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

Anonymous Author(s)

Affiliation

Address

email

1 Do-calculus for cyclic graphs

For completeness, we here repeat the rules of *do*-calculus for cyclic graphs, in the notation of [2], which generalizes [5]. We are given a causal graph G . To each node X_i which is intervened upon, we add an ‘intervention node’ I_{X_i} , with a directed edge from I_{X_i} to X_i that we clamp to the value x_i . The corresponding graph is called \hat{G}^+ . $\hat{G}_{do(\mathbf{W})}$ is now obtained by removing from \hat{G}^+ all incoming edges to variables that are part of \mathbf{W} , except those from the corresponding intervention nodes $I_{\mathbf{W}}$. We use shorthand

$$\mathbf{Y} \perp\!\!\!\perp_G^{\sigma} \mathbf{X} \mid \mathbf{Z}, do(\mathbf{W})$$

to indicate that \mathbf{Y} and \mathbf{X} are σ -separated by \mathbf{Z} in the graph $\hat{G}_{do(\mathbf{W})}$. σ -separation is a generalization of standard d-separation (see [2] for details).

Do-calculus now consists of the following three inference rules that can be used to map interventional and observational distributions.

1. Insertion/deletion of observation:

$$P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, do(\mathbf{W})) = P(\mathbf{Y}|\mathbf{Z}, do(\mathbf{W})) \text{ if } \mathbf{Y} \perp\!\!\!\perp_G^{\sigma} \mathbf{X} \mid \mathbf{Z}, do(\mathbf{W}).$$

2. Action/observation exchange:

$$P(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z}, do(\mathbf{W})) = P(\mathbf{Y}|\mathbf{X}, \mathbf{Z}, do(\mathbf{W})) \text{ if } \mathbf{Y} \perp\!\!\!\perp_G^{\sigma} I_{\mathbf{X}} \mid \mathbf{X}, \mathbf{Z}, do(\mathbf{W}).$$

3. Insertion/deletion of actions:

$$P(\mathbf{Y}|do(\mathbf{X}), \mathbf{Z}, do(\mathbf{W})) = P(\mathbf{Y}|\mathbf{Z}, do(\mathbf{W})) \text{ if } \mathbf{Y} \perp\!\!\!\perp_G^{\sigma} I_{\mathbf{X}} \mid \mathbf{Z}, do(\mathbf{W}).$$

Through consecutive application of these rules, we can try to turn any interventional probability of interest into an observational probability.

2 Shapley values for linear models

We will make use of the *do*-calculus rules above to derive the causal Shapley values for the four different models in Figure 1 in the main text. To this end, we consider the three models in Figure 1 that predict $f(x_1, x_2) = \beta_1 x_1 + \beta_2 x_2$ for general values of β_1 and β_2 . All three models have the same observational probability distribution, with $\mathbb{E}[X_i] = \bar{x}_i$ and $\mathbb{E}[X_{3-i}] = \alpha_i x_i$, for $i = 1, 2$, yet different causal structures. We will arrive at the main text’s results for the ‘chain’, ‘confounder’, and ‘cycle’

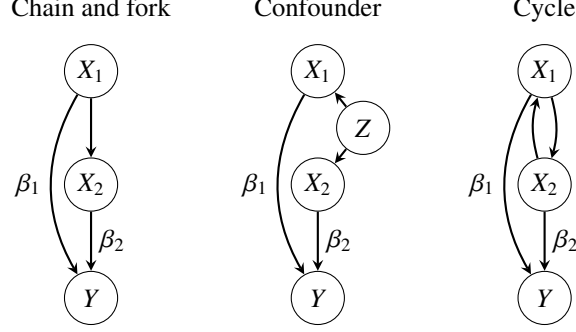


Figure 1: Three causal models with the same observational distribution over features, yet a different causal structure. To connect to the models in the main text, we set $\beta_1 = 0$ and $\beta_2 = \beta$, except that for the ‘fork’ we set $\beta_2 = 0$, $\beta_1 = \beta$, and then swap the indices.

by setting $\beta_1 = 0$, whereas for the ‘fork’ we set $\beta_2 = 0$ and swap the two indices. We then further need to take $\bar{x}_1 = \bar{x}_2 = 0$, and $\alpha = \alpha_2$.

Following the definitions in the main text, the contribution of feature i given permutation π is the difference in value function before and after setting the feature to its value:

$$\phi_i(\pi) = v(\{j : j \leq_\pi i\}) - v(\{j : j <_\pi i\}), \quad (1)$$

with value function

$$v(S) = \mathbb{E}[f(\mathbf{X}) | do(\mathbf{X}_S = \mathbf{x}_S)] = \int d\mathbf{X}_{\bar{S}} P(\mathbf{X}_{\bar{S}} | \hat{\mathbf{x}}_S) f(\mathbf{X}_{\bar{S}}, \mathbf{x}_S), \quad (2)$$

where we use shorthand $\hat{\mathbf{x}}$ for $do(\mathbf{X} = \mathbf{x})$. Combining these two definitions and substituting $f(\mathbf{x}) = \sum_i \beta_i x_i$, we obtain

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

The first term corresponds to the direct effect, the second one to the indirect effect. Symmetric causal Shapley values will follow by averaging these contributions for the two possible permutations $\pi = (1, 2)$ and $\pi = (2, 1)$. Conditional Shapley values result when replacing conditioning by intervention with conventional conditioning by observation, marginal Shapley values by not conditioning at all.

To start with the latter, we immediately see that for *marginal Shapley values* the indirect effect vanishes and the direct effect simplifies to

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

as also derived in [1].

For symmetric conditional Shapley values, we do get different contributions for the two different permutations, but by definition still the same for the three different models:

$$\begin{aligned} \phi_1(1, 2) &= \beta_1(x_1 - \mathbb{E}[X_1]) + \beta_2(\mathbb{E}[X_2 | x_1] - \mathbb{E}[X_2]) = \beta_1(x_1 - \bar{x}_1) + \beta_2\alpha_1(x_1 - \bar{x}_1) \\ \phi_2(1, 2) &= \beta_2(x_2 - \mathbb{E}[X_2 | x_1]) = \beta_2(x_2 - \bar{x}_2) - \beta_2\alpha_1(x_1 - \bar{x}_1). \end{aligned} \quad (3)$$

Here the first term in the contribution for the first feature corresponds to the direct effect and the second term to the indirect effect. The contribution for the second feature only consists of a direct effect. The contributions for the other permutation follow by swapping the indices and the final Shapley values by averaging to arrive at the *symmetric conditional Shapley values*

$$\begin{aligned} \phi_1 &= \beta_1(x_1 - \bar{x}_1) - \frac{1}{2}\beta_1\alpha_2(x_2 - \bar{x}_2) + \frac{1}{2}\beta_2\alpha_1(x_1 - \bar{x}_1) \\ \phi_2 &= \beta_2(x_2 - \bar{x}_2) - \frac{1}{2}\beta_2\alpha_1(x_1 - \bar{x}_1) + \frac{1}{2}\beta_1\alpha_2(x_2 - \bar{x}_2), \end{aligned} \quad (4)$$

where now the first two terms constitute the direct effect and the third term the indirect effect.

expectation	chain	confounder	cycle
$\mathbb{E}[X_1 \hat{x}_2]$	$\mathbb{E}[X_1]$	$\mathbb{E}[X_1]$	$\mathbb{E}[X_1 x_2]$
$\mathbb{E}[X_2 \hat{x}_1]$	$\mathbb{E}[X_2 x_1]$	$\mathbb{E}[X_2]$	$\mathbb{E}[X_2 x_1]$

Table 1: Turning expectations under conditioning by intervention into expectations under conventional conditioning by observation for the three models in Figure 1.

The asymmetric conditional Shapley values consider both permutations for the confounder and the cycle, and hence are equivalent to the symmetric Shapley values for those models. Yet for the chain, they only consider the permutation $\pi(1, 2)$ and thus yield $\phi = \phi(1, 2)$ from (5).

To go from the symmetric conditional Shapley values to the causal symmetric Shapley values, we follow the same line of reasoning, but have to replace $\mathbb{E}[X_2|x_1]$ by $\mathbb{E}[X_2|\hat{x}_1]$ and $\mathbb{E}[X_1|x_2]$ by $\mathbb{E}[X_1|\hat{x}_2]$. Table 1 tells whether the expectations under conditioning by intervention reduce to expectations under conditioning by observation (because of the second rule of *do*-calculus above) or to marginal expectations (because of the third rule). For the chain we have

$$P(X_2|\hat{x}_1) = P(X_2|x_1) \text{ since } X_2 \perp\!\!\!\perp_G I_{X_1} | X_1 \text{ (rule 2), yet } P(X_1|\hat{x}_2) = P(X_1) \text{ since } X_1 \perp\!\!\!\perp_G I_{X_2} \text{ (rule 3),}$$

for the confounder

$$P(X_2|\hat{x}_1) = P(X_2) \text{ since } X_2 \perp\!\!\!\perp_G I_{X_1} \text{ and } P(X_1|\hat{x}_2) = P(X_1) \text{ since } X_1 \perp\!\!\!\perp_G I_{X_2} \text{ (rule 3),}$$

and for the cycle

$$P(X_2|\hat{x}_1) = P(X_2|x_1) \text{ since } X_2 \perp\!\!\!\perp_G I_{X_1} | X_1 \text{ and } P(X_1|\hat{x}_2) = P(X_1|x_2) \text{ since } X_1 \perp\!\!\!\perp_G I_{X_2} | X_2 \text{ (rule 2).}$$

Consequently, for the confounder the symmetric and asymmetric causal Shapley values coincide with the marginal Shapley values (consistent with [4]) and for the cycle with the symmetric conditional Shapley values from (4). For the chain, the causal symmetric Shapley values become

$$\begin{aligned} \phi_1(1, 2) &= \beta_1(x_1 - \bar{x}_1) + \frac{1}{2}\beta_2\alpha_1(x_1 - \bar{x}_1) \\ \phi_2(1, 2) &= \beta_2(x_2 - \bar{x}_2) - \frac{1}{2}\beta_2\alpha_1(x_1 - \bar{x}_1), \end{aligned} \tag{5}$$

where the asymmetric causal Shapley values coincides with the asymmetric conditional Shapley values from (5).

Collecting all results and setting $\bar{x}_1 = \bar{x}_2 = \beta_1 = 0$, $\beta_2 = \beta$, and $\alpha_1 = \alpha$ (after swapping the indices for the ‘fork’), we arrive at the Shapley values reported in Figure 1 in the main text. Note that for most Shapley values, the indirect effect for the second feature vanishes because we chose to set $\beta_1 = 0$. The exceptions, apart from the marginal Shapley values, are the causal Shapley values for the chain and the confounder, as well as the asymmetric conditional Shapley values for the chain: these show no indirect effect for the second feature even for nonzero β_1 .

3 Proofs and corollaries on causal chain graphs

In this section we expand on the proof of Theorem 1 in the main text and add some corollaries to link back to other approaches for computing Shapley values.

The probability distribution for a causal chain graph reads

$$P(\mathbf{X}) = \prod_{\tau \in \mathcal{T}} P(\mathbf{X}_\tau | \mathbf{X}_{pa(\tau)}). \tag{6}$$

For each chain component, we further need to specify whether (surplus) dependencies within the component are due to confounding or due to mutual interactions. Given this information, we can turn any causal query into an observational distribution with the following interventional formula.

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

$$P(\mathbf{X}_{\bar{S}} | do(\mathbf{X}_S = \mathbf{x}_S)) = \prod_{\tau \in \mathcal{T}_{\text{confounding}}} P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \mathbf{x}_{pa(\tau) \cap S}) \times \prod_{\tau \in \overline{\mathcal{T}_{\text{confounding}}}} P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \mathbf{x}_{pa(\tau) \cap S}, \mathbf{x}_{\tau \cap S}). \quad (7)$$

73 *Proof.* Plugging in (6) and using shorthand $\hat{\mathbf{x}} = do(\mathbf{X} = \mathbf{x})$, we obtain

$$P(\mathbf{X}_{\bar{S}} | \hat{\mathbf{x}}_S) = P(\mathbf{X} | \hat{\mathbf{x}}_S) \stackrel{(1)}{=} \prod_{\tau \in \mathcal{T}} P(\mathbf{X}_{\tau} | \mathbf{X}_{pa(\tau)}, \hat{\mathbf{x}}_S) = \prod_{\tau \in \mathcal{T}} P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \hat{\mathbf{x}}_S)$$

74 where in the second step we implicitly made use of *do*-calculus rule (1): the conditional independen-
75 cies in the causal chain graph are preserved when we intervene on some of the variables.

76 Now rule (3) tells us that we can ignore any interventions from nodes in components further down
77 the causal chain graph as well as those from higher up that are shielded by the direct parents:

$$P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \hat{\mathbf{x}}_S) \stackrel{(3)}{=} P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \hat{\mathbf{x}}_{pa(\tau) \cap S}, \hat{\mathbf{x}}_{\tau \cap S}).$$

78 Rule (2) then states that conditioning by intervention upon variables higher up in the causal chain
79 graph is equivalent to conditioning by observation:

$$P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \hat{\mathbf{x}}_{pa(\tau) \cap S}, \hat{\mathbf{x}}_{\tau \cap S}) \stackrel{(2)}{=} P(\mathbf{X}_{\tau \cap \bar{S}} | \mathbf{X}_{pa(\tau) \cap \bar{S}}, \mathbf{x}_{pa(\tau) \cap S}, \hat{\mathbf{x}}_{\tau \cap S}).$$

80 For a chain component with dependencies induced by a common confounder, rule (3) applies once
81 more and makes that we can ignore the interventions:

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

82 For a chain component with dependencies induced by mutual interactions, rule (2) again applies and
83 allows us to replace conditioning by intervention with conditioning by observation:

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

84 □

85 Algorithm 1 provides pseudocode on how to estimate the value function $v(S)$ by drawing samples
86 from the interventional probability (7). It assumes that a user has specified a partial causal ordering
87 of the features, which is translated to a chain graph with components \mathcal{T} , and for each (non-singleton)
88 component τ whether or not surplus dependencies are the result of confounding. Other prerequisites
89 include access to the model $f(\cdot)$, the feature vector \mathbf{x} , (a procedure to sample from) the observational
90 probability distribution $P(\mathbf{X})$, and the number of samples N_{samples} .

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

92 **Corollary 1.** With all features combined in a single component and all dependencies induced by
93 confounding, as in [4], causal Shapley values are equivalent to marginal Shapley values.

94 *Proof.* *Do*-calculus rule (3) yields $P(\mathbf{X}_{\bar{S}} | \hat{\mathbf{x}}_S) = P(\mathbf{X}_{\bar{S}})$ for all subsets S , i.e., as if all features are
95 independent. □

96 **Corollary 2.** With all features combined in a single component and all dependencies induced by
97 mutual interactions, causal Shapley values are equivalent to conditional Shapley values as proposed
98 in [1].

99 *Proof.* Now, *do*-calculus rule (2) applies and gives $P(\mathbf{X}_{\bar{S}} | \hat{\mathbf{x}}_S) = P(\mathbf{X}_{\bar{S}} | \mathbf{x}_S)$ for all subsets S , which
100 boils down to conventional conditioning by observation. □

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

Algorithm 1 Compute the value function $v(S)$ under conditioning by intervention.

```

1: function VALUEFUNCTION( $S$ )
2:   for  $k \leftarrow 1$  to  $N_{\text{samples}}$  do
3:     for all  $j \leftarrow 1$  to  $|\mathcal{T}|$  do ▷ run over all chain components in their causal order
4:       if confounding( $\tau_j$ ) then
5:         for all  $i \in \tau_j \cap \bar{S}$  do
6:           Sample  $\tilde{x}_i^{(k)} \sim P(X_i | \tilde{\mathbf{x}}_{pa(\tau_j) \cap \bar{S}}, \mathbf{x}_{pa(\tau_j) \cap S})$  ▷ can be drawn independently
7:         end for
8:       else
9:         Sample  $\tilde{\mathbf{x}}_{\tau_j \cap \bar{S}}^{(k)} \sim P(\mathbf{X}_{\tau_j \cap \bar{S}} | \tilde{\mathbf{x}}_{pa(\tau_j) \cap \bar{S}}, \mathbf{x}_{pa(\tau_j) \cap S}, \mathbf{x}_{\tau_j \cap S})$  ▷ e.g., Gibbs sampling
10:      end if
11:    end for
12:  end for
13:   $v \leftarrow \frac{1}{N_{\text{samples}}} \sum_{k=1}^{N_{\text{samples}}} f(\mathbf{x}_S, \tilde{\mathbf{x}}_{\bar{S}}^{(k)})$ 
14:  return  $v$ 
15: end function

```

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

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

110 where in the last step we used interventional formula (7) in combination with the fact that $\mathcal{T}_{\text{confounding}} =$
 111 \emptyset . □

112 4 Additional illustrations on the bike rental data

113 Figure 2 shows sina plots for asymmetric conditional Shapley values (left) and marginal Shapley
 114 values (right), to be compared with the sina plots for symmetric causal Shapley values in Figure 3
 115 of the main text. In this case, the sina plots for asymmetric causal Shapley values are virtually
 116 indistinguishable from those for the asymmetric conditional Shapley values.

117 It can be seen that the marginal Shapley values strongly focus on temperature, largely ignoring the
 118 seasonal variables. The asymmetric Shapley values do the opposite: they focus on the seasonal
 119 variables, in particular *cosyear* and put much less emphasis on the temperature variables.

120 References

- 121 [1] Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when
 122 features are dependent: More accurate approximations to Shapley values. *arXiv preprint*
 123 *arXiv:1903.10464*, 2019.
- 124 [2] Patrick Forré and Joris M Mooij. Causal calculus in the presence of cycles, latent confounders
 125 and selection bias. In *Proceedings of the 35th Annual Conference on Uncertainty in Artificial*
 126 *Intelligence*, 2019.
- 127 [3] Christopher Frye, Ilya Feige, and Colin Rowat. Asymmetric Shapley values: incorporating causal
 128 knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*, 2019.
- 129 [4] Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in
 130 explainable AI: A causality problem. *arXiv preprint arXiv:1910.13413*, 2019.
- 131 [5] Judea Pearl. The *do*-calculus revisited. *arXiv preprint arXiv:1210.4852*, 2012.

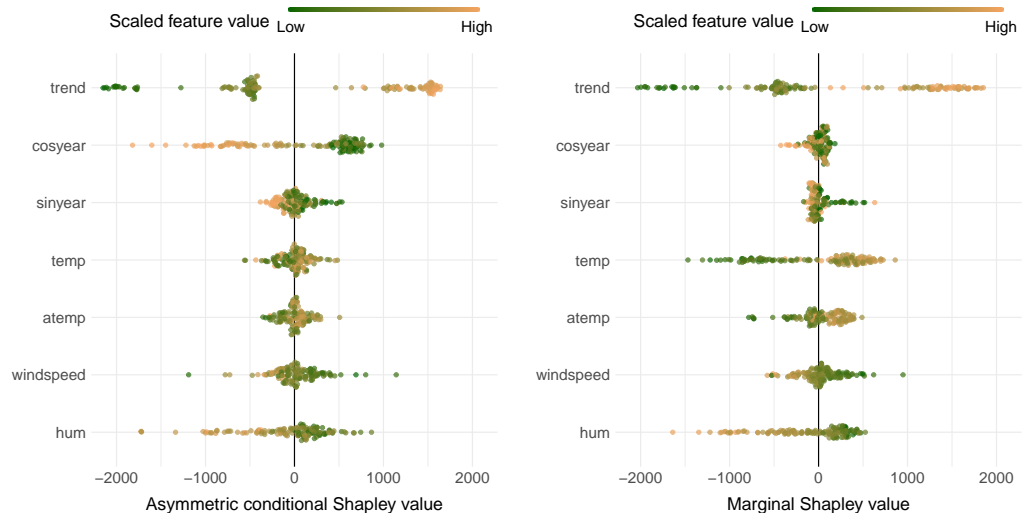


Figure 2: Sina plots of asymmetric (conditional) Shapley values (left) and marginal Shapley values (right). See Figure 3 in the main text for further details.