Incorporating Information into Shapley Values: Reweighting via a Maximum Entropy Approach

Anonymous Authors[1]

Abstract

In this article, we start by drawing a parallel between Shapley values as adopted in the area of eXplainable AI and some fundamental results from the area of graphical models. Specifically, we notice that both the marginal contributions needed for the computation of Shapley values and the graph produced by Pearl-Verma theorem rely on the choice of an ordering of the variables. For Shapley values, the marginal contributions are averaged over all orderings, while in causal inference methods/graphical models, the typical approach is to select orderings producing a graph with a minimal number of edges. We reconcile both approaches reinterpreting them from a maximum entropy perspective. Namely, Shapley values assume no prior knowledge about the orderings and treat them as equally likely. Conversely, causal inference approaches apply a form of Occam's razor and consider only orderings producing the simplest explanatory graphs. While Shapley values do not incorporate any available information about the model, we find that the blind application of Occam's razor also does not produce fully satisfactory explanations. Hence, we propose a variation of Shapley values based on entropy maximization to appropriately incorporate prior information about the model.

1. Introduction

Additive feature attribution algorithms are commonly used in eXplainable AI (XAI) to quantify how a model generates its output from the input features (Lundberg & Lee, 2017). However, assessing a feature's impact on a complex model is challenging due to interdependencies and non-linear interactions among features (Aas et al., 2021). As noted in (Lundberg & Lee, 2017; Štrumbelj & Kononenko, 2014; Datta et al., 2016), the problem is akin to cooperative games, where individual players contribute to the team's success, but how to isolate each player's contribution to the final outcome remains unclear. In game theory, Shapley values are a prevalent method to disentangle the individual contributions of each player to the team's payoff (Shapley, 1953). Specifically, Shapley values are given by the average marginal contribution of each player in every possible ordering of the players in the team. When addressing feature attribution, the authors in (Lundberg & Lee, 2017) quantified the contribution of each feature to the model's output by employing an analogy where the model output is viewed as the team's payoff, and each feature is considered as an individual player. Given a payoff function $v(S)$ defined for each subset of features $S \subseteq X = \{X_1, ..., X_N\}$ in a model, a possible way to compute the Shapley values for a specific instance of the features $X_1 = x_1$,...,$X_N = x_N$ is by using the following Algorithm 1.

Algorithm 1 Calculation of Shapley values

```
Require: Payoff $v$, instance of features $x = \{x_1, ..., x_N\}$
  for Every permutation $\pi$ of $\{1, ..., N\}$ do
    $z_i^\pi \leftarrow$ features preceding $x_i$ in the ordering $\pi$
    $\phi_i^\pi \leftarrow v\left(z_i^\pi \cup \{x_i\}\right) - v\left(z_i^\pi\right)$
  end for
  $\phi_i \leftarrow \frac{1}{N!} \sum_\pi \phi_i^\pi$
```

We refer to the quantity ϕ_i^π as the marginal contribution of the feature instance x_i in the ordering π whereas the Shapley value of the feature instance x_i is represented by the output ϕ_i in Algorithm 1, calculated as the average of these marginal contributions ϕ_i^π across all $N!$ possible orderings. While this algorithm offers the theoretical definition of Shapley values, it lacks efficiency and scalability for large feature sets, leading to various proposed approximation methods for practical estimation (Lundberg & Lee, 2017; Aas et al., 2021). Given this article's theoretical nature, our emphasis is on the theoretical definition of Shapley values, with no consideration for practical computations.

Given a model f, a common choice (Lundberg & Lee, 2017) for the payoff function is $v(S) = E[f(X)|S]$ which is the expectation of the output given the subset of features S. In essence, the computation of Shapley values involves considering the marginal contributions of each feature/player, which, though, are reliant on the ordering of the coalition. Assuming that there is no preferred natural ordering, Shapley values take the approach of averaging over all of them.

In addition to its simplicity, this approach has the appealing characteristic of satisfying several seemingly desirable properties that an additive feature attribution method should exhibit (Shapley, 1953). Indeed, Shapley values satisfy the following properties

- Efficiency: the contributions ϕ_i fully explain the outcome of an instance of features $X = x$ giving $\phi_0 + \sum_i \phi_i(x) = v(x)$, where $\phi_0 = E[f(X)]$.
- Missingness: If a feature has zero marginal contribution in each ordering, then its final contribution ϕ_i is zero $((\forall \pi)|\phi_i^{\pi}(x) = 0 \implies \phi_i(x) = 0)$.
- Symmetry: If for all orderings of the variables, swapping the variables X_i and X_j also swaps their marginal contributions, then $\phi_i(x) = \phi_j(x)$.
- Linearity/Additivity: Given two models f and g, the contribution of each feature x_i preserves the addition of the functions: $\phi_{i_{f+g}}(x) = \phi_{i_f}(x) + \phi_{i_g}(x)$, where $\phi_{i_f}(x)$ and $\phi_{i_g}(x)$ are the contributions of feature x_i to the output of model f and g respectively at the set point x.

Importantly, it can be proven that Shapley values are the only additive feature attribution method satisfying all these properties (Shapley, 1953; Lundberg & Lee, 2017). However, as mentioned in (Sundararajan & Najmi, 2020; Fryer et al., 2021), such uniqueness property is obviously subject to the choice of the payoff function. Indeed, different payoff functions $v(S)$ can lead to significantly different Shapley values. A considerable body of literature has investigated the implications of different choices for the payoff function (Janzing et al., 2020; Heskes et al., 2020; Giudici & Raffinetti, 2021; Ghalebikesabi et al., 2021; Yeh et al., 2022; Taufiq et al., 2023; Watson, 2022) and the type of explanations they lead to.

The work in (Janzing et al., 2020) proposes an alternative payoff function based on the expected value taken instead over the interventional distribution ($v(S) = E[f(X)|do(S)]$), arguing that in this way Shapley values' explanations are more aligned with a causal perspective. In a later collaborative work (Chen et al., 2020), the authors observe that both choices of the payoff function provide meaningful explanations assuming that they are used in the proper context. Authors in (Heskes et al., 2020), similar to (Janzing et al., 2020), consider a payoff function based on the interventional distribution and use do-calculus (Pearl, 2009) to calculate the payoff function when a causal partial ordering is available. On the other hand, the work in (Frye et al., 2020a) criticizes both interventional and observational Shapley values by arguing that the interventional payoff function is calculated off the data manifold where most machine learning models are unstable and that observational Shapley values are computationally expensive since they require estimating the conditional probability distribution. To address these criticisms, several works have proposed alternative payoff functions, such as Joint Baseline Shapley values (Yeh et al., 2022) and Manifold Shapley (Taufiq et al., 2023).

Additionally, the seemingly desirable aforementioned properties, in particular symmetry and additivity, have been questioned as well, since they may be pertinent to the task of dividing the payoff in a cooperative game, but not suitable in the context of explaining the decisions of a complex model (Sundararajan & Najmi, 2020; Frye et al., 2020b; Fryer et al., 2021; Jung et al., 2022; Yeh et al., 2022). The authors in (Fryer et al., 2021) provide several toy examples where enforcing these properties leads to unsatisfactory explanations. Meanwhile, authors in (Frye et al., 2020b) focus on the symmetry property, arguing that assigning equal contributions to identical (i.e., redundant) features conflicts with the goal of sparsity, which is often pursued in feature selection settings. In (Sundararajan & Najmi, 2020), the authors modify the definition of the missingness property and introduce a stronger property called "Dummy", which they deem preferable in the context of feature attribution. However, in (Chen et al., 2020) the authors present a counterexample, arguing that the dummy property may not be desirable in all applications, particularly in cases where the desired explanation is of interventional nature. In (Yeh et al., 2022), the authors depart from the context of cooperative games and seek to establish a list of desirable properties for a payoff function that should better fit the task of explaining a model. The work presented in (Ma & Tourani, 2020) provides a novel perspective by putting Shapley values in a context where the generative model is a Bayesian network. By investigating the explanation provided by Shapley values in this context, they argue that Shapley values are not always a good representative of the features' predictive power.

We follow a similar approach by considering structural equation models (SEMs) as our generative class, but our main contribution is to reinterpret Shapley values from a maximum entropy perspective by drawing an analogy with algorithms for causal discovery. Indeed the computation of standard Shapley values relies on averaging the marginal contributions of features over all possible orderings with equal weights. This is consistent with the assumption that no additional information is given about the probabilistic and causal relations among the features. However, causal inference algorithms typically make use of assumptions (e.g. faithfulness, see (Spirtes et al., 2000)) allowing one to consider some orderings of variables more likely than others. We show that it is possible to incorporate this kind of information with an appropriate choice of weights for different orderings. Interestingly, the work in (Frye et al., 2020b) uses a weighted approach to define alternative Shapley values as well, but their choice of weights is made a priori on the basis of a partial ordering relation consistent with a single causal Directed Acyclic Graph (DAG) underlying the data. Conversely, our approach determines a non-necessarily uniform weighting scheme after the computation of the marginal contributions of the features and is also capable of accommodating several causal DAGs providing different potential causal information about the data.

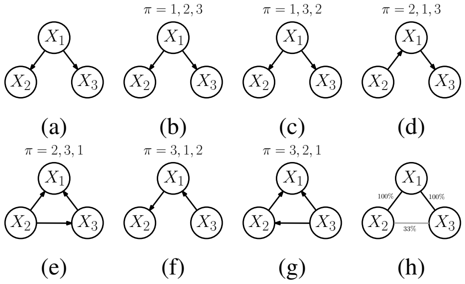
2. Faithfulness in causal inference

Apart from the computation of marginal contributions, another scenario in which different orderings of variables produce different results in the interpretation of their role in a model arises in the area of causal inference. This is exemplified by the application of Pearl-Verma Theorem (Theorem 2 in (Verma & Pearl, 1990a)), a fundamental result in the area of graphical models, to derive a graphical structure representing the conditional independence relations among a set of random variables. Pearl-Verma Theorem guarantees the existence and uniqueness of a DAG G with certain properties for a fixed ordering π over N variables. As a first property, the ancestry relation in G must be compatible with the ordering π. As a second property, G must be a minimal I-map (Verma & Pearl, 1990a; Koller & Friedman, 2009) for the joint probability distribution of the N variables, namely the relation of d-separation in the graph G has to imply conditional independence among the random variables (Koller & Friedman, 2009).

Pearl-Verma Theorem also provides a way to construct such a graph G. Let X_i be a variable and Z_i^π be the set of all variables preceding X_i in the ordering π. Pearl-Verma Theorem constructs a minimal I-map based on the ordering π by finding the largest set $P_i^\pi \subseteq Z_i^\pi$ such that X_i and $Z_i^\pi \backslash P_i^\pi$ are conditionally independent given P_i^π. The resulting graph G is obtained by connecting all nodes P_i^π to X_i. It is important to note that, as in the computation of the marginal contributions ϕ^π, the directed acyclic graph G output by the application of Pearl-Verma Theorem depends on the specific ordering π. In other words, the presence or absence of edges between variables in the graphical model G is in general determined by the choice of π.

The approaches followed by Shapley values and Pearl-Verma Theorem differ in how they address the dependence of their results on the ordering π. Shapley values address the issue by averaging the contributions over all possible orderings, while Pearl-Verma assumes an ordering a priori and outputs a DAG G. If we wanted to remove the dependence of the graph G on the a priori ordering, we could think of computing an ”average” graph over all possible orderings of the variables, similar to the approach used in Shapley values. This average graph could be a weighted graph where the weight $w_{ji} \in (0, 1]$ on the edge connecting X_i and X_j is given by the frequency with which that edge appears when applying Pearl-Verma Theorem on every possible ordering of the variables.

We elucidate the effect of this averaging procedure with an example. As a ground truth, assume a generative SEM with connectivity given by the 3-node graph in Figure 1(a). There are $3! = 6$ possible orderings for the three nodes in the graph. Applying Pearl-Verma to each ordering π, we obtain the directed acyclic graphs shown in Figures 1(b-g). If we disregard the orientation of the edges, and, for simplicity, only apply our averaging procedure to their skeletons (the unoriented version of these graphs), we obtain the weighted graph in Figure 1(h). Since all weights are different from zero, such a graph is complete. However, it has to be noted that the edges that have a 100% recovery rate are the same edges that are present in the true structure shown in Figure 1(a). This is because two adjacent nodes in the actual causal graph cannot be d-separated by any set Z, and therefore they will always be connected in any graph reconstructed by Pearl-Verma theorem, regardless of the ordering.



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 1: (a) is the original causal graph. (b-g) are the Pearl-Verma reconstruction given the ordering πs. (h) is the average graph of all the orderings.

This observation is a cornerstone of most algorithms or meta-algorithms for causal inference, such as IC (Inductive Causation) (Pearl & Verma, 1995), SGS or PC algorithms (Spirtes et al., 2000). Indeed, a basic assumption made by all these algorithms is that the joint probability distribution of the variables is faithful to its underlying causal structure (Spirtes et al., 2000; Pearl, 2009). Formally, faithfulness is defined by saying that the actual causal structure of the model is both an I-map and a D-map for the joint distribution (Koller & Friedman, 2009). However, a more captivating way of interpreting faithfulness is that the causal structure has the minimum number of edges possible given the observed conditional probability relations (Pearl, 2009). In other words, faithfulness is a form of Occam’s razor looking for simplest causal explanation, where ”simplicity” is defined as the graph’s edge count. Judea Pearl captures this idea perfectly in this quote:

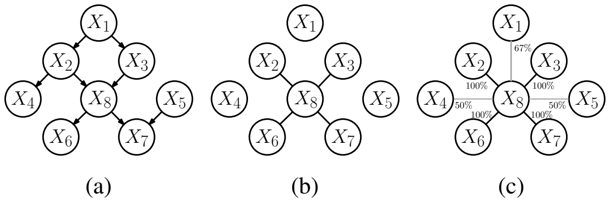
> *“[T]he scientist can never rule out the possibility that the underlying structure is a complete, acyclic, and* **arbitrar-**

> **ily ordered** *graph – a structure that (with the right choice of parameters) can mimic the behavior of any model,* **regardless of the variable ordering**. *However, following standard norms of scientific induction, it is reasonable to rule out any theory for which we find a simpler, less elaborate theory that is equally consistent with the data."* [J. Pearl, "Causality", Chapter 2, emphasis added.]

Let's highlight a deeper connection between the computation of Shapley values and causal inference by recasting a variation of IC algorithm that only recovers the edges adjacent to a single node in a probabilistic graphical model with faithful structure G. Without any loss of generality, let us assume that we are recovering the edges adjacent to X_{N+1} in the probabilistic model $P(X_1, ..., X_N, X_{N+1})$. The following Edge Discovery Algorithm 2 determines all edges adjacent to X_{N+1} under the faithfulness assumption. In the algorithm, the term $I(X_i, Z_i^\pi, X_{N+1})$ refers to a conditional independence test checking whether X_i is independent of X_{N+1} given the set Z_i^π. The output is the set of boolean variables E_i representing the presence of the unoriented edge $X_i - X_{N+1}$. The variable E_i is TRUE if and only if there is an edge connecting X_i and X_{N+1} in every ordering π of the variables $X_1, ..., X_N$. Interestingly, we observe that Algorithm 2 has striking similarities with the computation of Shapley values in Algorithm 1. Indeed, both algorithms loop over all possible orderings π of N variables and compute quantities associated with the same set Z_i^π defined as the set of variables preceding X_i in the ordering π. A more nuanced similarity lies in the fact that both algorithms check for a form of conditional independence. While Pearl-Verma tests for standard conditional independence, the computation of Shapley values, at least for the payoff $v(S) = E[f(X)|S]$, computes marginal contributions which, when vanishing, could be interpreted as a form of conditional mean independence (Teneggi et al., 2022). The fundamental difference between the two algorithms is in their outputs. Shapley value algorithm outputs the average (an evenly weighted sum) of the marginal contributions for all orderings, while Edge Discovery outputs the intersection (a product) of all edges appearing in all orderings. We can highlight this difference via an example. Assume a generative SEM faithful to the graph of Figure 2(a). Edge discovery applied to node X_8 outputs the graph of Figure 2(b) containing all and only the unoriented edges adjacent to node X_8 in the generative graph. Replacing the intersection (product) operation of Edge Discovery with an average (in the spirit of Shapley values) leads to the graph of Figure 2(c). In this case, node X_8 is connected to all other nodes and the weights on the edges are all strictly positive. Again, we observe that the edges which are present in all orderings are the edges actually present in the generative graph.

Algorithm 2 Edge Discovery

```
Require: Joint probability distribution $P$ over variables
  $X = \{X_1, ..., X_N, X_{N+1}\}$
  for every permutation $\pi$ of $\{1, ..., N\}$ do
    $Z_i^\pi \leftarrow$ Variables preceding $X_i$ in the ordering $\pi$
    if $I(X_i, Z_i^\pi, X_{N+1})$ then
      $E_i^\pi \leftarrow$ FALSE
    else
      $E_i^\pi \leftarrow$ TRUE
    end if
  end for
  $E_i \leftarrow \bigwedge_\pi E_i^\pi$
```



(a) (b) (c)

Figure 2: (a) represents the graphical representation of a SEM. (b) represents the unoriented structure recovered by Algorithm 2. (c) illustrates the unoriented graph obtained by applying an averaging scheme similar to the computation of Shapley values.

3. Adapting concepts of sparsity in causal inference to Shapley values

Causal inference algorithms typically postulate that the probabilistic model is faithful to its graph in order to reconstruct the edges linked to a single node. Under faithfulness, all orderings resulting in the minimum number of edges will consistently yield graphs with the same skeleton (Verma & Pearl, 1990b; Koller & Friedman, 2009). Conversely, in the standard computation of Shapley values no a priori assumptions are made and all orderings are taken into account and given an equal weight when producing the output. This comparison between the standard computation of Shapley values and causal inference algorithms seems to suggest that all orderings should be treated equally when no a priori knowledge is available, however, if some additional information about the model is given (e.g. faithfulness in a SEM), there could be reasons to favor some orderings over others. If we take a more general stance, we could now start investigating the implications of restricting the computation of Shapley values only to a subset of orderings, or weighting them in a non-necessarily uniform way. Algorithm 3 is an extension of Algorithm 1 that employs a more general weighting scheme to calculate the Shapley values of features. Specifically, Algorithm 3 uses weights $w^\pi \geqslant 0$, such that $\sum_\pi w^\pi = 1$, to account for the likelihood of different orderings or to incorporate prior knowledge about them. When the weights

Algorithm 3 Weighted Shapley

```
Require: Payoff $v$, instance of features $x = \{x_1, ..., x_N\}$
  for every permutation $\pi$ of $\{1, ..., N\}$ do
    $z_i^\pi \leftarrow$ features preceding $x_i$ in the ordering $\pi$
    $\phi_i^\pi \leftarrow v\left(z_i^\pi \cup \{x_i\}\right) - v\left(z_i^\pi\right)$
  end for
  $\phi_i \leftarrow \sum_\pi w^\pi \phi_i^\pi$
```

w^π are all equal to $\frac{1}{N!}$ we recover the standard definition of Shapley values, but alternative weighting choices could lead to explanations with different properties, as done in previous works (Frye et al., 2020b).

3.1. Sparsity is not fair to the spouses!

Inspired by the approach followed by causal inference algorithms, let's focus on a scenario where we compute the Weighted Shapley values by applying a similar form of Occam's razor principle. Namely, in Algorithm 3 we are going to consider only the orderings π leading to the sparsest vectors of marginal contributions ϕ^π. For example, if π gives rise to a ϕ^π that does not have the highest degree of sparsity, we set w^π to zero. Otherwise, we assign a positive weight. In order to examine the consequences of this choice, let's consider again a generative SEM faithful to the graph given by Figure 2(a) where we observe the variables $X_1, ..., X_7$ and we build a predictive model for the variable X_8. Assume that the prediction is $Y = f(X_1, ..., X_7)$ and is obtained by minimizing a least square criterion. Because of faithfulness, all the sparsest marginal contributions ϕ^π have the entries $\phi_2^\pi, \phi_3^\pi, \phi_6^\pi, \phi_7^\pi$ that are non-zero. Consequently the Weighted sparsest Shapley values $\phi_2, \phi_3, \phi_6, \phi_7$ are the only non-zero ones. This property can be stated in a more general way for linear SEMs via the following theorem.

Theorem 3.1. *Consider a generative linear SEM with the variables $X_1, ..., X_N, X_{N+1}$ faithful to the graph G^*. Let $f(X_1, ..., X_N) = E[X_{N+1}|X_1, ..., X_N]$ be the conditional expectation for X_{N+1}. Let ϕ_i be the Weighted Shapley values computed by assigning the weight $w^\pi = 0$ to any ordering π which does not provide a marginal contribution vector ϕ^π with the highest degree of sparsity. Then, $\phi_i \neq 0$ implies that X_i is either a parent or a child of X_{N+1} in G^*.*

Informally, Theorem 3.1 states that, when explaining a minimum least square error prediction model, limiting ourselves to the sparsest marginal contribution vectors leads to Weighted Shapley values ϕ where the non-zero entries are only associated with the parents and children of the predicted variable in G^*. The formal proofs of this theorem (and the following theorems) are in the Supplementary Material.

Even though such an explanation is the sparsest possible, there are reasons not to consider it as satisfactory. Indeed, it is a well-known fact that the least square prediction of a random variable in a probabilistic model faithful to a graph G^* relies not only on the parents and children, but also on the spouses of the predicted variable. We define spouses of a node, as the other nodes sharing at least a child with it. Spouses have an active contribution in the prediction by *explaining away* their effect on the shared children. In statistical modeling, ”explaining away” is a fundamental causal reasoning phenomenon that occurs when a particular variable or feature predicts the outcome of interest by diminishing the effect of other variables. Therefore, while the sparsest Weighted Shapley values may identify the immediate connections, neglecting the contribution of spouses results in explanations that fail to capture a complete picture of the prediction process.

3.2. Including ”explaining away”: Markov Blanket Shapley values

As observed in the previous section, the inclusion of spouses in the prediction is necessary as they play a fundamental role in explaining away the effect on their children. Specifically, parents provide information about observed causes for X_{N+1} and children provide information about observed effects of X_{N+1}. Given that this information is available, the spouses contribute to the prediction by providing information about alternative causes for the effects of X_{N+1} explaining away the effects of other causes. In order to incorporate the contribution of spouses to the prediction, we need to determine the appropriate orderings of variables to consider in computing the Weighted Shapley values. Since spouses contribute to the prediction merely by providing information about alternative explanations for observed effects, it is natural to consider orderings where the set of spouses follows the sets of parents and children. We recall that the set of parents, children, and spouses of a node is referred to as the Markov blanket of the node (Pearl, 1988; Koller & Friedman, 2009). The Markov blanket of a node is the minimal set that d-separates the node from the others. Thus, if we consider the orderings of the variables that begin with all the parents and children of the predicted node, adding solely its spouses would be sufficient to fully explain the outcome of the prediction since the Markov Blanket is the smallest set shielding the node from the rest of the variables in the model.

Indeed, if we consider all the orderings π of the variables starting with the set of parents and children (given by Theorem 3.1) of the predicted variable, and among those orderings choose the ones that have the sparsest marginal contribution vector ϕ^π to have weights $w^\pi \neq 0$, the nonzero entries of these ϕ^π correspond to the members of the Markov Blanket of the predicted variable. We can formally state this property in the following theorem.

Theorem 3.2. *Consider a generative linear SEM with the variables $X_1, ..., X_N, X_{N+1}$ faithful to the graph G^*. Let $f(X_1, ..., X_N) = E[X_{N+1}|X_1, ..., X_N]$ be the conditional expectation for X_{N+1}. In Algorithm 3, assign $w^\pi = 0$ to any ordering of the variables that do not start with parents and children of node X_{N+1}. Let vectors ϕ^π be the marginal contribution vectors computed by Algorithm 3 for all the orderings π that start with the parents and children of node X_{N+1}. Assign weight $w^\pi = 0$ to any ordering π which does not provide a marginal contribution vector ϕ^π with the highest degree of sparsity. Then, $\phi_i \neq 0$ implies that X_i is either a parent or child or a spouse of X_{N+1} in G^*.*

3.3. Weighting via maximum entropy principle

Both Theorem 3.1 and Theorem 3.2 provide sparsity properties for the Weighted Shapley values computed by Algorithm 3. They achieve sparsity by setting the weight $w^\pi = 0$ for any ordering π that is deemed irrelevant for the computation of the contributions. However, these results do not explicitly specify the weights for the orderings that are being considered. In comparison, standard Shapley values, computed by Algorithm 1, allocate equal weights to all orderings of the variables as no preference is given among them.

This approach can be seen as a maximum entropy principle, where weights are chosen to match the probability distribution that maximizes entropy in the absence of any other information (Jaynes, 1957; 1982). Following an entropy maximization approach we can define Markov Blanket Shapley values.

Definition 3.3. (Markov Blanket Shapley Values) Markov Blanket Shapley values are the output of Algorithm 3 setting $w^\pi = 0$ as per Theorem 3.2 and assigning equal weights to all other orderings.

4. Explaining the model or explaining what the model has learned

As already observed in (Chen et al., 2020), an important distinction has to be made about the type of explanation we are trying to obtain when computing additive feature attribution values. In the previous section, we were primarily concerned about explaining how a model predicts or estimates a variable X_{N+1} given the observations of the other variables $X_1, ..., X_N$ which led us to consider the payoff $v(S) = E[f(X)|S]$. The authors in (Chen et al., 2020) refer to this type of explanation as "observational", where the attribution values represent how much each observation contributes to the final estimate. Another type of explanation that may be of interest, also highlighted by (Chen et al., 2020), is of an "interventional" nature. In this case, the goal is to understand how manipulating the input variables affects the outcome, rather than simply assessing their predictive value. This form of explanation focuses on what the model has actually learned from a causal perspective. To capture the interventional impact of input variables on the output, we need to choose a different payoff function. As pointed out by (Heskes et al., 2020; Chen et al., 2020), in this situation a natural choice for the payoff function is $v(S) = E[f(X)|do(S)]$. This calculation relies on the knowledge of the underlying causal structure of the data (Pearl, 2009) as different causal structures lead to different interventional distributions. Results in (Heskes et al., 2020; Frye et al., 2020b) assume prior knowledge about the causal structure in the form of a graph or partial causal ordering of variables where the output is last in the ordering. However, these assumptions can be limiting in some situations given the various forms that prior knowledge about the causal structure can take.

As an example, assume that we obtain a dataset by sampling from a linear SEM with causal structure given by the unknown graph in Figure 3(a) and we train a linear model f that predicts the variable X_4. As a priori information, it is only given that the joint probability distribution of the variables is faithful to the causal graph. This allows one to infer the skeleton, representing the unoriented structure of the generative model, and also establish that X_2 is conditionally independent of X_3 given X_1. These constraints narrow down the candidate causal structures to those in Figure 3(b-d). Given our understanding of the chain graph based approach of (Heskes et al., 2020), this specific causal uncertainty cannot be represented in such a framework.

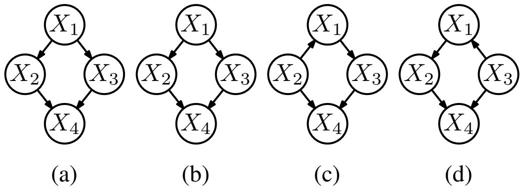


Figure 3: (a) Graphical structure of a causal model and candidate causal structures given faithfulness.

We can take each graph in Figures 3(b-d) separately and compute the interventional Shapley values using Algorithm 1 with the payoff function $v(S) = E[f(X)|do_{G_k}(x)]$ where G_k is the graph under consideration. In this way, each graph G_k provides potentially distinct plausible explanations.

We can again adopt our new framework based on entropy maximization and reconcile all these potentially conflicting explanations. Indeed, more generally, we can assume that prior knowledge about the causal structure of the variables is represented by a class of graphs for which we can calculate $v(S) = E[f(X)|do_{G_k}(S)]$. This class, denoted as $\mathcal{G}^* = \{G_1, G_2, \dots, G_M\}$, encompasses graphs compatible with the a priori knowledge of the phenomenon modeled by f. To calculate the interventional contribution of the input features while respecting our prior knowledge, we adopt a two-step approach. For each graph G_k ($k = 1, \dots, M$) in the set, we compute the associated Shapley value vector $\phi^{(k)}$ according to the payoff function $v(S) = E[f(X)|do_{G_k}(S)]$. Subsequently, we calculate the $\phi_{AS} = \frac{1}{M}\sum_k \phi^{(k)}$ and take it as the feature attribution vector. It's worth noting that, to have a non-zero contribution, an input feature must be an ancestor of the output variable in at least one of the graphs within $\mathcal{G}^*$. We formalize this idea with this definition.

Definition 4.1. (Ancestor Shapley Values) Given a set of graphs $\mathcal{G}^* = \{G_1, G_2, \dots, G_M\}$, Define $\phi^{(k)}$ as the Shapley value vector obtained using Algorithm 1 with the payoff function $v(S) = E[f(X)|do_{G_k}(S)]$. The uniform Ancestor Shapley values ϕ_{AS} are the average of $\phi^{(k)}$ overall M graphs in $\mathcal{G}^*$.

Ancestor Shapley values have been defined for simplicity on a set of graphs $\mathcal{G}^*$. However, they can be readily generalized considering any set $\mathcal{G}^*$ containing graphical structures for which interventional distributions can be computed (e.g. Chain Graphs (Lauritzen & Richardson, 2002; Heskes et al., 2020)).

Remark 4.2. If $\mathcal{G}^*$ contains all the complete DAGs where the output variable of the model has no descendants, the resulting Ancestor Shapley values end up coinciding with the standard Shapley values with payoff $v(S) = E[f(X)|S]$.

5. Discussion on lost and preserved properties

As discussed in Section 1, Shapley values are uniquely defined by the properties of Efficiency, Missingness, Symmetry, and Linearity. However, with non-uniformly weighted marginal contributions and a payoff function influenced by a priori knowledge, some of these properties may no longer hold. In this regard, we present the following theorem which characterizes the properties preserved by our proposed definition of Markov Blanket Shapley values and Ancestor Shapley values.

Theorem 5.1. *Markov Blanket Shapley values satisfy Efficiency, Missingness, and Symmetry. Ancestor Shapley values satisfy Missingness and Efficiency. Ancestor Shapley values satisfy Symmetry subject to the condition that if two variables X_i and X_j provide the same marginal contributions when swapped in every ordering π, then for any graph G_k in the set $\mathcal{G}^*$, swapping the nodes X_i and X_j in G_k results in a graph $G_\ell \in \mathcal{G}^*$.*

Hence, for the weighting schemes introduced in Definitions 3.3, we observe that Efficiency, Missingness, and Symmetry are preserved. In the context of Ancestor Shapley values, the symmetry of two input variables, as per standard definition mentioned in Section 1, yields identical interventional contributions when our prior knowledge about the causal structure aligns with this symmetry. However, if prior knowledge doesn't support the symmetry between the two input features, Ancestor Shapley values might end up being different.

In the context of the Efficiency property, it's essential to note that the computation of the payoff function, $v(x)$, varies depending on each causal graph $G_k \in \mathcal{G}^*$. For example, consider a scenario where the a priori causal knowledge about the phenomenon is compatible with a graph $G_k \in \mathcal{G}^*$ in which the output variable does not have any ancestors. This leads to $v(x) = E[f(X)|do_{G_k}(x_1, \dots, x_N)] = E[f(X)]$ which means that the value $\sum_{i=1}^{N} \phi_i$ is not necessarily unique across all graphs in $\mathcal{G}^*$, prompting us to introduce a new concept called intervenability. Intervenability quantifies the extent of interventional power over the output variable via the model's explanation through Shapley values, taking into account our prior knowledge.

Definition 5.2 (Intervenability)**.** Given a model f with input variables $X_1, \dots, X_N$ and output variable X_{N+1} and a priori information about the causal structure given in the form of a set of graphs $\mathcal{G}^* = \{G_1, \dots, G_M\}$, intervenability is defined in the following way:

$$\mathcal{I} = \frac{1}{M}\sum_{k=1}^{M} |E[f(X)|do_{G_k}(x)] - E[f(X)]| \, . \tag{1}$$

We also observe that in both Markov Blanket and Ancestor Shapley values the Linearity property is lost. Other desirable properties for linear attribution methods have been formulated as well. The authors in (Sundararajan & Najmi, 2020) propose an additional property, referred to as "Dummy." A feature X_i is dummy for f if for any two values x_i and x_i' we have $f(X_1, \dots, X_{i-1}, x_i, X_{i+1}, \dots, X_{N+1}) = f(X_1, \dots, X_{i-1}, x_i', X_{i+1}, \dots, X_{N+1})$. According to the Dummy property, every dummy feature should get a vanishing attribution. We have that Markov Blanket Shapley values satisfy Dummy, but Ancestor Shapley values do not.

Theorem 5.3. *Markov Blanket Shapley values satisfy Dummy.*

6. Comparison with other methods

While a more detailed numerical analysis can be found in the supplemental material, we briefly discuss some key insights here about different variations of interventional Shapley Values (SVs). Specifically, we would like to compare Ancestor SVs, Causal SVs (Heskes et al., 2020) and Asymmetric SVs (Frye et al., 2020b), using a minimalistic linear SEM, so that results can be verified against a known causal structure. This controlled setting eliminates complicating factors present in real-world data. We choose a linear SEM that is faithful to the graph 4(a), with a high correlation between the zero mean variables X_1 and X_2 ($\rho_{X_1X_2} = 0.98$). We train a linear model f that predicts variable X_3 using a least square error criterion. Assume we want an explanation for the point $x_1 = 0.57$, $x_2 = 0.55$, $f(x) = 0.39$. Using standard Shapley values with payoff $v(s) = E[f(X)|S]$ we find nearly identical Shapley values for X_1 and X_2 ($\phi = (0.2, 0.19)$). Now, we investigate the impact of causal a priori knowledge exploring three different scenarios. The numerical results are shown in Table 1.

Table 1: Comparison of different variations of interventional Shapley values. The last column represents intervenability.

	Asymmetric	Causal	Ancestor	$\mathcal{I}$
$\mathcal{G}_1^*$	$(0.39, 0.0)$	$(0.2, 0.19)$	$(0.2, 0.19)$	0.39
$\mathcal{G}_2^*$	-	$(0.2, 0.19)$	$(0.1, 0.29)$	0.39
$\mathcal{G}_3^*$	-	-	$(0.07, 0.19)$	0.26

First, assume knowledge of the exact causal structure. Namely the set $\mathcal{G}_1^*$ contains only the actual generative graph (Figure 4(a)). The resulting explanations given by the different methods are reported in Table 1. As already observed in (Heskes et al., 2020), Asymmetric SVs have the main drawback of concentrating interventional power in the root nodes of a DAG (X_1 in this case) neglecting the fact that intervening on other variables still has effect on the output. With perfect causal knowledge, Causal SVs and Ancestor SVs give identical explanations. This result aligns well with intuition as X_1 and X_2 are highly correlated and from the a priori knowledge it is known they are both ancestors of X_3. For the second case, the a priori information is that the joint probability distribution is faithful to the causal graph and that the output is the last in the causal ordering resulting in the graphs in set $\mathcal{G}_2^*$ of Figure 4(b). Asymmetric SVs are not defined with uncertain causal information. On the other hand, we can compute Causal SVs as defined in (Heskes et al., 2020) considering the partial causal ordering $\{(X_1, X_2)\}$ with mutual interactions between X_1 and X_2. Causal SVs produce the same explanation as in the previous case without taking into account the additional uncertainty. Conversely, Ancestor SVs give a larger contribution to X_2, because they take into account the fact that X_2 is definitely an ancestor of the output, while X_1 is not necessarily an ancestor. For the third case, we assume that the priori information boils down to just faithfulness of joint probability distribution to the casual graph, leading to the set $\mathcal{G}_3^*$ in Figure 4(c). Since no partial ordering of the features is compatible with all three graphs in the set, we can only calculate Ancestor SVs. We can see that both contributions have decreased from the previous case. This is because the additional graph in the set $\mathcal{G}_3^*$ suggests that the output variable might be uncontrollable from all inputs. Observe that this decreases intervenability from 0.39 for the previous cases to just 0.26.

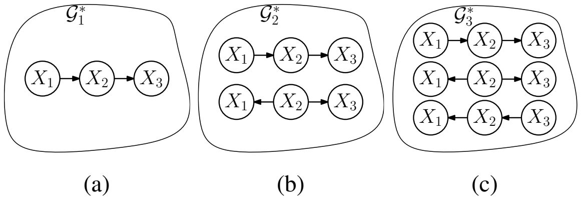


Figure 4: Different prior knowledge for the minimalistic SEM of Section 6

7. Conclusions

In this article, we start by exploring the parallels between the computation of Shapley values and causal discovery algorithms. We find that, methodologically, the primary distinction lies in their treatment of multiple plausible explanations for an outcome. This brings us to conclude that Standard Shapley values can be seen as a maximum entropy approach when no prior knowledge is available. We formalize two distinct approaches for calculating Shapley values in the presence of certain additional information about the underlying phenomenon. The first approach quantifies the role of a feature in the output of the model, while the second approach quantifies the potential interventional impact of a feature on the variable that the model is predicting. We argue that these approaches are inspired by the maximum entropy principle trying to reflect the additional information available.

8. Impact Statement

This article presents work whose goal is to advance the field of trustworthy machine learning. There might be potential societal consequences, none of which we feel must be specifically highlighted here.

References

Aas, K., Jullum, M., and Løland, A. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021.

Chen, H., Janizek, J. D., Lundberg, S., and Lee, S.-I. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

Datta, A., Sen, S., and Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pp. 598–617. IEEE, 2016.

Frye, C., de Mijolla, D., Begley, T., Cowton, L., Stanley, M., and Feige, I. Shapley explainability on the data manifold. In *International Conference on Learning Representations*, 2020a.

Frye, C., Rowat, C., and Feige, I. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems*, 33:1229–1239, 2020b.

Fryer, D., Strümke, I., and Nguyen, H. Shapley values for feature selection: The good, the bad, and the axioms. *IEEE Access*, 9:144352–144360, 2021.

Ghalebikesabi, S., Ter-Minassian, L., DiazOrdaz, K., and Holmes, C. C. On locality of local explanation models. *Advances in neural information processing systems*, 34: 18395–18407, 2021.

Giudici, P. and Raffinetti, E. Shapley-lorenz explainable artificial intelligence. *Expert systems with applications*, 167:114104, 2021.

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*, 33:4778–4789, 2020.

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, 2020.

Jaynes, E. T. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Jaynes, E. T. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982.

Jung, Y., Kasiviswanathan, S., Tian, J., Janzing, D., Blöbaum, P., and Bareinboim, E. On measuring causal contributions via do-interventions. In *International Conference on Machine Learning*, pp. 10476–10501. PMLR, 2022.

Koller, D. and Friedman, N. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Kumar, I. E., Venkatasubramanian, S., Scheidegger, C., and Friedler, S. Problems with shapley-value-based explanations as feature importance measures. In *International Conference on Machine Learning*, pp. 5491–5500. PMLR, 2020.

Lauritzen, S. L. and Richardson, T. S. Chain graph models and their causal interpretations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64 (3):321–348, 2002.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Ma, S. and Tourani, R. Predictive and causal implications of using shapley value for model interpretation. In *Proceedings of the 2020 KDD workshop on causal discovery*, pp. 23–38. PMLR, 2020.

Osborne, M. J. and Rubinstein, A. *A course in game theory*. MIT press, 1994.

Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pearl, J. and Verma, T. S. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pp. 789–811. Elsevier, 1995.

Remman, S. B., Strümke, I., and Lekkas, A. M. Causal versus marginal shapley values for robotic lever manipulation controlled using deep reinforcement learning. In *2022 American Control Conference (ACC)*, pp. 2683–2690. IEEE, 2022.

Shapley, L. A value for n-person games. *Annals of Mathematical Studies*, 28:307–317, 1953.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. *Causation, prediction, and search*. 2000.

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

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

Taufiq, M. F., Blöbaum, P., and Minorics, L. Manifold restricted interventional shapley values. *arXiv preprint arXiv:2301.04041*, 2023.

Teneggi, J., Bharti, B., Romano, Y., and Sulam, J. From shapley back to pearson: Hypothesis testing via the shapley value. *arXiv preprint arXiv:2207.07038*, 2022.

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

Verma, T. and Pearl, J. Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 255–270, 1990b.

Watson, D. Rational shapley values. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1083–1094, 2022.

Yeh, C.-K., Lee, K.-Y., Liu, F., and Ravikumar, P. Threading the needle of on and off-manifold value functions for shapley explanations. In *International Conference on Artificial Intelligence and Statistics*, pp. 1485–1502. PMLR, 2022.

A. Proofs of Theorems

A.1. Proof of Theorem 3.1

Proof. Since the generative structural equation model is linear, mean conditional independence is equivalent to conditional independence. Furthermore, since the model is faithful to its underlying directed acyclic graph, we can conclude that if two nodes are conditionally independent given some set of variables, they do not share an edge in the underlying graph. Therefore, when a variable exhibits a vanishing marginal contribution, it implies that it does not share an edge with X_{N+1}. It is evident that orderings of the variables that have the least number of non-zero marginal contributions exhibit $\phi_i^\pi \neq 0$ only for variables that are directly connected to X_{N+1} and thus are either a parent or child of X_{N+1}. □

A.2. Proof of Theorem 3.2

Proof. Similar to the argument presented in the proof of Theorem 1, for a linear structural equation model faithful to DAG G^*, a vanishing marginal contribution is equivalent to both conditional independence in the model and d-separation in the underlying DAG G^*. From the theory of graphical models, one node and its spouse can never be d-separated by any set that includes their common child (Verma & Pearl, 1990a; Pearl & Verma, 1995). Thus, for all orderings of the variables that start with parents and children of X_{N+1}, a spouse X_i of X_{N+1} will always exhibit $\phi_i^\pi \neq 0$. Furthermore, the set of parents, children, and spouses of a node forms the minimal set that d-separates it from all other variables in the graph. Hence, the non-zero entries of ϕ^π for orderings that start with parents and children of X_{N+1} which are also the sparsest only include parents, children, or spouses of X_{N+1}. □

A.3. Proof of Theorem 5.1

Proof. It is clear that Markov Blanket Shapley values satisfy efficiency since no variable or set of variables is excluded from the computation of Shapley values and only a different weighting scheme is applied to the ordering of the variables (We know that $\sum_{j=1} \phi_j^\pi = v(x) - E[f(X)]$ for all π). The missingness property holds for both Markov Blanket and Ancestor Shapley since the choice of non-negative weights does not affect the contribution ϕ_i of a variable if that variable has zero marginal contribution in every ordering of variables. For Markov Blanket Shapley Values the symmetry property is preserved because we are following a maximum entropy approach where all orderings with the same sparsity pattern receive equal weights. In the case of Ancestor Shapley values, consider two nodes X_i and X_j with identical marginal contributions when swapped in every ordering. The symmetry property follows since the class $\mathcal{G}^*$ contains a graph G_k if and only if swapping the two nodes X_i and X_j in it creates a graph G_ℓ that is still in $\mathcal{G}^*$. Ancestor Shapley values maintain efficiency. The value $v(x) = E[f(X)|do_{G_k}(x_1, x_2, \ldots, x_N)]$ may vary depending on the structure of the graph G_k. However, for each graph, we apply the standard Shapley value calculation with uniform weights for all possible orderings, as outlined in Algorithm 1 in the main text. Consequently, efficiency holds for each graph. if we consider the set $\mathcal{G}^*$ and define $v(x)$ for the set to as:

$$v(x) = \frac{1}{M} \sum_k E[f(X)|do_{G_k}(x)]. \tag{2}$$

Then,

$$\begin{aligned} \frac{1}{M} \sum_k \phi^{(k)} &= \frac{1}{M} \sum_k \sum_{i=0}^{N} \phi_i^{(k)} \\ &= \frac{1}{M} \sum_k E[f(X)|do_{G_k}(x)] \\ &= v(x). \end{aligned}$$

Thus, efficiency is preserved for Ancestor Shapley values. □

A.4. Proof of Theorem 5.3

Proof. The Dummy property states that variables that do not directly affect the computation of model f should receive a vanishing attribution. This property is satisfied by Markov Blanket Shapley values since this approach determines the minimal set that explains the model f.

In fact, the Dummy property embodies a condition that explicitly requires explanations to attribute potentially non-zero contributions only to the elements of the Markov Blanket in the case of linear structural equation models. □

B. Brief discussion on the linearity property

We mentioned that neither Ancestor Shapley values nor Markov Blanket Shapley values satisfy the Linearit/Additivity property. Among the four main properties satisfied by Standard Shapley values, it has been argued that Linearity is the one with less support from common intuition. Indeed, as observed by Osborne and Rubinstein (Osborne & Rubinstein, 1994) and reported in (Kumar et al., 2020)

> *[T]he notion of the sum of two games is not especially meaningful, the additivity [linearity] axiom has been described by game theorists as "mathematically convenient" and "not nearly so innocent" (Osborne & Rubinstein, 1994).*

This observation can be exemplified within the context of faithful Bayesian networks, as the generative class of models. When considering both observational and interventional explanations, enforcing linearity implies that the explanation of a function predicting the sum of two variables should be equivalent to the sum of their individual explanations. However, in the case of Bayesian networks, this does not always hold true, leading to situations where such an explanation may not be acceptable.

C. Numerical experiments: Markov Blanket vs. Sparsest Shapley values

This example aims to draw a comparison between the computed standard, sparsest, and Markov Blanket Shapley values for a structural equation model. This comparison confirms that averaging only the sparsest sets of marginal contributions (an extreme form of Occam's razor) can lead to unsatisfactory explanations.

Consider the following structural equation model

$$\begin{aligned}
X_1 &= e_1\\
X_2 &= 0.707 * X_1 + 0.707 * e_2\\
X_3 &= e_3\\
X_4 &= e_4\\
X_5 &= 0.707 * X_3 + 0.707 * e_5\\
X_6 &= 0.57 * X_4 + 0.57 * X_9 + 0.57 * e_6\\
X_7 &= 0.57 * X_5 + 0.57 * X_9 + 0.57 * e_7\\
X_8 &= 0.707 * X_9 + 0.707 * e_8\\
X_9 &= 0.57 * X_2 + 0.57 * X_3 + 0.57 * e_9
\end{aligned}$$

faithful to DAG shown in Figure 5, where $e = (e_1, \dots, e_9)^T$ are mutually independent Gaussian variables with mean equal to 0 and variance equal to 1.

We create a data set $\mathcal{D}$ with 10000 samples from this structural equation model. We train a linear regression model f that minimizes the mean square error of predicting X_9 from all the other variables in the structural equation model. We compute the Standard, Sparsest, and Markov Blanket Shapley values for the following data instance:

$$\begin{aligned}
&x_1 = 1.804, x_2 = 1.124, x_3 = 0.336, x_4 = -1.877,\\
&x_5 = 0.498, x_6 = -0.938, x_7 = 0.171, x_8 = -0.997,\\
&f(x_1, \dots, x_8) = 0.28
\end{aligned}$$

To compute the Standard Shapley values, we follow the steps outlined in Algorithm 1 in the article. In this example, our focus is on observational Shapley values and thus the computation of marginal contributions involves the computation of expected value over the conditional probability distribution. Indeed, the payoff function for observational Shapley values is defined as $v(S) = E[f(X)|S]$. Given that the generative model is linear, we employ linear regressions to estimate the conditional expectations. The resulting standard Shapley values for the data instance are reported in Figure 6(a). We observe

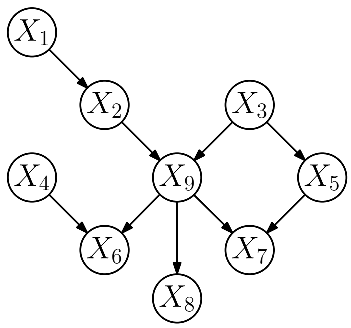


Figure 5: The underlying graph of structural equation model discussed in section C

that variable X_1 which does not play a role in any estimator derived through a mean square error minimization receives a larger Shapley value than most of the variables in the Markov Blanket of X_9. In other words, in this example we see that standard Shapley values do not satisfy the Dummy property defined in (Sundararajan & Najmi, 2020).

In order to compute the sparsest Shapley values, we follow the procedure described in Theorem 1. A practical way to determine whether ϕ_i^π is significantly different from zero would be to check the condition via a statistical test (e.g. via a F-test since the distribution is linear and Gaussian). However, in this example, for simplicity, we have assumed the presence of an oracle that can assess whether ϕ_i^π is vanishing or not, so that our results are not going to be affected by false positives/negatives in the computation of the statistical tests. Furthermore, during the computation process, we use uniformly distributed weights for $w^\pi \neq 0$. The explanation provided by averaging over orderings of the variables that have the sparsest ϕ^π is shown in Figure 6(b). We observe that variables with non-zero contributions are the exact same variables sharing an edge with X_9 in Figure 5 (X_2, X_3, X_6, X_7). In this case we notice that the Dummy property is verified, but as explained in the main article, the effect of explaining away from the spouses of X_9 is not taken into account.

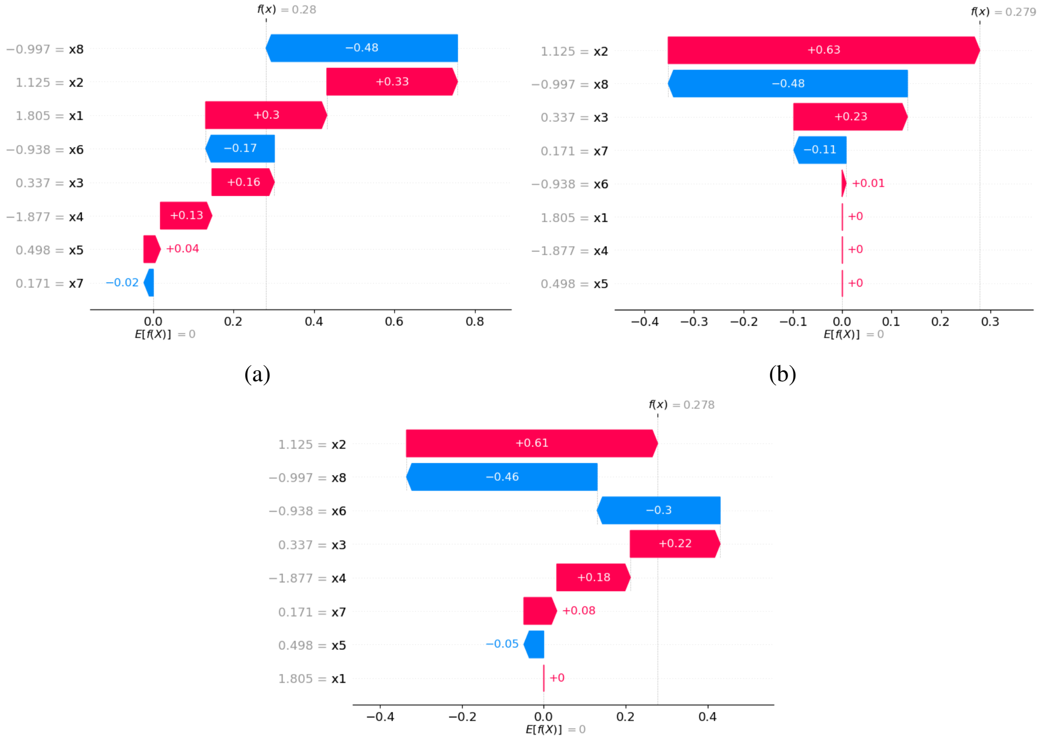
By leveraging the results obtained from sparsest Shapley values, we can compute Markov Blanket Shapley values from Definition 1 in the article. For Markov Blanket Shapley values, we solely consider orderings of the variables starting with parents and children of X_9 and among those orderings choose the ones with fewest non-zero entries in ϕ_i^π. The resulting explanation, presented in Figure 6(c), shows that all elements of the Markov Blanket of node X_9 receive a non-zero contribution and variable X_1 which is not part of the Markov Blanket receives zero contribution. In this case, we see that the Dummy property is still verified.

Furthermore, by comparing the results of sparsest Shapley values and Markov Blanket Shapley values shown in Figure 6(b,c), we can gain some insights into why the former fail to provide a fully satisfactory explanation of the prediction process. In model f, the regression coefficients of variables X_6 and X_7 are positive values. However, the explanation provided by sparsest Shapley values in Figure 6(b) suggests that their contribution to the outcome of the model is in the opposite direction of their realized value. This can be attributed to the exclusion of spouses, which are necessary to account for the phenomenon known as “explaining away”. Indeed, the regression coefficients associated with nodes X_4 and X_5 in model f are negative which is reflected in their contributions calculated by Markov Blanket Shapley values.

As an additional note, the evaluation of Markov Blanket Shapley values can be made more computationally attractive by using properties of factorized probability distributions. Specifically, if q is the number of parents and children of the predicted node, the computation of Markov Blanket Shapley values requires computing marginal contributions for $q! \times (N-q)!$ orderings of the variables instead of the standard $N!$. If our objective is to obtain a plausible explanation of the prediction process and, therefore, we solely intend to compute Markov Blanket Shapley values, exploring computationally efficient methods to determine the set of parents and children can expedite the calculation of Markov Blanket Shapley values without the necessity of computing sparsest Shapley values.

C.1. A more efficient calculation of Markov Blanket Shapley values

Causal inference techniques can guide us to find computationally efficient approaches for determining the elements of the Markov Blanket and the set of parents and children of the predicted node. Indeed, since the Markov Blanket of a node is the



(a)

(b)

(c)

Figure 6: (a), (b), and (c) report the Standard Shapley values, sparsest contributions, and the Markov Blanket Shapley values, respectively, for the variables in the example discussed in Section C.

minimal set that shields the node from the other nodes in a probabilistic model, the following Algorithm 4 can effectively identify the elements of the Markov Blanket with $\mathcal{O}(N)$ computational complexity.

Algorithm 4 Markov Blanket discovery

```
Require: Payoff $v(S) = E[f(X)|S]$, instance of features $x = \{x_1, ..., x_N\}$
  $MB \leftarrow \{\}$
  for instance of variable $x_i$ in $x$ do
    $z_i \leftarrow x/\{x_i\}$
    $\phi_i \leftarrow v(z_i \cup \{x_i\}) - v(z_i)$
    if $\phi_i \neq 0$ then
      Add $X_i$ to $MB$
    end if
  end for
```

In this algorithm, we focus on scenarios where the possibility of $\phi_i = 0$ arises solely from the condition where X_i is mean independent from X_{N+1} given all other variables. Other potential causes leading to $\phi_i = 0$ are disregarded in this context. Since the Markov Blanket of a node consists of its parents, children, and spouses, an algorithm that aims to discover the set of parents and children can iterate over this set rather than the set of all the features leading to more computationally efficient algorithms. Inspired by methods in causal discovery, we can devise the following algorithm to find the set of parents and children of node X_{N+1}(Algorithm 5).

This algorithm can identify the set of parents and children in $\mathcal{O}(|MB|^2)$, where $|MB|$ is the number of variables in the Markov Blanket identified from Algorithm 4.

Algorithm 5 Retrieving the set of Parents and Children

```
Require: Payoff $v(S) = E[f(X)|S]$, instance of features $x = \{x_1, ..., x_N\}$, and Markov Blanket of the predicted node $MB$
  $PaCh, Sp, Ch \leftarrow \{\}$
  for every $x_i \in MB$ do
    for every $x_j \neq x_i \in MB$ do
      $z_{ji} \leftarrow MB/\{x_i, x_j\}$
      $\phi_j \leftarrow v\left(z_{ji} \cup \{x_j\}\right) - v\left(z_{ji}\right)$
      if $\phi_j = 0$ then
        Add $X_j$ to $Sp$
        Add $X_i$ to $Ch$
      end if
    end for
  end for
  $PaCh \leftarrow MB/Sp$
```

D. Numerical experiments: Ancestor Shapley values

D.1. Minimalistic example

In this section, we will begin by providing a comprehensive description of the minimalistic example presented in the main body of the article. Consider the following linear SEM:

$$\begin{aligned}
X_1 &= e_1 \\
X_2 &= 0.98 * X_1 + 0.2 * e_2 \\
X_3 &= 0.707 * X_2 + 0.707 * e_3,
\end{aligned}$$

faithful to the directed graph shown in Figure 4(a). We assume $e = (e_1, \dots, e_3)^T$ are mutually independent Gaussian variables with mean equal to 0 and variance equal to 1. We create a data set $\mathcal{D}$ with 100000 samples from this structural equation model. We note that the variables X_1 and X_2 are chosen such that $\rho_{X_1 X_2} = 0.98$ indicating that these variables are strongly correlated. We train a linear model f that predicts variable X_3 using a least square error criterion. Assume we want a local explanation for the arbitrarily chosen point $x_1 = 0.57$, $x_2 = 0.55$, $f(x) = 0.39$ from an interventional perspective.

Calculating the standard Shapley values with payoff function $v(s) = E[f(X)|S]$, we find nearly identical Shapley values for X_1 and X_2 ($\phi = (0.2, 0.19)$). Note that ϕ_1 is slightly larger than ϕ_2 due to the slightly higher value of x_1 at the chosen test data point. Now, we investigate the impact of causal a priori knowledge exploring three different scenarios. The numerical results are shown in Table 1.

First, we assume that we have precise a priori knowledge of the causal structure. In this case, the set $\mathcal{G}_1^*$ exclusively contains the true generative graph, as illustrated in Figure 4(a). The respective explanations provided by various methods are documented in Table 1. We used the code from (Remman et al., 2022) to calculate the Causal Shapley values.

As previously noted in (Heskes et al., 2020), one of the primary drawbacks of Asymmetric Shapley values is their tendency to concentrate interventional power in the root nodes of a DAG, in this case X_1, neglecting the fact that intervening on other variables still has effect on the output. This occurs due to the use of a weighting scheme, as outlined in Algorithm 2 of the main article, which assigns uniformly distributed weights to orderings compatible with the graphical representation while assigning a weight of $w = 0$ to all other orderings of the variables.

In the case of our example, Asymmetric Shapley values would consider this weighting for the orderings:

$$w_{(1,2)} = 1, w_{(2,1)} = 0, \tag{3}$$

where $(1, 2)$ refers to ordering $\pi = (X_1, X_2)$. This results in giving all the interventional power to X_1.

For Causal Shapley values and Ancestor Shapley values, we see that both X_1 and X_2 receive nearly identical contributions. This outcome aligns with our intuition since the a priori knowledge indicates that both X_1 and X_2 are ancestors of the output variable and thus should both definitely have non-zero interventional contributions. Given their high correlation, it is reasonable to conclude that they should exhibit similar contributions from an interventional standpoint. We also note that with perfect causal knowledge, Causal Shapley values and Ancestor Shapley values give identical explanations.

For the second case, the a priori information is that the joint probability distribution is faithful to the causal graph. Furthermore, it is assumed that the output variable is the last in the causal ordering, resulting in the graphs in set $\mathcal{G}_2^*$ depicted Figure 4(b). Asymmetric Shapley values are not defined with uncertain causal information. On the other hand, we can compute Causal Shapley values as defined in (Heskes et al., 2020) considering the partial causal ordering $\{(X_1, X_2)\}$ with mutual interactions between X_1 and X_2. Causal Shapley values produce the same explanation as in the previous case without taking into account the additional uncertainty. Conversely, Ancestor Shapley values give a larger contribution to X_2, because they take into account the fact that X_2 is definitely an ancestor of the output, while X_1 is not necessarily an ancestor. We note that in both cases we have that $\mathcal{I} = f(x)$ because in both graphs parents of the output variables in the generative model are also its parents in all the candidate graphs in both $\mathcal{G}_1^*$ and $\mathcal{G}_2^*$.

For the third case, we assume that the priori information boils down to just faithfulness of joint probability distribution to the generative causal graph, leading to the set $\mathcal{G}_3^*$ in Figure 4(c). Since no partial ordering of the features is compatible with all three graphs in the set, we can only calculate Ancestor Shapley values. We can see that both contributions have decreased from the previous case. This is because the additional graph in the set $\mathcal{G}_3^*$ suggests that the output variable might be uncontrollable from all inputs. Observe that this decreases intervenability from 0.39 for the previous cases to just 0.26.

D.2. Dealing with unmeasured variables

In this section, we'll examine a simple example with an unmeasured variable to study its influence on Ancestor Shapley values in a linear SEM. Consider the following linear SEM:

$$\begin{aligned} U &= e_u \\ X_1 &= 0.707 * U + 0.707 * e_1 \\ X_2 &= 0.707 * X_1 + 0.707 * e_2 \\ X_3 &= 0.50 * U + 0.50 * X_2 + 0.5 * e_3. \end{aligned}$$

faithful to the directed graph shown in Figure 7(a). We assume $e = (e_u, e_1, \dots, e_3)^T$ are mutually independent Gaussian variables with mean equal to 0 and variance equal to 1. We create a data set $\mathcal{D}$ with 100000 samples from variables X_1, X_2, X_3 of this structural equation model. We train a linear model f that predicts variable X_3 from variables X_1, X_2 using a least square error criterion. Let's assume that we want an explanation for data point $X_1 = 1.213$ and $X_2 = 1.229$ and the prediction $f(X) = 1.045$ from an interventional perspective.

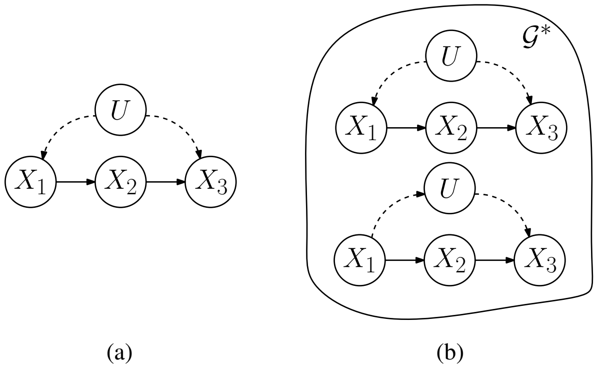
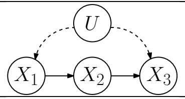
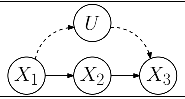


Figure 7: (a) is the true causal structure of the example given in section D.2. (b) is the set of the candidate graphs.

Let's assume that as a priori knowledge, we are given the set of candidate graphs in $\mathcal{G}^*$ as shown in Figure 7(b). This information suggests the presence of an unmeasured variable that contributes to the correlation between X_1 and X_3. However, it remains unclear whether this unmeasured variable is a confounder or a unobserved parallel path of causation from X_1 to X_3. We note that since the unmeasured variable is a parent of X_3, our approach for computing the payoff function $v(S) = E[f(X)|do(S)]$ simply requires the application of some rules of do-calculus.

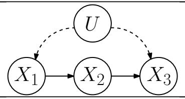
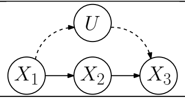
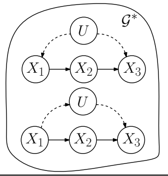
Table 2: *The calculation of the payoff function for each subset of the variables.*

	$E[f(X)\|do(x_1)]$	$E[f(X)\|do(x_2)]$	$E[f(X)\|do(x_1, x_2)]$
	$\int P(X_2\|x_1) \int f(X_1, X_2)P(X_1)dX_1dX_2 = 0.43$	$\int f(X_1, x_2)P(X_1)dX_1 = 0.62$	$\int f(X_1, x_2)P(X_1)dX_1 = 0.62$
	$\int f(x_1, X_2)P(X_2\|x_1)dX_2 = 0.86$	$\int f(X_1, x_2)P(X_1)dX_1 = 0.62$	$f(x_1, x_2) = 1.045$

The computation of the different interventions $do_{G_k}(x_1)$, $do_{G_k}(x_2)$ and $do_{G_k}(x_1, x_2)$ are reported in Table 2.

The final interventional contributions for each graph and also the Ancestor Shapley values for the a priori information $\mathcal{G}^*$ is presented in Table 3.

Table 3: *Shapley values computed for each graph using* $E[f(X)|do_{G_k}(x)]$*, and the Ancestor Shapley values for* $\mathcal{G}^*$ *given by averaging the Shapley values of the two graphs.*

	$\phi^{(k)}$	$E[f(X)\|do_{G_k}(x)] - E[f(X)]$
	(0.215, 0.405)	0.62
	(0.64, 0.403)	1.045
	ϕ	$\mathcal{I}$
	(0.428, 0.404)	0.83

This example highlights that Ancestor Shapley values provide a means to calculate Shapley values from an interventional perspective even when the given information includes hidden variables and the calculation of the payoff function is not straightforward.

Remark D.1. We observe that Ancestor Shapley values adopt the entropy maximization approach, assigning equal weight to all graphs. Yet, in cases where prior knowledge suggests preferences for certain graphs or deems them more probable than others, an analogous weighting scheme, akin to the one outlined in the Weighted Shapley Algorithm in the primary article, can be employed. This approach accommodates such information and facilitates the computation of interventional contributions that incorporate all available a priori knowledge.