\[ \epsilon_{\text{dml}}^{\text{vi}} = R_{M,N} + O_P\left\{ \sum_{p \in \{0,1\}} \sum_{j=1}^{M} e(\hat{h}_{S,j}^{i,p}) e(\hat{\theta}_{S,j}^{i,p}) \right\}. \]

Remark 4 (Properties of the Proposed Estimators). Error analyses in Thm. 3 exhibit consistency of IPW, REG, and DML estimators. Specifically, if nuisances are consistently estimated, \( \epsilon_{\text{ipw}}^{\text{vi}} = \epsilon_{\text{reg}}^{\text{vi}} = \epsilon_{\text{dml}}^{\text{vi}} = O_P(1) \), 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 \( \hat{\phi}_{\text{dml}}^{\text{vi}} \) converge to the true value if either \( e(\hat{h}_{S,j}^{i,p}) \) or \( e(\hat{h}_{S,j}^{i,p}) \) under Markovian, and either \( e(\hat{h}_{S,j}^{i,p}) \) or \( e(\hat{h}_{S,j}^{i,p}) \) under Direct-cause are accurate (doubly robustness). Also, \( \hat{\phi}_{\text{dml}}^{\text{vi}} \) converges at the root-\( N \) rate if all nuisances \( \hat{h}_{S,j}^{i,p}, \hat{h}_{S,j}^{i,p} \) under Markovian, and all nuisances \( \hat{h}_{S,j}^{i,p}, \hat{h}_{S,j}^{i,p} \) 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 is 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 \( \{ \phi_{\nu_i} \}_{i=1}^n \) as the ground-truth do-Shapley values.

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

Let \( \phi_{\nu_{i,k}}^{\text{est}} \in \{ \phi_{\nu_{i,k}}^{\text{dml}}, \phi_{\nu_{i,k}}^{\text{ipw}}, \phi_{\nu_{i,k}}^{\text{reg}} \} \) denote an estimated importance of the \( i \)th feature of \( j \)th samples (i.e., \( V_{i,k} \in \mathbb{V}(k) \in \mathcal{D} \)). We assess the quality of the estimator by computing the L1 error as \( L_1(\text{est}, k) := (1/n) \sum_{i=1}^n |\phi_{\nu_{i,k}}^{\text{est}} - \phi_{\nu_{i,k}}| \) (where \( n \) is the number of features). We ran the simulation for 50 randomly generated sets of samples; i.e., \( k \in \{1, 2, \cdots, 50\} \), and with sample size \( N := |\mathcal{D}| \in \{100, 250, 500, 750, 1000\} \) 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 \{Thm. 2, Markovian, Direct-cause\} 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’ is introduced at a \( N^{-\alpha} \) rate (i.e., \( \epsilon \sim \text{Normal}(N^{-\alpha}, N^{-2\alpha}) \)) 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 machine 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 of 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 with the robustness property of the DML (Remark 4). Also, the DML estimator exhibits the doubly robustness property: the estimator converges in both of 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 to -1 if the ranks are opposite. The true data generating function is \( Y = 3V_1 + 0.4V_2 + V_3 + U \) 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 can be noted, do-DML-Shapley identifies \( V_1 \) as the most important.

<table>
<thead>
<tr>
<th></th>
<th>Thm. 2</th>
<th>Markovian</th>
<th>Direct</th>
</tr>
</thead>
<tbody>
<tr>
<td>DML</td>
<td>1.0</td>
<td>0.8</td>
<td>0.93</td>
</tr>
<tr>
<td>Conditional</td>
<td>-0.28</td>
<td>-0.74</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Table 2: Comparison of the rank correlation.

**7. Conclusion**

We propose the do-Shapley as a causal contribution measure and provide theoretical justification through the axiomatic characterization (Thm. 1). Next, we provide conditions under which do-Shapley values can be inferred from non-experimental data in polynomial time (Thm. 2). We then propose three do-Shapley estimators (IPW, REG, DML) that are consistent. We show that the DML estimator has additional robustness property called doubly robustness and debiasness (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 grants from NSF IIS-1750807 (CA-REER).

**References**


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 \( [n] = \{1, 2, \cdots, n\} \) and the value function of \( S \), denoted \( \nu(S) \), the Shapley value of the player \( i \), denoted \( \phi_i = \phi_i(\nu) \), equals

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

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 \( \nu_1 \) and \( \nu_2 \), \( \phi_i(\nu_1 + \nu_2) = \phi_i(\nu_1) + \phi_i(\nu_2) \).

B. Details on Remark 3

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

1. Construct a graph \( G[X] \) composed of nodes in \( X \) and edges added as follows (Tian & Pearl, 2003).
   
   (a) add a directed edge \( V_i \rightarrow V_j \) in \( G[C] \) if there exists a directed path from \( V_i \) to \( V_j \) in \( G \) such that every vertex on the path is not in \( C \);
   
   (b) add a bidirected edge \( V_i \leftrightarrow V_j \) in \( G[C] \) if there exists a divergent path between \( V_i \) and \( V_j \) in \( G \) such that every vertex on the path is not in \( C \).

2. Construct the do-Shapley w.r.t. \( \{y, x\} \) on \( G[X] \). Specifically, for all \( x_i \in x \)

\[
\phi_{x_i} := \sum_{x_S \subseteq X \setminus x_i} \omega^X(S) \{\mathbb{E}[Y\mid do(x_S, i)] - \mathbb{E}[Y\mid do(x_S)]\}, \tag{B.1}
\]

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

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

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

**Proof.** It suffices to show that \( G[X] \) is a graph corresponding to \( P(X) \), because of \( \phi_{x_i} \) is the do-Shapley value defined on a graph corresponding to \( P(X) \). By (Koster et al., 2002), \( G[X] \) is a graph corresponding to \( P(X) \). \( \square \)
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 $G = \{V_1 \leftrightarrow V_2 \rightarrow Y\}$ 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 $v_1, v_2, E[Y|v_1] = E[Y] = v_1 - 1/2 \neq 0$, which leads that $\phi^n_{\text{cond}} \neq 0$. In contrast, $E[Y|do(v_1)] - E[Y] = E[Y|do(v_1, v_2)] - E[Y|do(v_2)] = 0$. Therefore, $\phi_{v_1} = 0$, implying that do-Shapley satisfies the causal irrelevance property, unlike 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 $M$: 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, $V_1 \leftarrow f_{V_1}(U_{V_1}) = U_{V_1}; V_2 \leftarrow f_{V_2}(V_1, U_{V_2}) = V_1 \vee U_{V_2};$ and $Y \leftarrow f_Y(V_2, U_Y) = V_2 \oplus U_Y$. A corresponding causal diagram is $G = \{V_1 \rightarrow V_2 \rightarrow Y, U_{V} \rightarrow V \forall V \in \{V_1, V_2, Y\}\}$. 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 $v_1 = 1$ and $v_2 = 0$, which makes $v_1 = v_2 = 1$. Let $\phi_{\text{ICC}}$ denote the ICC of $v_1$. Then, $\phi_{\text{ICC}} = 0.225$ and $\phi_{\text{ICC}} = 0.075$ 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

$$\nu_{do}(S) := E[Y|do(S)]$$

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 $[n]$, $\pi$ is an individual permutation in $\Pi([n])$, and $\text{pre}_\pi(i) := \{k \in [n] \text{ such that } k < i \in \pi([n])\}$. Then,

$$\sum_{i=1}^{n} \phi_{v_i}(\nu_{do}) = \frac{1}{n!} \sum_{\pi \in \Pi([n])} \sum_{i=1}^{n} \{\nu_{do} (\{i\} \cup \text{pre}_\pi(i)) - \nu_{do} (\text{pre}_\pi(i))\} = \frac{1}{n!} \sum_{\pi \in \Pi([n])} \{\nu_{do} ([n]) - \nu_{do} (\emptyset)\} = \nu_{do} ([n]) - \nu_{do} (\emptyset) = E[Y|do(S)] - E[Y] = E[Y|do(S)].$$
Now we consider the causal irrelevance property. Suppose \( V_i \) is causally irrelevant to \( Y \) in expectation for all witness \( w \subseteq \{v_j\} \). Then, the equality \( \nu_{do}(S \cup i) - \nu_{do}(S) = 0 \) holds immediately for all \( S \subseteq [n] \setminus \{i\} \).

Next we consider the causal symmetry property. Suppose \( v_i, v_j \) has the same causal explanatory power w.r.t. any witnesses \( w \subseteq \{v_i, v_j\} \). This leads \( \nu_{do}(\{i\} \cup S) = \nu_{do}(\{j\} \cup S) \) for any \( S \subseteq [n] \setminus \{i, j\} \). Then,

\[
\phi_{v_i}(\nu_{do}) = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \left( \frac{n-1}{|S|} \right) \left\{ \nu_{do}(S \cup i) - \nu_{do}(S) \right\}
= \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i, j\}} \left( \frac{n-1}{|S|} \right) \left\{ \nu_{do}(S \cup j) - \nu_{do}(S \cup i) \right\} + \frac{1}{n} \sum_{S \subseteq \{i, j\}} \left( \frac{n-1}{|S| + 1} \right) \left\{ \nu_{do}(S \cup \{i, j\}) - \nu_{do}(S \cup \{j\}) \right\}
= \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i, j\}} \left( \frac{n-1}{|S|} \right) \left\{ \nu_{do}(S \cup j) - \nu_{do}(S) \right\} + \frac{1}{n} \sum_{S \subseteq \{i, j\}} \left( \frac{n-1}{|S| + 1} \right) \left\{ \nu_{do}(S \cup \{i, j\}) - \nu_{do}(S \cup \{i\}) \right\}
= \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \left( \frac{n-1}{|S|} \right) \left\{ \nu_{do}(S \cup j) - \nu_{do}(S) \right\}
= \phi_{v_j}(\nu_{do}),
\]

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: \( \arg \min_{(\phi\nu_i)_{i=1}^n} \sum_{S \subseteq [n]} (\nu(S) - \nu_{do}(S))^2 \omega(S) \) (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 \( \{\phi_{v_i}\}_{v_i \in V} \) satisfying Axiom 1 is the \( do \)-Shapley.

**Proof.** Throughout the proof, we will define a canonical SCM as follow: Let \( T \subseteq [n] \) denote any fixed index set. A SCM is called canonical for \( T \) if \( \mathbb{E}[Y \mid do(V_S = 1) = 1 \text{ if } T \subseteq S, \text{ and } 0 \text{ otherwise.} \) We use \( \nu_{do}(S) \) denote the causal coalition function induced by the canonical SCM. Note \( \nu_{do}(S) = 1 \text{ if } T \subseteq S, \text{ and } 0 \text{ otherwise, by the definition of the canonical SCM.} \)

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

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

for some constants \( \{a^i_S\} \).

Now we focus on the causal irrelevance property. Suppose \( T \subseteq [n] \setminus \{i\} \). For any \( S \subseteq [n], \ T \subseteq S \implies (T \subseteq S \cup \{i\}) \). With \( i \notin T \), \( T \subseteq S \implies (T \subseteq S \cup \{i\}) \). Therefore, \( \nu_{do}(S) = \nu_{do}(S \cup \{i\}) \) for all \( S \subseteq [n] \). Then, by the causal irrelevance property, \( \phi_{v_i}(\nu_{do}^T) = 0 \) if \( T \subseteq [n] \setminus i \). Then, \( \phi_{v_i}(\nu_{do}^{[n] \setminus i}) = a^i_{[n] \setminus i} + a^i_{[n] \setminus i} = 0 \).

Suppose it has been shown that \( a^i_{T \cup \{i\}} + a^i_T = 0 \) for \( T \subseteq [n] \setminus i \) such that \( |T| \geq k \) for some \( k \). Then, for any \( S \subseteq [n] \setminus i \) such that \( |S| = k - 1 \),

\[
\phi_{v_i}(\nu_{do}^S) = \sum_{T \subseteq [n]} a^i_T \nu_{do}(T) = \sum_{T \subseteq [n]} a^i_T = \sum_{T \subseteq [n] \setminus \{i\}} (a^i_{T \cup \{i\}} + a^i_T)
= \left\{ \sum_{T \subseteq [n] \setminus \{i\}} (a^i_{T \cup \{i\}} + a^i_T) \right\} + (a^i_{[n] \setminus i} + a^i_{[n] \setminus i}) = a^i_{[n] \setminus i} + a^i_S,
\]
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 $S \subseteq [n]\{i\}$, by causal irrelevance property, $\phi_v(x_{do}) = 0$. This implies that $a^i_{S \setminus i} = a^i_S = 0$. Therefore, for any $T \subseteq [n]\{i\}$, $a^i_{T \setminus i} + a^i_T = 0$. 

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

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

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

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

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

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

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

Finally, we focus on the perfect assignment property. An attribution $\phi_{v_i}$ satisfies the perfect assignment property if and only if $\sum_{j \in V} p_{n-1} = 1$, and for any nonempty $T \subseteq [n]$, $\sum_{i \in T} p_{|T|-1} = \sum_{j \in T} p_{|T|}$ (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-\textit{Shapley} is identifiable if no variable in $V_i \in \{V\}$ 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 $S \subseteq [n]$, 

$$\mathbb{E}[Y|do(x_{do})] = \sum_{v \in U} \mathbb{E}[Y|v] Q[V \setminus V_S],$$

where $Q[V \setminus V_S] := Q[V \setminus V_S](v)$ is given as 

$$Q[V \setminus V_S] := \frac{P(v)}{Q[C(V_S)]} \prod_{k=1}^c \sum_{S_k} Q[C(S_k)],$$

where $Q[C(V_S)] = \prod_{v \in C(V_S)} P(v_{pre}(v_{do}))$ is a $C$-factor of a $C$-component $V_S$ ($C(V_S)$); $\{S_k\}_{k=1}^c$ is a $C$-partition of $V_S$; and $Q[C(S_k)] := \prod_{v \in C(S_k)} P(v_{pre}(v_{do}))$ is a $C$-factor of a $C$-component $C(S_k)$ for $S_k$. 

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

\[
    P(V|do(X)) = \frac{P(V)}{Q[C(X)]} \prod_{k=1}^{c} \sum_{x_k} Q[C(X_k)],
\]

where \( \{X_k\}_{k=1}^{c} \) is a C-partition of \( X \) in \( G \), and \( Q[C(X)] := \prod_{V_i \in C(X)} P(V_i|\text{pre}(V_i)) \) is a C-factor of a C-component of \( X \) (\( C(X) \)). Now, we show that \( V \) is identifiable and given as

\[
    P(V|do(X)) = \frac{P(V)}{Q[C(X_1)]} \sum_{x_1} Q[C(X_1)].
\]

Proposition 5.1 (Generalized Tian’s Adjustment – Complete identification criteria for \( P(V|do(X)) \)). \( P(V|do(X)) \) is identifiable if \( \forall X_a \in X \) and \( \text{Ch}(X_a) \) is not connected by bidirected paths. If identifiable, it’s given as

\[
    P(V|do(X)) = \frac{P(V)}{Q[C(X)]} \prod_{k=1}^{c} \sum_{x_k} Q[C(X_k)],
\]

where \( \{X_k\}_{k=1}^{c} \) is a C-partition of \( X \) in \( G \), and \( Q[C(X)] := \prod_{V_i \in C(X)} P(V_i|\text{pre}(V_i)) \) is a C-factor of a C-component of \( X \) (\( C(X) \)).

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

Suppose \( \forall X_a \in X \) is not connected with \( \text{Ch}(X_a) \) by bidirected paths. We first show that \( P(V|do(X_1)) \) (for any \( X_1 \in X \) a C-component in \( G(X) \)) is identifiable and given as

\[
    P(V|do(X_1)) = \frac{P(V)}{Q[C(X_1)]} \sum_{x_1} Q[C(X_1)].
\]

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

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

\[
    Q \left[ V \setminus X^{(i)} \right] = \frac{Q \left[ V \setminus X^{(i-1)} \right]}{Q[C(X_i)]} \sum_{x_i} Q[C(X_i)].
\]

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 \( X_a = \text{De}(X_i)_{G(C(X_i))} \forall V \setminus X^{(i-1)} \). To witness, consider the contradiction – for some \( X_a \in X_i \), there exists \( V_a \in \text{De}(X_i)_{G(C(X_i))} \) s.t. \( V_a \notin X_i \). First, \( V_a \) is connected with \( X_a \) by bidirected paths since \( V_a \in G(C(X_i)) \). Also, \( V_a \) is a descendent of \( X_a \), this means that a child of \( X_a \) is also in \( G(C(X_i)) \), connected by bidirected paths. This is a contradiction. Therefore, \( X_i = \text{De}(X_i)_{G(C(X_i))} \).

Now, we show that \( C(X_i)_{G(V \setminus X^{(i-1)})} = C(X_i)_G \). We start from an obvious observation – \( C(X_i)_{G(V \setminus X^{(i-1)})} \subset C(X_i)_G \).

We now prove \( C(X_i)_G \subset C(X_i)_{G(V \setminus X^{(i-1)})} \). For some \( V_a \in C(X_i)_G \), suppose \( V_a \notin C(X_i)_{G(V \setminus X^{(i-1)})} \). This means that bidirected paths connecting \( V_a \) to some nodes in \( X_1 \in X \) must be via other nodes in \( X_2 \in X \). This means that \( V_a \) and \( X_2 \) are connected by bidirected paths. However, given that \( X_2 \in X^{(i-1)} \) and \( X_1 \in X_i \), this is a contradiction, because they are in distinct C-partitions. Therefore, \( C(X_i)_{G(V \setminus X^{(i-1)})} = C(X_i)_G \).

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

\[
    Q \left[ V \setminus X^{(i)} \right] = \frac{P(V)}{\prod_{k=1}^{i} Q[C(X_k)]} \prod_{k=1}^{c} \sum_{x_k} Q[C(X_k)].
\]

We note \( Q[C(X^{(i)})] = \prod_{k=1}^{i} Q[C(X_k)] \), since

\[
    Q[C(X^{(i)})] = \prod_{V_i \in C(X^{(i)})} P(v_i|\text{pre}(v_i)) = \prod_{k=1}^{c} \prod_{V_i \in C(X_k)} P(v_i|\text{pre}(v_i)) = \prod_{k=1}^{c} Q[C(X_k)].
\]

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 \( S \subseteq [n] \),

\[
    (Y \perp \perp V_S|V_S^{(S')}_{G \setminus S}).
\]
Therefore,
\[ P(Y, V|do(V_S)) = P(Y|do(V_S), V_{\overline{S}})Q[V_S] = P(Y|V)Q[V_S], \]
which implies that
\[
E[Y|do(V_S)] = \sum_{V_{\overline{S}}} E[Y|V] \frac{Q[V]}{Q[C(V_S)]} \prod_{k=1}^{c} \sum_{s_k} Q[C(X_k)].
\]
This completes the proof. \(\square\)

**Corollary D.2** (Restated Corollary 1). *In the Markovian case*, \(E[Y|do(v_S)]\) is given as
\[
E[Y|do(v_S)] = \sum_{v_{\overline{S}}} E[Y|v_S, v_{\overline{S}}} \prod_{i \in S} P(v_i|\text{pre}(v_i)).
\]

**Proof.** In the Markovian case, \(C(W) = \emptyset\) for all \(W \subseteq V\). Then,
\[
P(Y, V_{\overline{S}}|do(V_S)) = \frac{P(V, Y)}{Q[C(V_S)]} \prod_{k=1}^{c} \sum_{s_k} Q[C(X_k)] = \frac{P(V, Y)}{Q[V_S]} = P(Y|V) \prod_{v_i \in V_{\overline{S}}} P(V_i|\text{pre}(V_i)).
\]
This completes the proof. \(\square\)

**Corollary D.2** (Restated Corollary 2). *In the Direct-cause case*, \(E[Y|do(v_S)]\) is given as
\[
E[Y|do(v_S)] = \sum_{v_{\overline{S}}} E[Y|v_S, v_{\overline{S}}} P(v_{\overline{S}}).
\]

**Proof.** In the Direct-cause case, \(Q[W] = P(W)\) for all \(W \subseteq V\) since there are no causal paths between a pair of variables in \(V\). Therefore, \(Q[V|V_S] = P(V|V_S) = P(V_{\overline{S}})\), which completes the proof. \(\square\)

### D.3. Proofs from Section 5

**Lemma D.1** (Restated Lemma 1). *Let \(S = \{m_1, \cdots, m_s\} \subseteq [n]\) denote an index set for \(V_S\). Let*
\[
\omega_k^S := \prod_{r=1}^{k} 1_{v_{m_r}}(V_{m_r})/h_r^S, \text{ for } k = s, \cdots, 1;
\]
\[
\omega^S := 1_{v_S}(V_S)/h^S,
\]
where \(h_r^S := P(V_{m_r}|\text{pre}(V_{m_r}))\) and \(h^S := P(V_S|V_{\overline{S}}).\) *Then, \(E[Y|do(v_S)] = E[Y|\omega]\) where \(\omega = \omega^S\) for the Markovian case, and \(\omega = \omega^S\) for the Direct-cause case.*

**Proof.** For the Markovian case,
\[
E[Y|do(v_S)] = \sum_{v_{\overline{S}}} \prod_{r=1}^{s} \frac{P(v)}{P(v_r|\text{pre}(v_{m_r}))} = E\left[ 1_{v_{m_r}}(V_{m_r})/h_r^S \right].
\]
For the Direct-cause case,
\[
E[Y|do(v_S)] = \sum_{v_{\overline{S}}} \frac{P(v)}{P(V_S|V_{\overline{S}})} = E\left[ 1_{v_S}(V_S)/P(V_S|V_{\overline{S}}) \right].
\]
\(\square\)
Lemma D.2 (Restated Lemma 2). Let \( S := \{m_1, \ldots, m_s\} \subseteq [n] \) denote an index set for \( V_S \). Let \( \theta^S_{s,1} := Y \). For \( k = s, s-1, \ldots, 1 \),

\[
\begin{align*}
\theta^S_{k,2} &:= \mathbb{E}[\theta^S_{k,1}|V_{m_k}, \text{pre}(V_{m_k})] \\
\theta^S_{k-1,1} &:= \mathbb{E}[\theta^S_{k,1}|\text{pre}(V_{m_k})],
\end{align*}
\]

Then, \( \mathbb{E}[Y|\text{do}(v_S)] = \mathbb{E}[\theta] \) where \( \theta = \theta^S_{0,1} \) for the Markovian case, and \( \theta = \theta^S_{0} \) for the Direct-cause case.

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

Lemma S.3. Suppose \( V' = \{Y\} \cup V \) where \( V' \) is an ordered set. Assume that \( Y \) is the last variable in the given order. Let \( V_S := \{V_{m_1}, \ldots, V_{m_s}\} \subseteq V \) (where \( \{m_1, \ldots, m_s\} \subseteq [n] \) denote a set of discrete variables. Let \( V_{\overline{S}} := V \setminus V_S \).

For each \( k = 2, \ldots, s \), let \( V_{t_k} := \{V_j \in V_{\overline{S}}: V_{m_{k-1}} \prec V_j \prec V_{m_k}\} \). Let \( V_{t_1} := \{V_j \in V_{\overline{S}}: V_j \prec V_{m_k}\} \) and \( V_{t_{s+1}} := \{V_j \in V_{\overline{S}}: V_m \prec V_j\} \).

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

\[
g_S(P) := \int_{V_{\overline{S}}} \mathbb{E}[Y|v] \prod_{i \in S} P(v_i|\text{pre}(v_i)) \, d[v_{\overline{S}}].
\]

Let \( \theta_{s,1} := Y \). For \( k = s, \ldots, 1 \), and

\[
\begin{align*}
\theta_{k,2} &:= \mathbb{E}[\theta_{k,1}|V_{m_k}, \text{pre}(V_{m_k})] \\
\theta_{k-1,1} &:= \mathbb{E}[\theta_{k,1}|\text{pre}(V_{m_k})],
\end{align*}
\]

Then, the following holds:

\[
g_S(P) = \mathbb{E}[\theta_{0,1}].
\]

Proof. Let

\[
\begin{align*}
A_k &:= \{\text{pre}(V_{m_k})\} \\
B_k &:= \{V_{t_{k+1}}, V_{t_{k+2}}, \ldots, V_{t_{s+1}}\} \\
C_k &:= \{V_{m_{k+1}}, V_{m_{k+2}}, \ldots, V_{m_s}\}.
\end{align*}
\]

For \( W \subseteq V \),

\[
q(W) := \begin{cases} 
\prod_{V_i \in W} P(v_i|\text{pre}(v_i)) & \text{if } W \neq \emptyset; \\
1 & \text{if } W = \emptyset.
\end{cases}
\]

Then, it suffices to show that

\[
\begin{align*}
\theta_{k,2} &= \int_{A_k, B_k} \mathbb{E}[Y|V_{m_k}, A_k, B_k, C_k] \, q(b_k) \, d[b_k, c_k] \\
\theta_{k-1,1} &= \int_{A_k, B_k} \mathbb{E}[Y|A_k, B_k, C_{k-1}] \, q(b_k) \, d[b_k, c_{k-1}],
\end{align*}
\]

because witnessing \( \mathbb{E}[\theta_{0,1}] = g_S(P) \) becomes trivial. Let \( \theta_{s,1} := Y \). 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,

\[
\begin{align*}
\theta_{k-2,1} &:= \mathbb{E}[\theta_{k-1,1}|V_{m_{k-1}}, \text{pre}(V_{m_{k-1}})] \\
\theta_{k-2,1} &:= \mathbb{E}[\theta_{k-1,1}|\text{pre}(V_{m_{k-1}})].
\end{align*}
\]
Then,
\[ \theta_{k-1,2} = E \left[ \int_{B_k, c_{k-1}} E \left[ Y | A_k, b_k, c_{k-1} \right] q(b_k) \Pi_{c_{k-1}}(C_{k-1}) d[b_k, c_{k-1}] \mid V_{m-1}, A_{k-1} \right] \]
\[ = \int_{B_{k-1}, c_{k-2}} E \left[ Y | V_{m-1}, A_{k-1}, b_{k-1}, c_{k-1} \right] q(b_{k-1}) \Pi_{c_{k-1}}(C_{k-1}) d[b_{k-1}, c_{k-1}] \]
and
\[ \theta_{k-2,1} = E \left[ \int_{B_k, c_{k-1}} E \left[ Y | A_k, b_k, c_{k-1} \right] q(b_k) \Pi_{c_{k-1}}(C_{k-1}) d[b_k, c_{k-1}] \mid v_{m-1}, A_{k-1} \right] \]
\[ = \int_{B_{k-1}, c_{k-2}} E \left[ Y | A_{k-1}, b_{k-1}, c_{k-2} \right] q(b_{k-1}) \Pi_{c_{k-2}}(C_{k-2}) d[b_{k-1}, c_{k-2}] \]

Therefore,
\[ \theta_{0,1} = \int_{B_1, c_0} E \left[ Y | A_1, b_1, c_0 \right] q(b_1) \Pi_{c_0}(C_0) d[b_1, c_0], \]
which gives the equality \( E[\theta_{0,1}] = g_S(P) \).

For the Direct-cause case,
\[ E[\theta^S_a] = \sum_{v_S} E[Y | v_S] P(v_S) = E[Y | do(v_S)], \]
which completes the proof.

**Lemma D.3** (Restated Lemma 3). Let
\[ \eta^s := \begin{cases} \{\theta^s_{0,1}\} \cup \{\theta^s_{k,1}, \theta^s_{k,2}\}_{k=1}^s \cup \{h_r^s\}_{r=1}^s \text{ (Markovian)} \\ \{\theta^S_a, \theta^S_b, h^S\} \text{ (Direct-cause)}, \end{cases} \]
defined in Defs. (4, 5) above, and
\[ V_S(V'; \eta^S) := \begin{cases} \theta^S_{0,1} + \sum_{k=1}^s \omega^S_k (\theta^S_{k,1} - \theta^S_{k,2}) \text{ (Markovian)} \\ \theta^S_a + \omega^S (Y - \theta^S_b) \text{ (Direct-cause)}, \end{cases} \]
where \( \omega^S_k := \prod_{r=1}^{k} \Pi_{v_m}(V_m) / h^S_r \) and \( \omega^S := \Pi_{v_S}(V_S) / h^S \). Then, \( E[Y | do(v_S)] = E[V_S(V'; \eta^S)] \).

**Proof.** For the Markovian case, it suffices to show that \( E[\theta^S_{k,1} - \theta^S_{k,2}] \) for any \( k = 1, 2, \cdots, s \). This holds since
\[ E[\theta^S_{k,1} - \theta^S_{k,2}] = E\left[ E[\theta^S_{k,1} - \theta^S_{k,2} | V_{m}, \text{pre}(V_{m})]\right] = E[\theta^S_{k,2} - \theta^S_{k,1}] = 0. \]
Therefore, \( E[V(V'; \eta^S)] = E[\theta^S_{0,1}] = E[Y | do(v_S)] \), where the 2nd equality holds by Lemma 2.

For the Direct-cause case, \( E[Y - \theta^S_a] = 0 \) by the definition of \( \theta^S_a \). Therefore, \( E[V(V'; \eta^S)] = E[\theta^S_b] = E[Y | do(v_S)] \).

**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 \( S_j,0 := \text{pre}_{\pi_i}(i) \) and \( S_j,1 := \{i\} \cup S_{j,0} \). Let \( \tilde{\eta}^{S,0}, \tilde{\eta}^{S,1,2} \) denote \( L_2 \)-consistent estimates for all nuisances \( \{\eta^{S,o}, \eta^{S,1,2}\}_{j=1}^M \) defined in Def. 6. Let \( R_{MN} := O_P(M^{-1/2} + N^{-1/2}) \). Let \( e(\hat{g}) := \|\hat{g} - g\| \) denote an error for a nuisance estimates for any \( g \in \tilde{\eta} \) and \( g \in \eta \). For the do-Shapley estimators defined in Def. 7, suppose the estimators \( T^{\text{rest}}(S) \) are bounded. Let \( \phi^{\text{rest}}_{\hat{v}_i} := \phi^{\text{rest}}_{\hat{v}_i} - \phi_{v_i} \) (where \( \text{est} \in \{ipw, \text{reg}, \text{dml}\} \)).
Under the Markovian case,

\[ e_{vi}^{\text{ipw}} = R_{M,N} + O_P\{ \sum_{p \in \{0,1\}} \sum_{j=1}^M e(\hat{\omega}_{S,j,p}^{S_j})\}, \]

\[ e_{vi}^{\text{reg}} = R_{M,N} + O_P\{ \sum_{p \in \{0,1\}} \sum_{j=1}^M e(\hat{\theta}_{0,1}^{S_j,p})\}, \]

\[ e_{vi}^{\text{dml}} = R_{M,N} + O_P\{ \sum_{p \in \{0,1\}} \sum_{j=1}^{s_j} e(\hat{h}_k^{S_j,p}) e(\hat{\theta}_k^{S_j,p})\}. \]

Under the Direct-cause case,

\[ e_{vi}^{\text{ipw}} = R_{M,N} + O_P\{ \sum_{p \in \{0,1\}} \sum_{j=1}^M e(\hat{\omega}_{S,j,p}^{S_j})\}, \]

\[ e_{vi}^{\text{reg}} = R_{M,N} + O_P\{ \sum_{p \in \{0,1\}} \sum_{j=1}^M e(\hat{\theta}_{2}^{S_j,p})\}, \]

\[ e_{vi}^{\text{dml}} = R_{M,N} + O_P\{ \sum_{p \in \{0,1\}} \sum_{j=1}^M e(\hat{h}_k^{S_j,p}) e(\hat{\theta}_k^{S_j,p})\}. \]

**Proof.** In the proof, we will use a notation \( E_{D-P} [f(V)] \) for \( f(V) := E_D [f(V)] - E [f(V)] \). We use \( N := |D| \). Also, for any quantity \( A, B, A \lesssim B \) if there is a constant \( c \) s.t. \( A \leq cB \). We first introduce a useful tool for analyzing errors of the proposed estimator.

**Lemma S.4.** Let \( \eta_0 \) denote some nuisance and \( \hat{\eta} \) denote its \( L_2 \) consistent estimate. Let \( f(V; \eta) \) denote an arbitrary function having a bounded second moment for any fixed \( \eta \). Suppose samples used for constructing \( \hat{\eta} \) and for evaluating \( f(V; \hat{\eta}) \) are independent.

\[ E_D [f(V; \hat{\eta})] - E [f(V; \eta_0)] = O_P(N^{-1/2}) + E [f(V; \hat{\eta}) - f(V; \eta_0)] . \]

**Proof.** We first note that

\[ E_D [f(V; \hat{\eta})] - E [f(V; \eta_0)] = E_{D-P} [f(V; \eta_0)] - E_{D-P} [f(V; \hat{\eta}) - f(V; \eta_0)] + E [f(V; \hat{\eta}) - f(V; \eta_0)] . \]

First, \( E_{D-P} [f(V; \eta_0)] = O_P(N^{-1/2}) \) by the classical central limit theorem. Second, \( E_{D-P} [f(V; \hat{\eta}) - f(V; \eta_0)] = O_P(N^{-1/2}) \) under given conditions by (Kennedy et al., 2020, Lemma 2).

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

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

\[ \tilde{\phi}_{vi} := \frac{1}{n!} \sum_{\pi \in \Pi([n])} \left\{ E[Y|do(v_{\pi(i)}, i)] - E[Y|do(v_{\pi_0(i)})] \right\} , \]

where \( \Pi([n]) \) is a set of all possible permutations of \([n]\), \( \pi \) is an individual permutation in \( \Pi([n]) \), \( \pi_0(i) := \{ k \in [n] \text{ such that } k < i \text{ in } \pi([n]) \} \).

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

\[ \tilde{\phi}_{vi} := \frac{1}{M} \sum_{j=1}^M \left\{ E[Y|do(v_{\pi_j(i)}, i)] - E[Y|do(v_{\pi_0(i)})] \right\} , \quad (D.5) \]

where \( M \) is the number of randomly generated permutation of \([n]\) and \( \pi_j \) denotes \( k \)th permutation. Convergence of \( \tilde{\phi}_{vi} \) is guaranteed by the following result:
Lemma S.5.

\[ \hat{\phi}_{vi} - \phi_{vi} = O_P(M^{-1/2}). \]  

**Proof.** Let \( Z(\sigma) := E[Y|do(v_i, pre_{\pi_\sigma(i)})] - E[Y|do(v_{pre_{\pi_\sigma(i)}})] \) denote a random variable where the randomness is over the permutation \( \sigma \), where \( P(\sigma) = \frac{1}{M} \). Then, \( E_P[Z(\sigma)] = \phi_{vi} \). By the given assumption, \( Z(\sigma) \) and are bounded random variables. Let \( B \) denote such bound. Then, by (Lattimore & Szepesvári, 2020, Corollary 5.5),

\[ \hat{\phi}_{vi} > \phi_{vi} - \sqrt{\frac{2B^2 \log(1/\delta)}{M}} \quad \text{and} \quad \hat{\phi}_{vi} < \phi_{vi} + \sqrt{\frac{2B^2 \log(1/\delta)}{M}} \]

in probability \((1 - \delta)\), which implies that \( \hat{\phi}_{vi} \) converges in \( \sqrt{M} \) rate. This completes the proof. \( \square \)

Let \( S_{j,a} := pre_{\pi_j}(i) \) and \( S_{j,b} := \{i\} \cup pre_{\pi_j}(i) \). By Def. 7, Eqs. (D.5,D.6),

\[ \phi_{vi}^{est} - \phi_{vi} = \hat{\phi}_{vi} - \hat{\phi}_{vi} + \phi_{vi} \]

\[ \quad = \frac{1}{M} \sum_{i=1}^j \left( \{T^{est}(S_{j,b}) - E[Y|do(v_{S_{j,b}})]\} + \{T^{est}(S_{j,a}) - E[Y|do(v_{S_{j,a}})]\} \right) + O_P(M^{-1/2}). \]  

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

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

\[ T^{ipw}(S) - E[Y|do(v_S)] = \begin{cases} O_P(N^{-1/2}) + O_P\left(\left\| \hat{\omega}_S - \omega_S \right\| \right) & \text{(Markovian)} \\ O_P(N^{-1/2}) + O_P\left(\left\| \hat{\omega}_S - \omega_S \right\| \right) & \text{(Direct-cause)} \end{cases}. \]  

**Proof.** We will prove only for the Markovian case, since the exactly same proof is applied for the Direct-cause case. First, \( E[Y|do(v_S)] = E[Y\omega_S] \). From Lemma S.4, it suffices to show that \( E[Y\hat{\omega}_S - Y\omega_S] = O_P\left(\left\| \hat{\omega}_S - \omega_S \right\| \right) \). It can be shown by

\[ E[Y\hat{\omega}_S - Y\omega_S] \leq \|Y\| \|\hat{\omega}_S - \omega_S\| \leq \|\omega_S - \omega_S\|, \]

where the first inequality by Cauchy-Schwarz inequality and the second by the boundness of \( Y \). \( \square \)

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

\[ T^{reg}(S) - E[Y|do(v_S)] = \begin{cases} O_P(N^{-1/2}) + O_P\left(\left\| \hat{\theta}_{0,1} - \theta_{0,1} \right\| \right) & \text{(Markovian)} \\ O_P(N^{-1/2}) + O_P\left(\left\| \hat{\theta}_a - \theta_a \right\| \right) & \text{(Direct-cause)} \end{cases}. \]  

**Proof.** We will prove only for the Markovian case, since the exactly same proof is applied for the Direct-cause case. We note that \( E[\hat{\theta}_{0,1}^S] = E[Y|do(v_S)] \) by Lemma 2. From Lemma S.4, it suffices to show that \( E[\hat{\theta}_{0,1}^S - \theta_{0,1}^S] = O_P(\|\hat{\theta} - \theta\|) \). It holds by Cauchy-Schwarz inequality. \( \square \)

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

\[ T^{dml}(S) - E[Y|do(v_S)] = \begin{cases} O_P(N^{-1/2}) + \sum_{j=1}^s O_P\left(\left\| \hat{\theta}_{j,2}^S - \theta_{j,2}^S \right\| \right) \|h_j^S - h_j^S\| & \text{(Markovian)} \\ O_P(N^{-1/2}) + O_P\left(\left\| \hat{\theta}_a^S - \theta_a^S \right\| \right) \|h_S - h_S\| & \text{(Direct-cause)} \end{cases}. \]
Proof. We note that $\mathbb{E} \left[ \mathcal{V}(\mathbf{V}', \eta^S) \right] = \mathbb{E}[Y|do(\mathbf{V}_S)]$ by Lemma 3. From Lemma S.4, it suffices to show that

$$
\mathbb{E} \left[ \mathcal{V}(\mathbf{V}', \eta^S) - \mathcal{V}(\mathbf{V}', \eta^S) \right] = \left\{ \begin{array}{ll}
\sum_{j=1}^s O_P \left( \left\| \hat{\theta}_{j,2}^S - \theta_{j,2}^S \right\| \left\| \hat{h}_j^S - h_j^S \right\| \right) & \text{(Markovian)} \\
O_P \left( \left\| \hat{\theta}_a^S - \theta_a^S \right\| \left\| \hat{h}_S - h_S \right\| \right) & \text{(Direct-cause)}
\end{array} \right.
$$

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

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

where $\omega_{j;k} := \prod_{r=j}^k \frac{1_{v_m}(V_m)}{h_r^S}$. Let $Q_{s+1} := Y$ and $\omega_{j+1} := 0$. We note that $Q_1 = \mathcal{V}(\mathbf{V}', \eta^S)$, and $\mathbb{E}[Q_1] = \mathbb{E}[Y|do(\mathbf{V}_S)]$. Also, the following holds, by the definition of $\theta_{k-1,1}, \theta_{k,2}$:

$$
\mathbb{E} \left[ \theta_{k-1,1} \right] = \mathbb{E} \left[ 1_{v_m_k}(V_{m_k}) \theta_{k,2} \right],
$$

$$
\mathbb{E} \left[ \hat{\theta}_{k-1,1} \right] = \mathbb{E} \left[ 1_{v_m_k}(V_{m_k}) \hat{\theta}_{k,2} \right].
$$

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

$$
Q_j = \theta_{j-1,1} + \omega_{j;j} (Q_{j+1} - \theta_{j,2}).
$$

To witness, consider the followings:

$$
Q_j = \theta_{j-1,1} + \omega_{j;j}(\theta_{j,1} - \theta_{j,2}) + \omega_{j;j+1}(\theta_{j+1,1} - \theta_{j+1,2}) + \omega_{j;j+2}(\theta_{j+2,1} - \theta_{j+2,2}) + \cdots
$$

$$
Q_{j+1} = \theta_{j,1} + \omega_{j+1;j+1}(\theta_{j+1,1} - \theta_{j+1,2}) + \omega_{j+1;j+2}(\theta_{j+2,1} - \theta_{j+2,2}) + \cdots
$$

$$
\omega_{j;j}Q_{j+1} = \omega_{j;j}\theta_{j,1} + \omega_{j;j+1}(\theta_{j+1,1} - \theta_{j+1,2}) + \omega_{j;j+2}(\theta_{j+2,1} - \theta_{j+2,2}) + \cdots.
$$

Then,

$$
Q_j = \omega_{j;j}Q_{j+1} - \omega_{j;j}\theta_{j,1} + \theta_{j-1,1} + \omega_{j;j}(\theta_{j,1} - \theta_{j,2})
$$

$$
= \theta_{j-1,1} + \omega_{j;j} (Q_{j+1} - \theta_{j,2}).
$$

Finally, we will witness the following holds:

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

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

$$
\mathbb{E} \left[ \hat{Q}_s - Q_s \right] = \mathbb{E} \left[ \hat{Q}_s - \theta_{s-1,1} \right] = \mathbb{E} \left[ \hat{\theta}_{s-1,1} + \hat{\omega}_{s:s}(Y - \hat{\theta}_{s,2}) - \theta_{s-1,1} \right]
$$

$$
= \mathbb{E} \left[ \hat{\theta}_{s-1,1} + 1_{v_{m_s}}(V_{m_s}) \frac{(Y - \hat{\theta}_{s,2}) - \theta_{s-1,1}}{\hat{\pi}_s} \right]
$$

$$
= \mathbb{E} \left[ 1_{v_{m_s}}(V_{m_s})(\theta_{s,2} - \theta_{s,2}) + 1_{v_{m_s}}(V_{m_s})(\hat{\theta}_{s,2} - \hat{\theta}_{s,2}) \right]
$$

$$
= O_P \left( \left\| \theta_{s,2} - \hat{\theta}_{s,2} \right\| \left\| \hat{\pi}_s - \pi_s \right\| \right).
For any \( j = s - 1, \cdots, 1 \),

\[
\mathbb{E} \left[ Q_j - Q_j \right] = \mathbb{E} \left[ \hat{Q}_j - \hat{\theta}_{j-1,1} \right] = \mathbb{E} \left[ \hat{\theta}_{j-1,1} - \theta_{j-1,1} + \hat{\omega}_{j,j} \left( Q_{j+1} - \hat{\theta}_{j,2} \right) \right]
\]

\[
= \mathbb{E} \left[ \hat{\theta}_{j-1,1} - \theta_{j-1,1} + \hat{\omega}_{j,j} \left( Q_{j+1} - \theta_{j,1} \right) + \hat{\omega}_{j,j} \left( \theta_{j,1} - \hat{\theta}_{j,2} \right) \right]
\]

\[
= \mathbb{E} \left[ \hat{\omega}_{j,j} \left( Q_{j+1} - \theta_{j,1} \right) \right] + \mathbb{E} \left[ \mathbb{I}_{V_{m_j}} \left( V_{m_j} \right) \left( \hat{\theta}_{j,2} - \theta_{j,2} \right) \right]
\]

\[
= \mathbb{E} \left[ \hat{\omega}_{j,j} \left( Q_{j+1} - \theta_{j,1} \right) \right] + \mathbb{E} \left[ \frac{1}{P(V_{m_j} | W_{m_j})} \left\{ \theta_{j,2} - \hat{\theta}_{j,2} \right\} \left\{ \hat{h}_j - h_j \right\} \right]
\]

\[
\leq \mathbb{E} \left[ \left( Q_{j+1} - \theta_{j,1} \right) \right] + \mathbb{E} \left[ \left\{ \theta_{j,2} - \hat{\theta}_{j,2} \right\} \right] \left\{ \hat{h}_j - h_j \right\}
\]

\[
\leq \mathbb{E} \left[ \left( Q_{j+1} - \theta_{j,1} \right) \right] + \left\| \theta_{j,2} - \hat{\theta}_{j,2} \right\| \left\| \hat{h}_j - h_j \right\|
\]

If we assume \( \mathbb{E} \left[ Q_r - \theta_{r-1,1} \right] = \sum_{k=r}^{s} O \left( \left\| \theta_{k,2} - \hat{\theta}_{k,2} \right\| \left\| \hat{h}_k - h_k \right\| \right) \) for \( r = j + 1, \cdots, s \), then it’s easy to witness that it holds for \( r = j \), too. Therefore, by an induction, the equality holds for all \( r = 1, 2, \cdots, 1 \). This completes the proof for Markovian case.

For Direct-cause case,

\[
\mathbb{E} \left[ \mathcal{V}(V'; \eta^S) - \mathcal{V}(V; \eta^S) \right] = \mathbb{E} \left[ \mathbb{I}_{V_{m_j}} \left( \mathcal{V}(V') \right) - \mathbb{I}_{V_{m_j}} \left( \mathcal{V}(V) \right) \right]
\]

\[
= \mathbb{E} \left[ \mathbb{I}_{V_{m_j}} \left( \mathcal{V}(V') \right) \right] - \mathbb{E} \left[ \mathbb{I}_{V_{m_j}} \left( \mathcal{V}(V) \right) \right]
\]

\[
= \mathbb{E} \left[ \mathcal{V}(V') - \mathcal{V}(V) \right]
\]

\[
\leq \left\| \mathcal{V}(V') - \mathcal{V}(V) \right\| = \left\| \theta_{j,2} - \hat{\theta}_{j,2} \right\| \left\| \hat{h}_j - h_j \right\|
\]

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

\[
V_1 = U_{V_1} \oplus U
\]

\[
V_3 = U_{V_3} \lor U
\]

\[
V_2 = (V_1 \land V_3) \lor U_{V_2}
\]

\[
Y = 3V_1 + 0.4V_2 + V_3 + U_Y.
\]
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 \( \{\phi_{ipw_{v_i}}, \phi_{reg_{v_i}}, \phi_{dml_{v_i}}\} \) 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 \( \phi_{est_{v_i,k}} \in \{\phi_{dml_{v_i}}, \phi_{ipw_{v_i}}, \phi_{reg_{v_i}}\} \) denote an estimated importance of the \( i \)th feature of \( j \)th samples (i.e., \( V_{i,k} \in V_{(k)} \in D \)). As in Section 6, we assess the quality of the estimator by computing the \( L_1 \) error as

\[
L_1(\text{est}, k) := (1/n) \sum_{i=1}^{n} |\phi_{est_{v_i,k}} - \phi_{v_i,k}|,
\]

(where \( n \) is the number of features). We ran the simulation for 100 randomly samples; i.e., \( k \in \{1, 2, \cdots, 100\} \), and with sample size \( N := |D| \in \{100, 1000, 5000, 10000\} \) 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).
We compute the feature importance as proposed in (Molnar, 2020, Chap. 9.6.5), where the importance of the th feature is defined as:

\[ I_j := \frac{1}{|\mathcal{D}|} \sum_{(V_{(j)}) \in \mathcal{D}} |\phi_j(V_{(j)})|, \]

where \( \mathcal{D} \) is the data set. We also recommend checking the code \texttt{data_generator_1.py}, \texttt{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 \( \{\phi_i^{\text{dml}}\}_{i=1}^n \) outperforms \( \{\phi_i^{\text{ipw}}, \phi_i^{\text{reg}}\}_{i=1}^n \) 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 \( \{\phi_i^{\text{dml}}\}_{i=1}^n \) 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:

\[ I_j := \frac{1}{|\mathcal{D}|} \sum_{V_{(j)} \in \mathcal{D}} |\phi_j(V_{(j)})|, \]
where $\phi_i$ is the Shapley value, and $D' \subseteq D$ is a subset of samples.

In our experiments, we randomly selected 100 samples and compare the feature importance using the $do$-DML-Shapley ($\phi_{dml}^i$) and the ICC approach, denoted $\phi_{icc}^i$. 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 $\{\phi_{dml}^i(V(j))\}_{V(j) \in D'}$ and $\{\phi_{icc}^i(V(j))\}_{V(j) \in D'}$.

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.

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

<table>
<thead>
<tr>
<th></th>
<th>S</th>
<th>P</th>
<th>I</th>
<th>M</th>
<th>D</th>
<th>L</th>
<th>E</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$do$-DML</td>
<td>0.07</td>
<td>0.12</td>
<td>0.05</td>
<td>0.05</td>
<td>0.03</td>
<td>0.06</td>
<td>0.03</td>
<td>0.12</td>
<td>0.13</td>
</tr>
<tr>
<td>ICC</td>
<td>0.12</td>
<td>0.13</td>
<td>0.12</td>
<td>0.13</td>
<td>0.13</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
</tr>
</tbody>
</table>

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