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 the model output and an average baseline output 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 output of a possibly complex and highly non-linear machine learning model. 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 model output for individual data points to the model's features, that can be combined to obtain better global understanding of the model structure [7]. Shapley values are based on a principled mathematical foundation [13] 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., [6, 15]. Recent interests have been triggered by Lundberg and Lee's breakthrough paper [8] that unifies Shapley values and other popular local model-agnostic approaches such as LIME [12], while at the same time introducing more efficient computational procedures.

When applied to explain the output of a machine learning model, Shapley values consider the difference between the model's output when knowing all feature values and its baseline output when knowing none of the feature values and spread 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 [15] estimate this expectation by assuming that the features are independent. This is also the approach taken in [8], 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. [4] 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 [7] 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. [3], 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 [4, 7] in proposing an active interpretation of Shapley values. (1) We will generalize their approach and show that when features are causally related, conditioning by intervention does not reduce to unconditional observational expectations. This makes causal Shapley values truly different from marginal and conditional Shapley values and a more direct way to incorporate causal knowledge, orthogonal to asymmetric Shapley values. (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 [5], we propose a practical approach for computing causal Shapley based 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 data point with feature vector $\mathbf{x}$ and corresponding model output $f(\mathbf{x})$. We compare this output with the average output

$$f_0 = \mathbb{E}f(\mathbf{X}) = \int d\mathbf{X}\, P(\mathbf{X})f(\mathbf{X}) \,,$$

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. To better understand the output of the model for our specific feature vector $\mathbf{x}$, we would like to attribute the difference between $f(\mathbf{x})$ and 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, we would like to 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 output $f(\mathbf{x})$. Equation (1) is referred to as the efficiency property [13], which appears to be a sensible desideratum for any attribution method and we therefore take here as our starting point.

We can think of (at least) two different interpretations on how to go from knowing none of the feature values in f_0 to knowing all feature values in $f(\mathbf{x})$.

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 [9] 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.

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 [13].

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 model output 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.

Our active, interventional interpretation of Shapley values coincides with that in [4, 7]. When all dependencies between features are the result of confounding, conditioning by intervention reduces to no conditioning at all, $P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S)) = P(\mathbf{X}_{\bar{S}})$ for any subset S, and causal Shapley values simplify to marginal Shapley values. However, as we will show in the next sections, when the features are causally related, for example, when one feature drives another or when dependencies between features are better explained through mutual interactions instead of through confounding, the argumentation for unconditional expectations breaks down.

When applied to incorporate causal knowledge, the asymmetric Shapley values introduced in [3] 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 [3] as asymmetric conditional Shapley values, to contrast them with asymmetric causal Shapley values that implement both ideas.

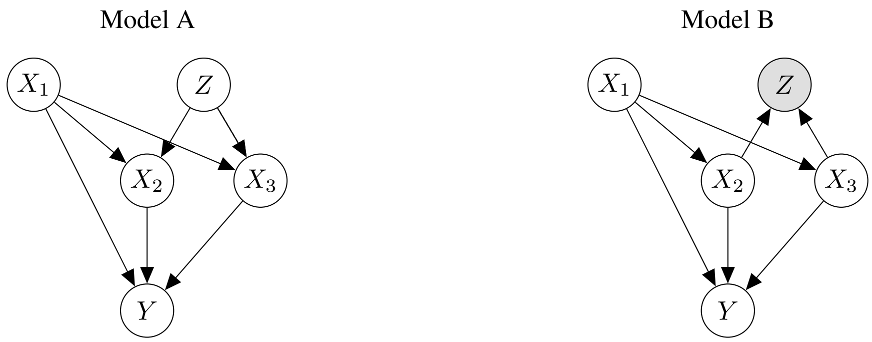

Figure 1: Two causal models. In both, X_1 causes X_2 and X_3. In Model A the excess correlation between X_2 and X_3 is induced by a common confounder Z, in Model B by selection bias.

3 Decomposing Shapley values into direct and indirect effects

Having a causal interpretation of Shapley values, we can decompose our explanation to reflect the contribution of 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. This addition has two effects: a direct effect because now we know the value of x_i and an indirect effect because adding $do(X_i = x_i)$ to the interventional set may change the distribution of the other features. For notational convenience, we write $\underline{S} = \{j : j \prec_\pi i\}$ and $\bar{S} = \{j : j \succ_\pi i\}$, and get:

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

The direct effect measures the expected change in model output 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.

Let $\mathcal{CG}$ denote a causal graph, with $j \succ_{\mathcal{CG}} i$ if and only if there is a causal path from ancestor j to descendant i. For any pair of features (i, j) with $j \succ_{\mathcal{CG}} i$, the indirect effect of feature i through feature j is added to the Shapley value for i, if and only if $j \succ_\pi i$. This goes at the expense of the direct effect of feature j, essentially because when feature j is set to its value, the direct effect is computed relative to the expectation conditioned upon intervention with the 'in-coalition' features, which then necessarily already includes x_i.

Asymmetric (causal) Shapley values only incorporate permutations π with $j \succ_\pi i$ if $j \succ_{\mathcal{CG}} i$. With symmetric causal Shapley values, $j \succ_\pi i$ in half of the permutations π: in the other half feature j is intervened upon before feature i and there is no indirect effect to be accounted for. This makes, as we will see, the indirect part of asymmetric Shapley values roughly a factor two times the indirect part of symmetric Shapley values.

4 Illustration

For illustration, we consider two causal models in Figure 1. They have a different causal structure, but the same dependency structure (all features are dependent) and we assume that the probability distribution $P(\mathbf{X})$ is exactly the same for Model A and Model B. Our estimate of the output Y is a linear function of the features:

$$f(\mathbf{x}) = \beta_0 + \sum_{i=1}^{3} \beta_i x_i \, .$$

Too much detail in this section: move parts to supplement? If so, which parts? Combining (2) and (3), we obtain, after some rewriting

$$\phi_i(\pi) = \beta_i \left(x_i - \mathbb{E}[X_i | op(\mathbf{x}_{j:j \prec_\pi i})] \right) + \sum_{k \succ_\pi i} \beta_k \left(\mathbb{E}[X_k | op(\mathbf{x}_{j:j \preceq_\pi i})] - \mathbb{E}[X_k | op(\mathbf{x}_{j:j \prec_\pi i})] \right) .$$

For marginal Shapley values only the first term before the sum remains, yielding

$$\phi_i = \phi_i(\pi) = \beta_i (x_i - \mathbb{E}[X_i]) ,$$

as also derived in [1].

Analytically computing the conditional Shapley values is tedious, but conceptually straightforward. To write the equations in a compact form, we define $\bar{x}_{k|S} = \mathbb{E}[X_k | \mathbf{x}_S]$ and combine all expectations in a single matrix $\bar{\mathcal{X}}$:

$$\bar{\mathcal{X}} = \begin{pmatrix} x_1 & x_2 & x_3 \\ \bar{x}_1 & \bar{x}_2 & \bar{x}_3 \\ \bar{x}_{1|2} & \bar{x}_{2|1} & \bar{x}_{3|1} \\ \bar{x}_{1|3} & \bar{x}_{2|3} & \bar{x}_{3|2} \\ \bar{x}_{1|2,3} & \bar{x}_{2|1,3} & \bar{x}_{3|1,2} \end{pmatrix} .$$

Any vector $\boldsymbol{\phi}$ with the Shapley values of the three features can be written as a linear combination of these expectations times the regression coefficients $\boldsymbol{\beta}$. With definition of the matrix

$$\mathcal{C}^{\text{conditional}} = \frac{1}{6} \left(\begin{array}{rrrrr|rrrrr|rrrrr} 6 & -2 & -1 & -1 & -2 & 0 & -2 & 2 & -1 & 1 & 0 & -2 & 2 & -1 & 1 \\ 0 & -2 & 2 & -1 & 1 & 6 & -2 & -1 & -1 & -2 & 0 & -2 & -1 & 2 & 1 \\ 0 & -2 & -1 & 2 & 1 & 0 & -2 & -1 & 2 & 1 & 6 & -2 & -1 & -1 & -2 \end{array} \right) ,$$

the conditional Shapley values for any linear model with three variables and a uniform distribution over permutations can be written as

$$\boldsymbol{\phi}^{\text{conditional}} = \mathcal{C}^{\text{conditional}} \, \text{vec}(\text{diag}(\boldsymbol{\beta}) \bar{\mathcal{X}}) . \tag{5}$$

Skip this explanation? Vectorization stacks the columns on top of one another to end up with a 15-dimensional column vector. The vertical bars in the matrix $\mathcal{C}^{\text{conditional}}$ indicate the three blocks, with the first 5 columns in the matrix mapping to the first column of $\bar{\mathcal{X}}$ with expectations of X_1, the next 5 columns to the expectations of X_2, and the final 5 columns to the expectations of X_3.

Skip this paragraph? By summing every column of $\mathcal{C}^{\text{conditional}}$, we can perform the sanity check that efficiency indeed holds. The first column of each block (which relates to the feature values themselves) adds up to 1, the second (corresponding to the marginal expectations) to -1, and the other three columns to zero, as they should. Since Shapley values are constructed by always comparing two (possibly) different expectations, each row within each block sums up to zero.

Putting X_1 before X_2 and X_3, and X_2 and X_3 on equal footing, asymmetric Shapley values only consider the two permutations where x_1 is observed before x_2 and x_3, leading to (we divide by 6 to make it easier to compare with the other Shapley values)

$$\mathcal{C}^{\text{asymmetric}} = \frac{1}{6} \left(\begin{array}{rrrrr|rrrrr|rrrrr} 6 & -6 & 0 & 0 & 0 & 0 & -6 & 6 & 0 & 0 & 0 & -6 & 6 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & 0 & -3 & 0 & -3 & 0 & 0 & -3 & 0 & 3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -3 & 0 & 3 & 6 & 0 & -3 & 0 & -3 \end{array} \right) .$$

Using the standard rules for *do*-calculus [9], we show in Table 1 how the expectations under conditioning by intervention reduce to expectations under conditioning by observation. Since in Model A the correlation between X_2 and X_3 is due to confounding, the interventional expectations and thus Shapley values simplify considerably:

$$\mathcal{C}^{\text{causal,A}} = \frac{1}{6} \left(\begin{array}{rrrrr|rrrrr|rrrrr} 6 & -6 & 0 & 0 & 0 & 0 & -3 & 3 & 0 & 0 & 0 & -3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 6 & -3 & -3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 6 & -3 & 3 & 0 & 0 \end{array} \right) .$$

For Model B, on the other hand, features X_2 and X_3 do affect each other when intervened upon, which makes that compared to the conditional Shapley values only the first block changes:

$$\mathcal{C}^{\text{causal,B}} = \frac{1}{6} \left(\begin{array}{rrrrr|rrrrr|rrrrr} 6 & -6 & 0 & 0 & 0 & 0 & -2 & 2 & -1 & 1 & 0 & -2 & 2 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 6 & -2 & -1 & -1 & -2 & 0 & -2 & -1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & -1 & 2 & 1 & 6 & -2 & -1 & -1 & -2 \end{array} \right) .$$

expectation	model A	model B
$\hat{x}_{1\|2}$	$\bar{x}_1$	
$\hat{x}_{1\|3}$	$\bar{x}_1$	
$\hat{x}_{1\|2,3}$	$\bar{x}_1$	
$\hat{x}_{2\|1}$	$\bar{x}_{2\|1}$	
$\hat{x}_{2\|3}$	$\bar{x}_2$	$\bar{x}_{2\|3}$
$\hat{x}_{2\|1,3}$	$\bar{x}_{2\|1}$	$\bar{x}_{2\|1,3}$
$\hat{x}_{3\|1}$	$\bar{x}_{3\|1}$	
$\hat{x}_{3\|2}$	$\bar{x}_3$	$\bar{x}_{3\|2}$
$\hat{x}_{3\|1,2}$	$\bar{x}_{3\|1}$	$\bar{x}_{3\|1,2}$

Table 1: Turning expectations under conditioning by intervention, $\hat{x}_{i|S} = \mathbb{E}[x_i|do(\mathbf{X}_S = \mathbf{x}_S)]$, into expectations under conventional conditioning by observation, $\bar{x}_{i|S} = \mathbb{E}[x_i|\mathbf{X}_S]$, for the two models in Figure 1. Needed? Can put it next to Figure 1 to save space.

To make this more concrete, let us assume that $P(\mathbf{X})$ follows a multivariate normal distribution corresponding to the causal model

$$X_1 \sim \mathcal{N}(0;1) \text{ and } (X_2, X_3|X_1) \sim \mathcal{N}\left((\alpha_2 X_1, \alpha_3 X_1); \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

where for notational convenience, we chose zero means and unit variance for all noise variables. The first feature drives the second and third feature with coefficients α_2 and α_3. The confounding in Model A or selection bias in Model B leads to correlation ρ on top of the correlation induced by the joint dependence on the first feature. Straightforward calculations yield

$$\bar{\mathcal{X}} = \begin{pmatrix} x_1 & x_2 & x_3 \\ 0 & 0 & 0 \\ \frac{\alpha_2}{\sigma_2^2} x_2 & \alpha_2 x_1 & \alpha_3 x_1 \\ \frac{\alpha_3}{\sigma_3^2} x_3 & \frac{\gamma}{\sigma_3^2} x_3 & \frac{\gamma}{\sigma_2^2} x_2 \\ \frac{\delta_{23} x_2 + \delta_{32} x_3}{1 - \rho^2 + \delta_{23}\alpha_2 + \delta_{32}\alpha_3} & (\alpha_2 - \rho\alpha_3) x_1 + \rho x_3 & (\alpha_3 - \rho\alpha_2) x_1 + \rho x_2 \end{pmatrix}.$$

with $\delta_{ij} = \alpha_i - \rho\alpha_j$, $\sigma_i^2 = 1 + \alpha_i^2$ (the marginal variance for feature i), and $\gamma = \rho + \alpha_2\alpha_3$ (the total covariance between the second and third feature). Plugging this and the expressions for the different $\mathcal{C}$ matrices into (5), we obtain the asymmetric Shapley values

$$\begin{aligned} \phi_1^{\text{asymmetric}} &= \beta_1 x_1 + (\alpha_2\beta_2 + \alpha_3\beta_3) x_1 \\ \phi_2^{\text{asymmetric}} &= \beta_2 x_2 - \alpha_2\beta_2 x_1 + \frac{\rho}{2}(\alpha_3\beta_2 - \alpha_2\beta_3) x_1 + \frac{\rho}{2}(\beta_3 x_2 - \beta_2 x_3) \\ \phi_3^{\text{asymmetric}} &= \beta_3 x_3 - \alpha_3\beta_3 x_1 + \frac{\rho}{2}(\alpha_2\beta_3 - \alpha_3\beta_2) x_1 + \frac{\rho}{2}(\beta_2 x_3 - \beta_3 x_2), \end{aligned}$$

and the causal Shapley values, for Model A,

$$\begin{aligned} \phi_1^{\text{causal,A}} &= \beta_1 x_1 + \frac{1}{2}(\alpha_2\beta_2 + \alpha_3\beta_3) x_1 \\ \phi_2^{\text{causal,A}} &= \beta_2 x_2 - \frac{1}{2}\alpha_2\beta_2 x_1 \\ \phi_3^{\text{causal,A}} &= \beta_3 x_3 - \frac{1}{2}\alpha_3\beta_3 x_1. \end{aligned}$$

and, for Model B, (really ugly…can we prevent this?)

$$
\begin{aligned}
\phi_1^{\text{causal, B}} &= \beta_1 x_1 + \frac{1}{2}(\alpha_2\beta_2 + \alpha_3\beta_3)x_1 - \frac{\rho}{6}(\alpha_3\beta_2 + \alpha_2\beta_3)x_1 - \frac{1}{6}\left(\frac{\alpha_3\delta_{23}}{\sigma_3^2}\beta_2 x_3 + \frac{\alpha_2\delta_{32}}{\sigma_2^2}\beta_3 x_2\right) \\
\phi_2^{\text{causal, B}} &= \beta_2 x_2 - \frac{1}{2}\alpha_2\beta_2 x_1 + \frac{\rho}{6}(2\alpha_3\beta_2 - \alpha_2\beta_3)x_1 + \\
&\qquad \frac{1}{6}\left(2\frac{\alpha_2\delta_{32}}{\sigma_2^2}\beta_3 x_2 - \frac{\alpha_3\delta_{23}}{\sigma_3^2}\beta_2 x_3\right) + \frac{\rho}{2}(\beta_3 x_2 - \beta_2 x_3) \\
\phi_3^{\text{causal, B}} &= \beta_3 x_3 - \frac{1}{2}\alpha_3\beta_3 x_1 + \frac{\rho}{6}(2\alpha_2\beta_3 - \alpha_3\beta_2)x_1 + \\
&\qquad \frac{1}{6}\left(2\frac{\alpha_3\delta_{23}}{\sigma_3^2}\beta_2 x_3 - \frac{\alpha_2\delta_{32}}{\sigma_2^2}\beta_3 x_2\right) + \frac{\rho}{2}(\beta_2 x_3 - \beta_3 x_2) \,.
\end{aligned}
$$

In this linear model, the asymmetric Shapley value for the first feature adds its indirect causal effects on the output through the second and third feature, $\alpha_2\beta_2 x_1 + \alpha_3\beta_3 x_1$, to its direct effect, $\beta_1 x_1$. The causal Shapley values for the first feature are somewhat more conservative: they essentially claim only half of the indirect effects through the other two features. Move the rest of this paragraph elsewhere? Now overlap with direct/indirect in next section. This is a direct consequence of taking a uniform distribution over all permutations: for any pair of features i and j, the feature value x_i is set before and after x_j for exactly half of the number of permutations. Which distribution over permutations to prefer, a uniform one or one that respects the causal structure, depends on the question the practitioner tries to answer and possibly on the application. For example, when a causal link represents a temporal relationship, it may make no sense to set a feature value before the values of all features preceding it in time have been set. In that case, it would be wise to consider a non-uniform distribution over permutations as in [3]. On the other hand, for causal models without temporal interpretation, e.g., describing presumed causal relationships between personal and biomedical variables related to Alzheimer [14] or between social and economic characteristics in census data [2], deviating from a uniform distribution over permutations (and hence sacrificing the symmetry property) seems unnecessary. With or without uniform distribution over permutations, applying *do*-calculus instead of conditioning by observation is a natural way to incorporate causal information.

The Shapley values for Model A are different from those for Model B, even though the observable probability distribution $P(\mathbf{X})$ is exactly the same. Those for Model A simplify a lot, because in this model any excess correlation between X_2 and X_3 beyond the correlation resulting from the common parent X_1 results from a confounder. This correlation vanishes when we intervene on either X_2 or X_3. The contributions of the second and third feature are therefore just their direct effect, minus half of the indirect effect, which already has been attributed to the first feature. In Model B, on the other hand, for most expectations conditioning by intervention reduces to conditioning by observation on the same variables and does not further simplify to conditioning on less or even no variables as for Model A. Since the causal Shapley values consider all six permutations, in contrast to the asymmetric Shapley values which only consider two of them, expectations such as $\mathbb{E}[X_2|x_3]$ now also enter the equation, which considerably complicates the analytical expressions.

5 Causal chain graphs

Added a theorem, corollaries, a figure, and an algorithm. Still need to decide which ones 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., [10]). For example, if Model A or B also includes a causal link from X_2 to X_3, even knowing the probability distribution $P(\mathbf{X})$ and the causal structure is insufficient: it is impossible to express, for example, $P(X_3|do(X_2 = x_2))$ in terms of $P(\mathbf{X})$, essentially because, without knowing the parameters of the causal model, there is no way to tell which part of the observed dependence between X_2 and X_3 is due to the causal link and which due to the confounding or selection bias.

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

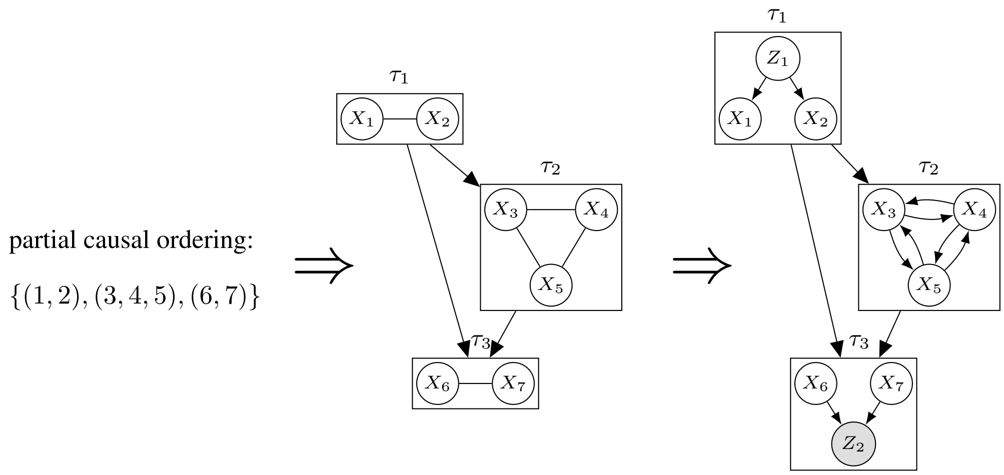

Figure 2: From partial ordering to causal chain graph. Features on equal footing are combined into a fully connected chain component. How to handle interventions within each component depends on the generative process that best explains the (surplus) dependencies. In this example, the dependency between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

general applicability. We therefore follow the same line of reasoning as in [3] 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 [5]:

$$P(\mathbf{X}_{\bar{S}} | do(\mathbf{X}_S = \mathbf{x}_S)) = \prod_{j \in \bar{S}} P(X_j | \mathbf{X}_{pa(j) \cap \bar{S}}, \mathbf{x}_{pa(j) \cap S}) \,, \tag{6}$$

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, for example, in Figure 1 where we have the partial ordering $(\{1\},\{2,3\})$: the first feature precedes the second and third feature in the causal ordering, with the second and third feature on equal footing, i.e., without specifying whether the second causes the third or vice versa. Here causal chain graphs [5] 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, as in Model A 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 [5]. In this case, setting the value of a feature does affect the distribution of the variables within the same component. The same applies to the case of selection bias, as in Model B in Figure 1.

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

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

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 [10] 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 selection bias or mutual interactions, rule (2) again applies:

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

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

Theorem 1 connects to observations made and algorithms proposed in recent papers.

Corollary 2. *With all features combined in a single component and all dependencies induced by confounding, as in [4], causal Shapley values are equivalent to marginal Shapley values.*

Proof. We immediately get $P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S)) = P(\mathbf{X}_{\bar{S}})$ for all subsets S, i.e., as if all features are independent. □

Corollary 3. *With all features combined in a single component and all dependencies induced by selection bias or mutual interactions, causal Shapley values are equivalent to conditional Shapley values as proposed in [1].*

Proof. Now $P(\mathbf{X}_{\bar{S}}|do(\mathbf{X}_S = \mathbf{x}_S)) = P(\mathbf{X}_{\bar{S}}|\mathbf{x}_S)$ for all subsets S, which boils down to conventional conditioning by observation. □

Corollary 4. *When we only take into account permutations that match the causal ordering and assume that all dependencies within chain components are induced by selection bias or mutual interactions, the resulting asymmetric causal Shapley values are equivalent to the asymmetric conditional Shapley values as defined in [3].*

Algorithm 1 Compute the value function $v(S)$ under conditioning by intervention. Include in main text? Seems rather obvious…

function VALUEFUNCTION(S)
  for $k \leftarrow 1$ to N_{samples} **do**
    for all $j \leftarrow 1$ to $|\mathcal{T}|$ **do** ▷ run over all chain components in their causal order
      if confounding(τ_j) **then**
        for all $i \in \tau_j \cap \bar{S}$ **do**
          Sample $\tilde{x}_i^{(k)} \sim P(X_i | \tilde{\mathbf{x}}^{(k)}_{pa(\tau_j) \cap \bar{S}}, \mathbf{x}_{pa(\tau_j) \cap \bar{S}})$ ▷ can be drawn independently
        end for
      else
        Sample $\tilde{\mathbf{x}}^{(k)}_{\tau_j \cap \bar{S}} \sim P(\mathbf{X}_{\tau_j \cap \bar{S}} | \tilde{\mathbf{x}}^{(k)}_{pa(\tau_j) \cap \bar{S}}, \mathbf{x}_{pa(\tau_j) \cap \bar{S}}, \mathbf{x}_{\tau_j \cap S})$ ▷ e.g., Gibbs sampling
      end if
    end for
  end for
  $v \leftarrow \dfrac{1}{N_{\text{samples}}} \displaystyle\sum_{k=1}^{N_{\text{samples}}} f(\mathbf{x}_S, \tilde{\mathbf{x}}^{(k)}_{\bar{S}})$
  return v
end function

Proof. Following [3], asymmetric Shapley values only include those permutations π for which $i \prec_\pi j$ for all known ancestors i of descendants j in the causal graph. For those permutations, we are guaranteed to have $\tau \prec_{\mathcal{CG}} \tau'$ for all $\tau, \tau' \in \mathcal{T}$ such that $\tau \cap S \neq \emptyset, \tau' \cap \bar{S} \neq \emptyset$. That is, the chain components that contain features from S are never later in the causal ordering of the chain graph $\mathcal{CG}$ than those that contain features from $\bar{S}$. We then have

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

where in the last step we used the fact that $\mathcal{T}_{\text{confounding}} = \emptyset$. □

Conditioning by intervention in causal chain graphs boils down to conditioning by observation, where features within components that are later in the causal ordering are excluded from the conditioning set. The asymmetric Shapley values in [3], since still based on conditioning by observation, need to choose a non-uniform distribution over permutations to achieve the same. Our analysis shows that this restriction to permutations that are consistent with the causal ordering is not needed and can be considered a separate choice, orthogonal to conditioning by observation or by intervention.

Turn the above into a Theorem? Bit hard to make concrete.

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 selection, 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, 3]. 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 [3].

6 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?

7 Discussion

Conditional and causal Shapley values attribute the difference between the model output and the expected model output to the individual features assuming that their values are passively *observed* and actively *set*, respectively. Both interpretations are valid, but only the latter enables us to leverage additional available information regarding the causal structure.

Marginal Shapley values are perfectly fine when dependencies are purely the result of confounding, as argued in [4]. However, when these dependencies are the result of causal relationships or even due to selection bias or mutual feedback, then expectations under conditioning by intervention do not simplify to marginal expectations.

We proposed a novel algorithm to compute 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 equal footing. Conditioning by intervention becomes conditioning by observation, but only on ancestors and possibly, depending on the interpretation, on features within the same component. Any existing computational approach for conditional Shapley values can be easily adapted and combined with computationally efficient approaches such as KernelSHAP [8] and TreeExplainer [7]. If anything the expectations simplify, since there is no conditioning by observation on the descendants.

Our analysis may also provide a better understanding of asymmetric Shapley values: for all permutations that are consistent with the causal ordering, conditioning by intervention boils down to conditioning by observation. However, the current definition in [3] implicitly assumes that dependencies between features that are on equal footing is due to mutual feedback or selection bias, not common confounding. Whether or not to only consider permutations that match the causal ordering depends on the application and possibly one's preference. This paper shows that it is unnecessary to forgo symmetry in order to arrive at a causal interpretation of Shapley values. Roughly speaking, asymmetric (causal/conditioning) Shapley values attribute all indirect effects of a causal variable through a mediator on the output to the causal variable (is this the right term?), subtracting them from the Shapley value of the mediator. Symmetric Shapley values are more conservative and attribute only half to the causal variable.

Discuss non-manipulable causes as in [11]?

Compare with counterfactual explanations?

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] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7801–7808, 2019.

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

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

[5] 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.

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

[7] 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.

[8] 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.

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

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

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

[12] 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.

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

[14] Xinpeng Shen, Sisi Ma, Prashanthi Vemuri, and Gyorgy Simon. Challenges and opportunities with causal discovery algorithms: Application to Alzheimer's pathophysiology. *Scientific reports*, 10(1):1–12, 2020.

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