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. Shapley values are well calibrated to a user's intuition when features are independent, but may lead to undesirable, counter-intuitive explanations when the independence assumption is violated.

In this paper, we propose a novel framework for computing Shapley values generalizes recent work aiming to either lift or argue for 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 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 [17]. Shapley values are based on a principled mathematical foundation [27] 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., [15, 31]. Recent interests have been triggered by Lundberg and Lee's breakthrough paper [19] that unifies Shapley values and other popular local model-agnostic approaches such as LIME [26], while at the same time introducing more efficient computational procedures.

Humans have a strong tendency to reason in causal terms [28], where explanation and causation are intimately related: explanations often appeal to causes, and causal claims often answers questions about why or how something occurred [16]. The specific domain of causal responsibility studies how people attribute an effect to one or more causes, all of which may have contributed to the observed effect [29]. Causal attributions by humans strongly depend on a subject's understanding of the generative model that explains how different causes lead to the effect, for which the relations between these causes are essential [6].

Most explanation methods, however, act as if features are independent. Even so-called counterfactual approaches, that strongly rely on a causal intuition, make this simplifying assumption (e.g., [33]) and ignore that, in the real world, a change in one input feature may lead to a change in another. This independence assumption also underlies early Shapley-based approaches, such as [31, 3], and is made explicit as an approximation for computational reasons in [19]. 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*. Janzing et al. [8], following [3], discuss a causal interpretation of Shapley values, in which they replace conventional conditioning by observation with conditioning by intervention, as in Pearl's *do*-calculus [23]. This, somewhat surprisingly, leads them to conclude that marginal Shapley values are to be preferred over conditional ones. This argument is also picked up by [17] when implementing interventional Tree SHAP. Finally, Frye et al. [5] propose *asymmetric Shapley values* as a way to incorporate causal knowledge by restricting the possible permutations of the features when computing the Shapley values to those 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 [3, 8, 17] in proposing an active interpretation of Shapley values. (1) Through a different line of reasoning, we will derive 'causal' Shapley values that aim to explain the causal effect of features on the prediction, taking into account their causal relationships, which makes them principally 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 [13], 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

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 assume that we are given a machine learning model $f(\cdot)$ that can generate predictions for any feature vector $\mathbf{x}$. Our goal is to provide an explanation for an individual prediction $f(\mathbf{x})$, that takes into account the causal relationships between the features.

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 to the different features $i \in N$ with $N = \{1, \ldots, n\}$ and n the number of features:

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

where ϕ_i is 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 not knowing any of the feature values. Equation (1) is referred to as the *efficiency property* [27], which appears to be a sensible desideratum for any attribution method and we therefore take here as our starting point.

To go from knowing none of the feature values, as for f_0, to knowing all feature values, as for $f(\mathbf{x})$, we add feature values one by one, actively setting the features to their values in a particular order π. We define the contribution of feature i given 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 $j \preceq_\pi i$ if j precedes i or is equal to i, and where we choose the value function

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

Here S is the subset of 'in-coalition' indices with known feature values $\mathbf{x}_S$. To compute the expectation, we average over the 'out-of-coalition' or dropped features $\mathbf{X}_{\bar{S}}$ with $\bar{S} = N \setminus S$, the complement of S. To explicitly take into account that we actively *set* the features to their values, we condition 'by intervention' for which we resort to Pearl's *do*-calculus [22]. Since the sum over features i in (2) is telescoping, 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 (with shorthand i for the singleton $\{i\}$):

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

Besides efficiency, the Shapley values uniquely satisfy three other desirable properties [27].

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.

Replacing conditioning by intervention with conventional conditioning by observation, i.e., averaging over $P(\mathbf{X}_{\bar{S}}|\mathbf{x}_S)$ instead of $P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S))$ in (3), we arrive at the approach of [1, 18]. A third option is to ignore the feature values $\mathbf{x}_S$ and take the unconditional, marginal distribution $P(\mathbf{X}_{\bar{S}})$. We refer to the corresponding Shapley values as causal, conditional, and marginal, respectively. We will argue that causal Shapley values are the only ones that measure the total effect of an input feature on the model's prediction and reduce to marginal or conditional Shapley values only in special cases.

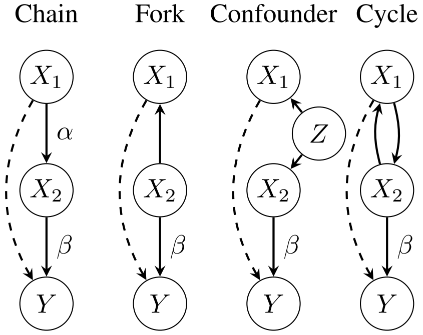
Our active, interventional interpretation of Shapley values from the outset appears to coincide with that in [3, 8, 17]. However, by formally distinguishing between true features (corresponding to one of the data points) and the features plugged as input into the model, Janzing et al. [8] choose to ignore any dependencies between the features in the real world, and 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 and interventional Shapley values in the interpretation of [3, 8, 17] conveniently simplify to marginal Shapley values.

When applied to incorporate causal knowledge, the asymmetric Shapley values introduced in [5] 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 [5] 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 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 $\underline{S} = \{j : j \prec_\pi i\}$

	D		*E*		*R*	
	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*	*E*	*R*	*E*	*R*
Fork	*D*	*E*	*D*	*D*	*D*
Confounder	*D*	*E*	*E*	*D*	*D*
Cycle	*D*	*E*	*E*	*E*	*E*

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. We assume a linear model that happens to ignore the first feature: $f(x_1, x_2) = \beta x_2$. 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. Each letter in the bottom table corresponds to one of the patterns of direct and indirect effects detailed in the top table: 'direct' (*D*, only direct effects), 'evenly split' (*E*, credit for an indirect effect split evenly between the features), and 'root cause' (*R*, all credit for the indirect effect to the root cause).

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. For marginal Shapley values, the indirect effect vanishes: by construction they can 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. We assume to have trained a linear model $f(x_1, x_2)$ that happens to largely, or even completely to simplify the formulas, ignore 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. A derivation can be found in the supplement.

To argue which explanations makes sense in which cases, we follow [20] in calling upon classical norm theory [9]. Classical norm theory states that humans, when asked for an explanation of an effect, contrast the actual observation with a counterfactual, more normal alternative. What is considered normal, depends on the context. Shapley values can be given the exact same interpretation [20]: they measure the difference in prediction between knowing and not knowing the value of a particular feature, where the choice of what's normal translates to the choice of an appropriate reference distribution to average over when the value is still unknown.

In this perspective, marginal Shapley values as in [3, 8, 17] correspond to a very simplistic and even counterintuitive interpretation of what's normal. 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 and another in August. Marginal Shapley values end up with the exact same explanation for the predicted bike rental on both days, completely ignoring that the temperature in April is higher than normal for the time of year and in August lower than normal. Just like marginal Shapley values, symmetric conditional Shapley values as in [1] do not distinguish between any of the four causal structures. They do take into account the dependency between the two features, but then fail to acknowledge that an *intervention* on feature X_1 in the fork and the confounder, does not change distribution of feature X_2.

For the confounder and the cycle, asymmetric Shapley values put X_1 and X_2 on an equal footing and then coincide with their symmetric counterparts. Asymmetric conditional Shapley values from [5] have no means to distinguish between the cycle and the confounder, unrealistically assigning credit to X_1 in the latter case. For the chain and the fork, asymmetric Shapley values only consider the context in which the root cause is set first. This makes that, in our bike rental example of the chain, asymmetric Shapley values first give full credit to season, attributing to temperature only what is left over. Although in general this distribution of credit seems unnecessarily unfair, when dealing with a temporal chain of events, as for example in one of the examples in [5]), it may align with theories on how humans credit causality in a chain of events [30].

When computing the contribution of, for example, X_2, symmetric causal Shapley values always consider two contexts – one in which X_1 is intervened upon before X_2 and one in which X_1 is intervened upon before X_2 – and then average over the results in these two contexts. This strategy appeals to the theory that humans "sample counterfactual scenarios" [7] to estimate causal strength, which dates back to [14]. With the possible exception of asymmetric causal Shapley values for temporal causal structures, the symmetric causal Shapley value are the only ones that lead to intuitive causal explanations in all four models.

4 Causal chain graphs

In the ideal situation, a practitioner has access to a fully specified causal model that can be plugged in (3) to compute or sample from every interventional probability of interest. In practice, such a requirement is hardly realistic. In fact, even if a practitioner could specify a complete causal structure and we have full access to the observational probability $P(\mathbf{X})$, there is no guarantee that any causal query is identifiable (see e.g., [23]). Furthermore, requiring so much prior knowledge could be detrimental to the method's general applicability. In this section, we describe a pragmatic approach that is applicable when we have access to a (partial) causal ordering plus a bit of additional information to distinguish confounders from mutual interactions, as well as a training set to estimate (relevant parameters of) $P(\mathbf{X})$.

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. With no further conditional independences, the parents of j are all nodes that precede j in the causal ordering. For causal DAGs, we have the interventional formula [13]:

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

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.

When we cannot give a complete ordering between the individual variables, but still a partial ordering, causal chain graphs [13] 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. See Figure 2 for an illustration of the procedure. The probability distribution

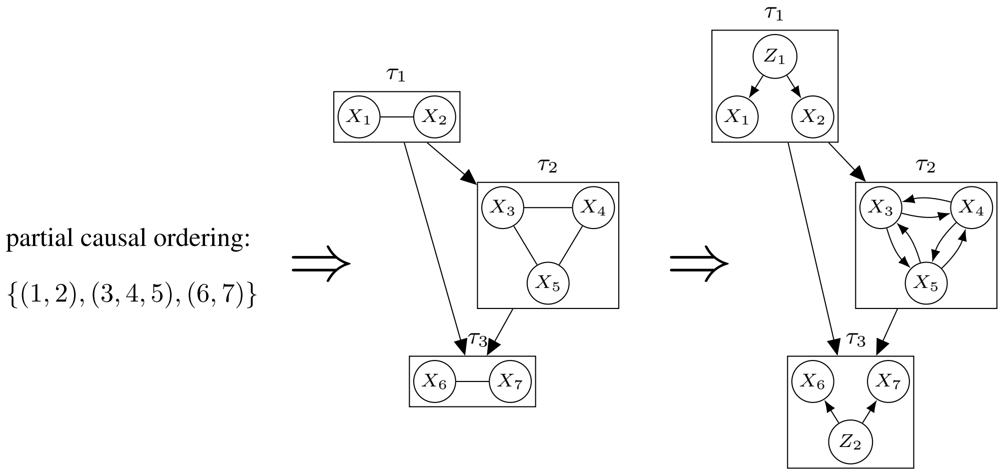


Figure 2: From partial ordering to causal chain graph. Features on equal footing are combined into a fully connected chain component. How to handle interventions within each component depends on the generative process that best explains the (surplus) dependencies. In this example, the dependencies between 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.

$P(\mathbf{X})$ in a chain graph 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 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, as in the confounder in Figure 1, 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 [13], as in the cycle of Figure 1. In this case, setting the value of a feature does affect the distribution of the variables within the same component.

Any expectation by intervention needed to compute the causal Shapley values can be translated to an expectation by observation, by making use of the following theorem (see the supplement for a more detailed proof and some corollaries linking back to other types of Shapley values as special cases).

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

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

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 [23] 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 $P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \mathbf{x}_{pa(\tau) \cap S}, \mathbf{x}_{\tau \cap S}))$. □

To compute these observational expectations, we can rely on the various methods that have been proposed to compute conditional Shapley values [1, 5]. 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 [5].

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. 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, even when integrated with computationally efficient approaches such as KernelSHAP [19] and TreeExplainer [17].

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 [24]?

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 [26], detect flaws and biases [12], and even address the legal "right for an explanation" as formulated in the GDPR [32].

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 [28] and causal-model theories fit well to how humans, for example, classify objects [25]. 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 [21].

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 and hence must sacrifice some accuracy. It is important to better understand what these limitations are [11]. Model-agnostic general purpose explanation tools are often applied without properly understanding their limitations and over-trusted [10], and could possibly even be misused just to check a mark in internal or external audits. Automated explanations can further give an unjust sense of transparency, sometimes referred to as the 'transparency fallacy' [4]. Last but not least, tools for explainable AI are still mostly used as an internal resource by engineers and developers to identify and reconcile errors [2]. A key challenge is therefore 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] Anupam Datta, Shayak Sen, and Yair Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 598–617. IEEE, 2016.

[4] 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.

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

[6] Tobias Gerstenberg, Noah Goodman, David Lagnado, and Joshua Tenenbaum. Noisy Newtons: Unifying process and dependency accounts of causal attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, 2012.

[7] Thomas F Icard, Jonathan F Kominsky, and Joshua Knobe. Normality and actual causal strength. *Cognition*, 161:80–93, 2017.

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

[9] Daniel Kahneman and Dale T Miller. Norm theory: Comparing reality to its alternatives. *Psychological review*, 93(2):136, 1986.

[10] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[11] I Elizabeth Kumar, Suresh Venkatasubramanian, Carlos Scheidegger, and Sorelle Friedler. Problems with shapley-value-based explanations as feature importance measures. *arXiv preprint arXiv:2002.11097*, 2020.

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

[13] 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.

[14] David Lewis. Causation. *The journal of philosophy*, 70(17):556–567, 1974.

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

[16] Tania Lombrozo and Nadya Vasilyeva. Causal explanation. *Oxford Handbook of Causal Reasoning*, pages 415–432, 2017.

[17] 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.

[18] Scott M Lundberg, Gabriel G Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[19] 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.

[20] Luke Merrick and Ankur Taly. The explanation game: Explaining machine learning models with cooperative game theory. *arXiv preprint arXiv:1909.08128*, 2019.

[21] 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.

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

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

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

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

[26] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “Why should I trust you?” 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.

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

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

[29] Elliott Sober. Apportioning causal responsibility. *The Journal of philosophy*, 85(6):303–318, 1988.

[30] Barbara A Spellman. Crediting causality. *Journal of Experimental Psychology: General*, 126(4):323, 1997.

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

[32] 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.

[33] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.