Causal Shapley Values: Exploiting Causal Knowledge to Explain Individual Predictions of Complex Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Shapley values underlie one of the most popular model-agnostic methods within explainable artificial intelligence. These values are designed to attribute the difference between a model's prediction and an average baseline to the different features used as input to the model. Being based on solid game-theoretic principles, Shapley values uniquely satisfy several desirable properties, which is why they are increasingly used to explain the predictions of possibly complex and highly non-linear machine learning models. However, they are typically computed under the assumption that features are independent, which ignores any causal structure between the features and can lead to unreliable explanations.

In this paper, we propose a novel framework for computing Shapley values that generalizes recent work aiming to relax or defend the independence assumption. By employing Pearl's *do*-calculus, we show how these 'causal' Shapley values can be derived for general causal graphs without sacrificing any of their desirable properties. Moreover, causal Shapley values enable us to separate the contribution of direct and indirect effects. We provide a practical implementation for computing causal Shapley values based on causal chain graphs and illustrate their utility on several real-world examples.

1 Introduction

Complex machine learning models like deep neural networks and ensemble methods like random forest and gradient boosting machines may well outperform simpler approaches such as linear regression or single decision trees, but are notably harder to interpret. This can raise practical, ethical, and legal issues, most notably when applied in critical systems, e.g., for medical diagnosis or autonomous driving. The field of explainable AI aims to address these issues by enhancing the interpretability of complex machine learning models.

The Shapley-value approach, that we also focus on in this paper, has quickly become one of the most popular model-agnostic methods within explainable AI. It can provide local explanations, attributing changes in predictions for individual data points to the model's features, that can be combined to obtain better global understanding of the model structure [9]. Shapley values are based on a principled mathematical foundation [18] and satisfy various desiderata (see also Section 2). They have been applied for explaining statistical and machine learning models for quite some time, see e.g., [8, 20]. Recent interests have been triggered by Lundberg and Lee's breakthrough paper [10] that unifies Shapley values and other popular local model-agnostic approaches such as LIME [17], while at the same time introducing more efficient computational procedures.

When applied to explain the predictions of a machine learning model, Shapley values consider the difference between the model's prediction when knowing all feature values and its baseline prediction when knowing none of the feature values and split this difference among the features that are used as

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

input to the model. A crucial subroutine of the approach needs to compute or estimate the expected model output when some features are known, while others are dropped. Early approaches, such as [20] estimate this expectation by assuming that the features are independent. This is also the approach taken in [10], but mainly for computational reasons. We will refer to these as marginal Shapley values. Aas et al. [1] argue and illustrate that marginal Shapley values may lead to incorrect explanations when features are highly correlated, motivating what we will refer to as conditional Shapley values. Even more recently, Janzing et al. [5] suggest the contrary, when stating that marginal rather than conditional expectations provide the right notion of dropping features. They make a distinction between conditioning by observation and conditioning by intervention, and argue that the latter is to be preferred and then boils down to marginal expectations. This argument is also picked up by [9] when implementing interventional Tree SHAP. Where marginal and conditional Shapley values correspond to a uniform distribution over all possible permutations of the features, so-called asymmetric Shapley values, introduced by Frye et al. [4], propose to incorporate causal knowledge by choosing a non-uniform distribution of permutation consistent with a (partial) causal ordering. In line with [1], they then apply conventional conditioning by observation to make sure that the resulting explanations respect the data manifold.

In this paper, we will follow [5, 9] in proposing an active interpretation of Shapley values. (1) Following a different line of reasoning, we will derive 'causal' Shapley values that aim to explain the causal effect of features on the prediction and are truly different from marginal and conditional Shapley values. Compared to asymmetric Shapley values, causal Shapley values provide a more direct, orthogonal way to incorporate causal knowledge. (2) We extend the concept of Shapley values with the possibility to decompose feature attributions in direct and indirect effects. (3) Making use of so-called causal chain graphs [7], we propose a practical approach for computing causal Shapley values and illustrate this on several real-world examples.

2 A causal interpretation of Shapley values

Complete rewrite, attempting to argue for causal Shapley values instead of just proposing them. Old illustration out, new one in, but probably far too detailed. Suggestions on what to keep are welcome!

In this section, we will introduce the causal, interventional interpretation of Shapley values and contrast this to other approaches, such as conditional and asymmetric Shapley values. We consider a typical supervised machine learning scenario, in which we are given examples of feature vectors $\mathbf{x}$ with corresponding targets y, assumed to be drawn from an unknown distribution $P(Y|\mathbf{x})$. After training, we have access to a model $f(\mathbf{x})$, e.g., a deep neural network or an ensemble of trees, that can provide a prediction for every possible feature vector $\mathbf{x}$. We assume that we aim for a better understanding of prediction $f(\mathbf{x})$ for a particular feature vector, not because this prediction is just some arbitrary function of some inputs, but because it provides us relevant information about the actual distribution $P(Y|\mathbf{x})$. In most applications of supervised learning, we mainly care about the expectation $\mathbb{E}[Y|x]$ and a successfully trained model approximates some (link) function of this expectation:

$$f(\mathbf{x}) \approx g(\mathbb{E}[Y|\mathbf{x}]) \,,$$

with just the identity $g(z) = z$ for regression and for binary classification either the identity (as in [1]) or the logit $g(z) = \log(z/(1-z))$ (as in [10]).

Attribution methods, with Shapley values as their most prominent example, provide a local explanation of individual predictions by attributing the difference between $f(\mathbf{x})$ and a baseline f_0 in a sensible way to the different features $i \in N$ with $N = \{1, \ldots, n\}$ and n the number of features. That is, they write

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} \phi_i \,, \tag{1}$$

where we will refer to ϕ_i as the contribution of feature i to the prediction $f(\mathbf{x})$. For the baseline f_0 we will take the average prediction $f_0 = \mathbb{E}f(\mathbf{X})$ with expectation taken over some (for now assumed to be known) probability distribution $P(\mathbf{X})$, corresponding to the situation in which we would not know any of the feature values. Equation (1) is referred to as the efficiency property [18], which appears to be a sensible desideratum for any attribution method and we therefore take here as our starting point.

An obvious way to go from knowing none of the feature values, as for f_0, to knowing all feature values, as for $f(\mathbf{x})$, would be to add feature values one by one. Here we can think of (at least) two possible interpretations.

Passive. We interpret the feature vector $\mathbf{x}$ as a passive observation. Feature values come in one after the other and the contribution of feature i should reflect the difference in expected value of $f(\mathbf{X})$ after and before *observing* its feature value x_i.

Active. We interpret the feature vector $\mathbf{x}$ as the result of an action. Feature values are imposed one after the other and the contribution of feature i relates to the difference in expected value of $f(\mathbf{X})$ after and before *setting* its value to x_i.

Following the above sequential reasoning, the contribution of each feature depends on the order π in which the feature values arrive or are imposed. We write the contribution of feature i given the permutation π as

$$\phi_i(\pi) = v(\{j : j \preceq_\pi i\}) - v(\{j : j \prec_\pi i\}) \,, \tag{2}$$

with $j \prec_\pi i$ if j precedes i in the permutation π and where we define the value function

$$v(S) = \mathbb{E}\left[f(\mathbf{X})|op(\mathbf{x}_S)\right] = \int d\mathbf{X}_{\bar{S}}\, P(\mathbf{X}_{\bar{S}}|op(\mathbf{X}_S = \mathbf{x}_S)) f(\mathbf{X}_{\bar{S}}, \mathbf{x}_S) \,. \tag{3}$$

Here S is the subset of indices of features with known 'in-coalition' feature values $\mathbf{x}_S$. To compute the expectation, we still need to average over the 'out-of-coalition' or dropped feature values $\mathbf{X}_{\bar{S}}$ with $\bar{S} = N \setminus S$, the complement of S. The operator $op()$ specifies how the distribution of the 'out-of-coalition' features $\mathbf{X}_{\bar{S}}$ depends on the 'in-coalition' feature values $\mathbf{x}_S$. To arrive at the passive interpretation, we set $op()$ to conventional conditioning by observation, yielding $P(\mathbf{X}_{\bar{S}}|\mathbf{x}_S)$. For the active interpretation, we need to condition by intervention, for which we resort to Pearl's *do*-calculus [13] and write $P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S))$. A third option is to ignore the feature values $\mathbf{x}_S$ and just take the unconditional, marginal distribution $P(\mathbf{X}_{\bar{S}})$. We refer to the corresponding Shapley values as conditional, causal, and marginal, respectively. Below we will discuss when which type of Shapley value to apply.

Since the sum over features i in (2) is telescoping, it can be immediately seen that the efficiency property (1) holds for any permutation π. Therefore, for any distribution over permutations $w(\pi)$ with $\sum_\pi w(\pi) = 1$, the contributions

$$\phi_i = \sum_\pi w(\pi)\phi_i(\pi) \tag{4}$$

still satisfy (1). An obvious choice would be to take a uniform distribution $w(\pi) = 1/n!$. We then arrive at the standard definition of Shapley values:

$$\phi_i = \sum_{S \subseteq N \setminus i} \frac{|S|!(n - |S| - 1)!}{n!} \left[v(S \cup i) - v(S)\right] \,,$$

where we use shorthand i for the singleton $\{i\}$. Besides efficiency, these Shapley values uniquely satisfy three other desirable properties [18].

Linearity: for two value functions v_1 and v_2, we have $\phi_i(\alpha_1 v_1 + \alpha_2 v_2) = \alpha_1\phi_i(v_1) + \alpha_2\phi_i(v_2)$. This guarantees that the Shapley value of a linear ensemble of models is a linear combination of the Shapley values of the individual models.

Null player (dummy): if $v(S \cup i) = v(S)$ for all $S \subseteq N \setminus i$, then $\phi_i = 0$. A feature that never contributes to the prediction (directly nor indirectly, see below) receives zero Shapley value.

Symmetry: if $v(S \cup i) = v(S \cup j)$ for all $S \subseteq N \setminus \{i, j\}$, then $\phi_i = \phi_j$. Symmetry holds for marginal, conditional, and causal Shapley values.

Efficiency, linearity, and null player still hold for a non-uniform distribution of permutations, but symmetry is then typically lost. Properties such as null player and symmetry are defined with reference to the value function or, equivalently, w.r.t. binary variables indicating for each feature whether it is 'in-coalition' or 'out-of-coalition', so not with respect to the function $f(\mathbf{x})$ itself as in [21]. In the latter case, only marginal Shapley values satisfy the null player property, but are

not even guaranteed to satisfy the symmetry property: even though $f(x_1, x_2) = f(x_2, x_1)$, for all possible values x_1 and x_2, we can get $\phi_1 \neq \phi_2$ when $P(x_1) \neq P(x_2)$. Remove?

To understand which interpretation and then also type of Shapley values to prefer, we go back to what $f(\mathbf{x})$ is actually supposed to represent: some function of the expectation of the target Y given the features values $\mathbf{x}$. When the features are assumed to be potential causes of the target Y, we have $\mathbb{E}[Y|\mathbf{x}] = \mathbb{E}[Y|do(\mathbf{X} = \mathbf{x})]$. In this case, causal Shapley values, providing an explanation in terms of actively setting features to their values, appear to be most natural and informative choice, as we will also illustrate through examples in the next section. However, when the features are themselves mere consequences of the target Y, an active, interventional interpretation makes no sense, since then $\mathbb{E}[Y|do(\mathbf{X} = \mathbf{x})] = \mathbb{E}[Y] \neq \mathbb{E}[Y|\mathbf{x}]$. As an example, Y could represent the presence or absence of a disease and $\mathbf{x}$ the outcome of various tests. It is perfectly fine to build a predictive model for the probability of the disease given test outcomes, but an explanation discussing how this probability changes when we actively set a test outcome to its value makes no sense. When some features are causes of the target Y and others consequences, we can consider actively setting the causes and passively observing the consequences. This more complicated setting is beyond the scope of the current paper.

Our active, interventional interpretation of Shapley values from the outset appears to coincide with that in [5, 9]. However, by considering $f(\mathbf{x})$ an arbitrary function, instead of a learned relationship between features and a target, and by formally distinguishing between true features (corresponding to one of the data points) and the features plugged into the model, Janzing et al. [5] conclude that, in our notation, $P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S)) = P(\mathbf{X}_{\bar{S}})$ for any subset S. As a result, any expectation under conditioning by intervention reduces to a marginal expectation. Although the chosen construction is technically correct, it then also appears to throw the baby out with the bathwater: the (then reduced to marginal) Shapley values by construction can only incorporate direct effects and fail to provide insight into the *total effect* of setting a feature to its value.

When applied to incorporate causal knowledge, the asymmetric Shapley values introduced in [4] choose $w(\pi) \neq 0$ in (4) only for those permutations π that are consistent with the causal structure between the features, i.e., are such that a known causal ancestor always precedes its descendants. They provide somewhat of a mix between an active, interventional (incorporating causal structure into the allowed permutations) and passive, observational (conditioning by observation) approach. This idea, to restrict the allowed permutations when computing the Shapley values, can be considered orthogonal to the replacement of conditioning by observation with conditioning by intervention. We will therefore refer to the approach of [4] as asymmetric conditional Shapley values, to contrast them with asymmetric causal Shapley values that implement both ideas.

3 Decomposing Shapley values into direct and indirect effects

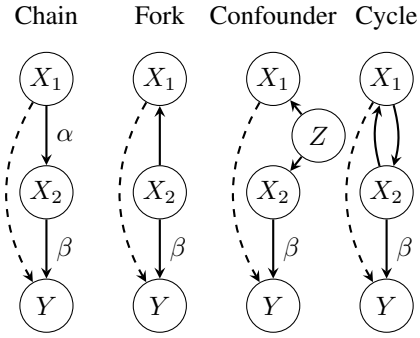
The contribution $\phi_i(\pi)$ of a particular permutation π and feature i in (2) measures the difference in value function with and without adding X_i to the 'in-coalition' features. With shorthand notation $\underline{S} = \{j : j \prec_\pi i\}$ and $\bar{S} = \{j : j \succ_\pi i\}$, we can decompose this total effect into a direct and an indirect effect:

$$\begin{aligned}
\phi_i(\pi) &= \mathbb{E}[f(\mathbf{X}_{\bar{S}}, \mathbf{x}_{\underline{S}\cup i})|do(\mathbf{X}_{\underline{S}\cup i} = \mathbf{x}_{\underline{S}\cup i})] - \mathbb{E}[f(\mathbf{X}_{\bar{S}\cup i}, \mathbf{x}_{\underline{S}})|do(\mathbf{X}_{\underline{S}} = \mathbf{x}_{\underline{S}})] && \text{(total effect)} \\
&= \mathbb{E}[f(\mathbf{X}_{\bar{S}}, \mathbf{x}_{\underline{S}\cup i})|do(\mathbf{X}_{\underline{S}} = \mathbf{x}_{\underline{S}})] - \mathbb{E}[f(\mathbf{X}_{\bar{S}\cup i}, \mathbf{x}_{\underline{S}})|do(\mathbf{X}_{\underline{S}} = \mathbf{x}_{\underline{S}})] + && \text{(direct effect)} \\
&\quad \mathbb{E}[f(\mathbf{X}_{\bar{S}}, \mathbf{x}_{\underline{S}\cup i})|do(\mathbf{X}_{\underline{S}\cup i} = \mathbf{x}_{\underline{S}\cup i})] - \mathbb{E}[f(\mathbf{X}_{\bar{S}}, \mathbf{x}_{\underline{S}\cup i})|do(\mathbf{X}_{\underline{S}} = \mathbf{x}_{\underline{S}})] && \text{(indirect effect)}
\end{aligned}$$

The direct effect measures the expected change in prediction when the stochastic feature X_i is replaced by its feature value x_i, without changing the distribution of the other 'out-of-coalition' features. The indirect effect measures the difference in expectation when the distribution of the other 'out-of-coalition' features changes due to the additional intervention $do(X_i = x_i)$. Direct and indirect Shapley values can be computed by taking a, possibly weighted, average over all permutations. Conditional Shapley values can be decomposed in the same way. Marginal Shapley values by construction only represent the direct effect.

To illustrate the difference between the various Shapley values, we consider four causal models on two features. They are constructed such that they have the same $P(\mathbf{X})$, with $\mathbb{E}[X_2|x_1] = \alpha x_1$ and $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$, but with different causal explanations for the dependency between X_1 and X_2. In the causal chain X_1 could, for example, represent season, X_2 temperature, and Y bike rental. The

Shapley values	*D*		*S*		*A*	
	direct	indirect	direct	indirect	direct	indirect
ϕ_1	0	0	0	$\frac{1}{2}\beta\alpha x_1$	0	$\beta\alpha x_1$
ϕ_2	βx_2	0	$\beta x_2 - \frac{1}{2}\beta\alpha x_1$	0	$\beta x_2 - \beta\alpha x_1$	0



	marginal	conditional		causal	
		symmetric	asymmetric	symmetric	asymmetric
Chain	*D*	*S*	*A*	*S*	*A*
Fork	*D*	*S*	*D*	*D*	*D*
Confounder	*D*	*S*	*S*	*D*	*D*
Cycle	*D*	*S*	*S*	*S*	*S*

Figure 1: Direct and indirect Shapley values for four causal models with the same observational distribution over features (such that $\mathbb{E}[X_1] = \mathbb{E}[X_2] = 0$ and $\mathbb{E}[X_2|x_1] = \alpha x_1$), yet a different causal structure. Having been trained on combinations of features and corresponding targets, the resulting linear model happens to lead to predictions that ignore the first feature: $f(x_1, x_2) = \beta x_2$. The top table shows the three different patterns that can occur: 'direct' (*D*), 'symmetric' (*S*), and 'asymmetric' (*A*). The bottom table gives for each of the four causal models on the left the marginal, conditional, and causal Shapley values, where the latter two are further split up in symmetric and asymmetric.

fork inverts the arrow between X_1 and X_2, where now Y may represent hotel occupation, X_2 season, and X_1 temperature. In the chain and the fork, different data points correspond to different days. For the confounder and the cycle, X_1 and X_2 may represent obesity and sleep apnea, respectively, and Y hours of sleep. The confounder model implements the assumption that obesity and sleep apnea have a common confounder Z, e.g., some genetic predisposition. The cycle, on the other hand, represents the more common assumption that there is a reciprocal effect, with obesity affecting sleep apnea and vice versa [12]. In the confounder and the cycle, different data points correspond to different subjects. Is this translation to actual "real-world" cases useful/needed?

Given a training set with combinations of features x_1 and x_2 and corresponding targets y, we have trained a linear model $f(x_1, x_2)$ to try and capture the relationship between the features and the target. Suppose that it now so happens that this trained model largely, or even completely to simplify the formulas, ignores the first feature, and boils down to the prediction function $f(x_1, x_2) = \beta x_2$. Figure 1 shows the explanations provided by the various Shapley values for each of the causal models in this extreme situation.

Suggestions to shorten/remove some of the following paragraphs?

Obviously, the marginal Shapley values from [5, 9] do not make any difference between the four causal models and attribute the prediction fully to the direct effect of x_2, not taking any causal relationships between the features into account and hence, in this example, never giving any credit to x_1. The symmetric conditional Shapley values from [1] also do not make a difference between the models, but do assign an indirect effect of size $\frac{1}{2}\beta\alpha x_1$ to x_1 because of its correlation to x_2, which is subtracted from the direct effect of x_2. The factor $\frac{1}{2}$ stems from the fact that symmetric Shapley values incorporate both permutations $(1, 2)$ and $(2, 1)$ for all four models, each with weight $\frac{1}{2}$. For permutation $(1, 2)$, there can be an indirect effect, but for permutation $(2, 1)$ there is none since the prediction does not depend on x_1. Asymmetric Shapley values only incorporate the permutation $(1, 2)$ for the chain and $(2, 1)$ for the fork, both with weight 1. Consequently, for the chain, the indirect part of the asymmetric conditional Shapley value from [4] is twice the indirect part of the symmetric conditional Shapley value. For the fork, the asymmetric conditional Shapley value reduces to the marginal one. Asymmetric conditional Shapley values cannot make a distinction between the

confounder and the cycle: for both they treat x_1 and x_2 on an equal footing and then boil down to symmetric conditional Shapley values.

Causal Shapley values fully take the causal structure into account to compute the total effect of a feature. Although for the fork and the confounder, conditioning by observation leads to an indirect Shapley value, conditioning by intervention does not, since for both models $\mathbb{E}[X_2|do(X_1 = x_1)] = 0$ and causal Shapley values reduce to marginal Shapley values. This is consistent with the argumentation in [5], yet contrary to the asymmetric conditional Shapley values from [4] for the confounder. Where symmetric causal Shapley values are the same for the chain and the cycle, the indirect part of the asymmetric causal Shapley values for the chain is twice that for the cycle.

We would argue that the marginal Shapley values, by only estimating the direct effect, provide a limited view. Consider for example the case of the chain, with X_1 representing season, X_2 temperature, and Y bike rental, and two days with the same temperature of 20 degrees Celsius, one in April (when the temperature is higher than normal for the time of year) and another in August (when it is lower than normal). Marginal Shapley values in the interpretation of [5, 9] treat $f(x_1, x_2) = \beta x_2$ as an arbitrary function and end up with the exact same explanation for the predicted bike rental on both days. Causal Shapley values take into account that $f(x_1, x_2)$ aims to represent $\mathbb{E}[Y|do(\mathbf{X} = \mathbf{x})]$ and then provide different explanations for the two days. Asymmetric causal Shapley values attribute the expected bike rental given the season fully to the Shapley value for season (which in August is higher than in April) and then the difference to the temperature: positive for April and negative for August. Symmetric Shapley values are more conservative, since they incorporate not just the permutation with season set before temperature, but also the permutation with temperature set before season, and give season and temperature each half the credit of the indirect effect.

With the risk of running into a circular argument, symmetric causal Shapley values appear to be a fair generalization of the original idea behind Shapley values. Since for a complex nonlinear function, the effect of adding a feature to the 'in-coalition' features depends on the features that are already part of the coalition, just considering a single permutation leads to an arbitrary result. Similarly, the presumed effect of a feature depends on the order in which interventions are applied, even for a simple linear model. Averaging over all permutations then seems to be the fair procedure, to not accidentally prioritize one feature over another. Asymmetric Shapley values deliberately do prioritize ancestors over their descendants in the causal order. When properly interpreted, this may lead to more natural explanations when the causal order matches the natural order in which features are set to their respective values, e.g., when causal links represent temporal relationships as in one of the examples of [4].Does this argumentation make sense? If so, is this the right place or better somewhere else?

4 Causal chain graphs

Algorithm is out. Still need to decide which theorems/proofs to keep and then adapt the text accordingly by adding the appropriate references and removing repetitions.

Computing causal Shapley values not only requires knowledge of the probability distribution $P(\mathbf{X})$, but also of the underlying causal structure. And even then, there is no guarantee that any causal query is identifiable (see e.g., [14]). Furthermore, and perhaps more importantly, requiring a practitioner to specify a complete causal structure, possibly even including some of its parameters, would be detrimental to the method's general applicability. We therefore follow the same line of reasoning as in [4] and assume that a practitioner may be able to specify a causal ordering, but not much more.

In the special case that a complete causal ordering of the features can be given and that all causal relationships are unconfounded, $P(\mathbf{X})$ satisfies the Markov properties associated with a directed acyclic graph (DAG) and can be written in the form

$$P(\mathbf{X}) = \prod_{j \in N} P(X_j | \mathbf{X}_{pa(j)}) \,,$$

with $pa(j)$ the parents of node j. If no further conditional independences are assumed, the parents of j are all nodes that precede j in the causal ordering. For causal DAGs, we have the interventional formula [7]:

$$P(\mathbf{X}_{\bar{S}} | do(\mathbf{X}_S = \mathbf{x}_S)) = \prod_{j \in \bar{S}} P(X_j | \mathbf{X}_{pa(j) \cap \bar{S}}, \mathbf{x}_{pa(j) \cap S}) \,, \tag{5}$$

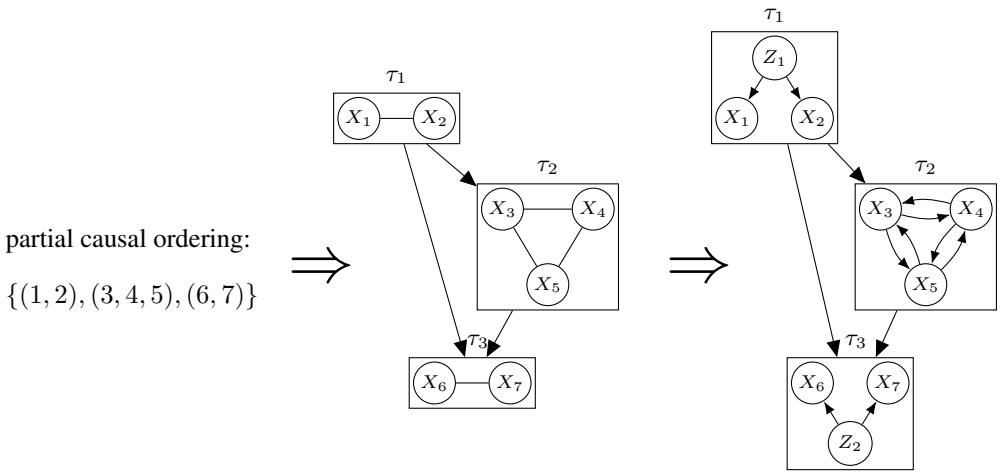


Figure 2: From partial ordering to causal chain graph. Features on equal footing are combined into a fully connected chain component. How to handle interventions within each component depends on the generative process that best explains the (surplus) dependencies. In this example, the dependencies between X_1 and X_2 in chain component τ_1 and X_6 and X_7 in τ_3 are assumed to be the result of a common confounder. The surplus dependencies in τ_2 are assumed to be caused by mutual interactions. Attempt to illustrate the main ideas. Could be nice, but probably not enough space?

with $pa(j) \cap T$ the parents of j that are also part of subset T. The interventional formula can be used to answer any causal query of interest. We will often approximate the expectations needed to compute the Shapley values through sampling, which is particularly straightforward for causal DAGs under conditioning by intervention. Variables are sampled consecutively by following the causal ordering. The probability distribution for a feature then only depends on the values of its parents, which by then is either sampled or fixed. Since the intervention blocks the influence of all descendants, there is no need for an MCMC approach such as Gibbs sampling: the values of all features can be sampled in a single pass through the graph.

We may not always be willing or able to give a complete ordering between the individual variables, but rather a partial ordering, as in Figure 2. Here causal chain graphs [7] come to the rescue. A causal chain graph has directed and undirected edges. All features that are treated on an equal footing are linked together with undirected edges and become part of the same chain component. Edges between chain components are directed and represent causal relationships. The probability distribution $P(\mathbf{X})$ now factorizes as a "DAG of chain components":

$$P(\mathbf{X}) = \prod_{\tau \in \mathcal{T}} P(\mathbf{X}_\tau | \mathbf{X}_{pa(\tau)}) \,,$$

with each τ corresponding to a chain component, consisting of all features that are treated on an equal footing.

How to compute the effect of an intervention now depends on the interpretation of the generative process leading to the (surplus) dependencies between features within each component. If we assume that these are the consequence of marginalizing out a common confounder, intervention on a particular feature will break the dependency with the other features. We will refer to the set of chain components for which this applies as $\mathcal{T}_{\text{confounding}}$. Another possible interpretation is that the undirected part corresponds to the equilibrium distribution of a dynamic process resulting from interactions between the variables within a component [7]. In this case, setting the value of a feature does affect the distribution of the variables within the same component.

Theorem 1. *For causal chain graphs, we have the interventional formula*

$$P(\mathbf{X}_{\bar{S}} | do(\mathbf{X}_S = \mathbf{x}_S)) = \prod_{\tau \in \mathcal{T}_{\text{confounding}}} P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \mathbf{x}_{pa(\tau) \cap S}) \times \prod_{\tau \in \mathcal{T}_{\overline{\text{confounding}}}} P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \mathbf{x}_{pa(\tau) \cap S}, \mathbf{x}_{\tau \cap S}) \,.$$

Proof.

$$
\begin{aligned}
P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S)) &\overset{(1)}{=} \prod_{\tau \in \mathcal{T}} P(\mathbf{X}_{\tau \cap \bar{S}}|\mathbf{X}_{pa(\tau)\cap \bar{S}}, do(\mathbf{X}_S = \mathbf{x}_S)) \\
&\overset{(3)}{=} \prod_{\tau \in \mathcal{T}} P(\mathbf{X}_{\tau \cap \bar{S}}|\mathbf{X}_{pa(\tau)\cap \bar{S}}, do(\mathbf{X}_{pa(\tau)\cap S} = \mathbf{x}_{pa(\tau)\cap S}), do(\mathbf{X}_{\tau \cap S} = \mathbf{x}_{\tau \cap S})) \\
&\overset{(2)}{=} \prod_{\tau \in \mathcal{T}} P(\mathbf{X}_{\tau \cap \bar{S}}|\mathbf{X}_{pa(\tau)\cap \bar{S}}, \mathbf{x}_{pa(\tau)\cap S}, do(\mathbf{X}_{\tau \cap S} = \mathbf{x}_{\tau \cap S})) \,,
\end{aligned}
$$

where the number above each equal sign refers to the standard *do*-calculus rule from [14] that is applied. For a chain component with dependencies induced by a common confounder, rule (3) applies once more and yields $P(\mathbf{X}_{\tau \cap \bar{S}}|\mathbf{X}_{pa(\tau)\cap \bar{S}}, \mathbf{x}_{pa(\tau)\cap S})$, whereas for a chain component with dependencies induced by mutual interactions, rule (2) again applies and gives $P(\mathbf{X}_{\tau \cap \bar{S}}|\mathbf{X}_{pa(\tau)\cap \bar{S}}, \mathbf{x}_{pa(\tau)\cap S}, \mathbf{x}_{\tau \cap S}))$.

Proof probably needs to be extended, which is fine when it goes to the supplement anyway. □

So, to be able to compute the expectations in the Shapley equations under an interventional interpretation, we need to specify (1) a partial order and (2) whether any dependencies between features that are treated on an equal footing are most likely the result of mutual interaction or of a common confounder. Based on this information, any expectation by intervention can be translated to an expectation by observation.

To compute these expectations, we can rely on the various methods that have been proposed to compute conditional Shapley values [1, 4]. Following [1], we will assume a multivariate Gaussian distribution for $P(\mathbf{X})$ that we estimate from the training data. Alternative proposals include assuming a Gaussian copula distribution, estimating from the empirical (conditional) distribution (both from [1]) and a variational autoencoder [4].

5 Experiments

From here on just rough text and ideas.

Show that it works. For now: example on bike rental. Do we predict more bike shares on a warm, but cloudy day in August because of the season or because of the weather? ADNI as another example? Tried German Credit Data, but hard to see differences between causal and conditioning, mainly because the features, such as gender and age, that can be considered causes of some of the others, hardly affect the prediction. Other suggestions? Currently using a relatively straightforward adaptation of the code of [1]. How to describe this? Do we need to publish the code? Do we need to show results for asymmetric Shapley values as well? If so, need to dig deeper into the code. Also: currently no code for handling discrete variables. Could connect to Ruifei's Gaussian copula's for mixed missing data, if needed?

6 Discussion

This paper introduced causal Shapley values, a model-agnostic approach to split a model's prediction of the target variable for an individual data point into contributions of the features that are used as input to the model, where each contribution aims to estimate the total effect of that feature on the target and can be decomposed into a direct and an indirect effect. We contrasted causal Shapley values with (interventional interpretations of) marginal and (asymmetric variants of) conditional Shapley values. We proposed a novel algorithm to compute these causal Shapley values, based on causal chain graphs. All that a practitioner needs to provide is a partial causal order (as for asymmetric Shapley values) and a way to interpret dependencies between features that are on an equal footing. Existing code for computing conditional Shapley values is easily generalized to causal Shapley values, without additional computational complexity. Even when integrated with computationally efficient approaches such as KernelSHAP [10] and TreeExplainer [9], computing conditional and causal Shapley values can be considerably more expensive than computing marginal Shapley values due to the need to sample from conditional instead of marginal distributions.

We considered the typical situation in which the examples provided in the training set are purely observational and a machine learning model has learned to represent $\mathbb{E}[Y|\mathbf{x}]$. To provide a sensible causal explanation, we then need to assume that all features are actual causes of the target values. Future research may consider cases in which some of the features are causes of the target variable and others mere consequences, possibly combined with active scenarios in which feature vectors are actively generated and input to an oracle to obtain the corresponding target variable. Last but not least, user studies should explore to what extent explanations provided by causal Shapley values align with the needs and requirements of practitioners in real-world settings.

Discuss non-manipulable causes as in [15]?

Compare with counterfactual explanations?

Broader Impact

Our research, which aims to provide an explanation for complex machine learning models that can be understood by humans, falls within the scope of explainable AI (XAI). On the positive side, XAI methods like ours can help to open up the infamous "black box" of complicated machine learning models like deep neural networks and decision tree ensembles. A better understanding of the predictions generated by such models may provide higher trust [17], detect flaws and biases [6], and even address the legal "right for an explanation" as formulated in the GDPR [22].

Causality is essential to understanding any process and system, including complex machine learning models. Humans have a strong tendency to reason about their environment in causal terms [19] and causal-model theories fit well to how humans, for example, classify objects [16]. In that sense, explanation approaches like ours, that appeal to a human's capability for causal reasoning could be considered a step in the right direction [11].

Despite their good intentions, explanation methods do come with associated risks. Almost by definition, any sensible explanation of a complex machine learning system involves some simplification. This explanation can give an unjust sense of transparency, sometimes referred to as the 'transparency fallacy' [3]. There is a related risk that model-agnostic general purpose tools like ours will be misused to check a mark in internal or external audits, claiming transparency by just referring to a tool that automatically provides explanations. For now, tools for explainable AI are mostly used as an internal resource by engineers and developers to identify and reconcile errors [2]. A key challenge is to better align explanation methods to the actual needs of external end users.

References

[1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *arXiv preprint arXiv:1903.10464*, 2019.

[2] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 648–657, 2020.

[3] Lilian Edwards and Michael Veale. Slave to the algorithm: Why a right to an explanation is probably not the remedy you are looking for. *Duke L. & Tech. Rev.*, 16:18, 2017.

[4] Christopher Frye, Ilya Feige, and Colin Rowat. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*, 2019.

[5] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causality problem. *arXiv preprint arXiv:1910.13413*, 2019.

[6] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

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

[8] Stan Lipovetsky and Michael Conklin. Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry*, 17(4):319–330, 2001.

[9] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence*, 2(1):2522–5839, 2020.

[10] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[11] Brent Mittelstadt, Chris Russell, and Sandra Wachter. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288, 2019.

[12] Chong Weng Ong, Denise M O'Driscoll, Helen Truby, Matthew T Naughton, and Garun S Hamilton. The reciprocal interaction between obesity and obstructive sleep apnoea. *Sleep medicine reviews*, 17(2):123–131, 2013.

[13] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[14] Judea Pearl. The do-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

[15] Judea Pearl. Does obesity shorten life? Or is it the soda? On non-manipulable causes. *Journal of Causal Inference*, 6(2), 2018.

[16] Bob Rehder. A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6):1141, 2003.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?"s Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[18] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[19] Steven Sloman. *Causal models: How people think about the world and its alternatives*. Oxford University Press, 2005.

[20] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

[21] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. *arXiv preprint arXiv:1908.08474*, 2019.

[22] European Union. Eu general data protection regulation (gdpr): Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), oj 2016 l 119/1, 2016.