PWSHAP: A Path-Wise Explanation Model for Targeted Variables

Lucile Ter-Minassian [*1] **Oscar Clivio** [*1] **Karla Diaz-Ordaz** [2] **Robin J. Evans** [1] **Chris Holmes** [1 3]

Abstract

Predictive black-box models can exhibit high-accuracy but their opaque nature hinders their uptake in safety-critical deployment environments. Explanation methods (XAI) can provide confidence for decision-making through increased transparency. However, existing XAI methods are not tailored towards models in sensitive domains where one predictor is of special interest, such as a treatment effect in a clinical model, or ethnicity in policy models. We introduce Path-Wise Shapley effects (PWSHAP), a framework for assessing the targeted effect of a binary (e.g. treatment) variable from a complex outcome model. Our approach augments the predictive model with a user-defined directed acyclic graph (DAG). The method then uses the graph alongside on-manifold Shapley values to identify effects along causal pathways whilst maintaining robustness to adversarial attacks. We establish error bounds for the identified path-wise Shapley effects and for Shapley values. We show PWSHAP can perform local bias and mediation analyses with faithfulness to the model. Further, if the targeted variable is randomised we can quantify local effect modification. We demonstrate the resolution, interpretability and true locality of our approach on examples and a real-world experiment.

1. Introduction

Recent years have seen an increase in the demand for transparency on machine learning-based decisions. In safety-sensitive settings particularly, practitioners need to understand how a model reasons in order to ensure its safe deployment in the future. In many scenarios, their attention focuses on assessing the importance of a specific predictor variable such as a treatment in a clinical model, or ethnicity with regards to model fairness. Ultimately, as humans naturally have a *causal* approach to model explainability, users may want to understand how the treatment[1] causally impacts the outcome i.e. through what mechanisms and if this matches their prior assumptions on the causal relationships in the data. Here, we focus on the following question: in the presence of a general black-box ML model, how can we compute feature attributions according to the causal beliefs encapsulated by the posited DAG? We provide a framework for locally explaining a treatment's effect in such settings, with the following goals: *reliability*, *safety*, *interpretability* and *high resolution*.

Reliability, safety, interpretability in XAI An XAI model should aim at generating explanations that are *reliable*, *safe* and *interpretable*. Reliability, also known as being "true to the model", implies that the explanations do reveal the functional dependence of the model and are robust to distributional shifts. Safety relates to the ability to protect the framework from hazard, in particular attempts to fool models with deceptive data, also known as adversarial attacks. As defined in (Miller, 2019), interpretability is the degree to which a human can understand the cause of a decision.

High resolution for safety-critical XAI In addition to the XAI goals cited above, resolution is often necessary when explaining a model to ensure its fairness. Following (Chiappa, 2019), we illustrate our point using a simple causal structure inspired by the Berkeley admissions dataset. Consider a predictive model for college entry with three features: sex, exam results and department. The sensitive attribute, sex, potentially impacts the predicted admission through both fair and unfair causal pathways. Sex may indirectly and fairly impact admission due to some individuals applying to more competitive departments. However, there may also be an effect through an unfair direct path, representing prejudice on the part of the admissions officer. This motivates *path-specific* measures of feature importance, instead of an overall single score that groups all paths together.

[*]Equal contribution [1]Department of Statistics, University of Oxford, Oxford, UK [2]Department of Statistical Science, University College London, London, UK [3]The Alan Turing Institute, London, UK. Correspondence to: Lucile Ter-Minassian <lucile.ter-minassian@stats.ox.ac.uk>, Oscar Clivio <oscar.clivio@stats.ox.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]We use the example of "treatment" in clinical models as illustrative of a central binary predictor variable.

Our solution: Path-Wise Shapley (PWSHAP)

We introduce a method for explaining the local effect of a binary treatment under an assumed causal graph. We assume that (i) the treatment is an ancestor of the outcome in the directed acyclic graph (DAG) and that (ii) the DAG is compatible with the data, i.e. that it respects all conditional independences that could be found by running conditional independence tests in the data. The posited DAG may come from prior domain knowledge, or indeed be learnt from data and represents the user's beliefs. Our aim is not to understand the model's "internal DAG", and thus we do not assume that the posited DAG corresponds to the DAG of the underlying model.

We show how augmenting the predictive model with such a causal DAG supports a novel targeted extension of SHAP values, allowing for the decomposition of the black-box treatment effect into interpretable path-wise Shapley effects. We provide stand-alone theoretical results for decomposing the original on-manifold Shapley value (i.e. Shapley with a conditional reference distribution) of a treatment feature into path-wise local causal effects. We claim that our method achieves the four goals presented above: reliability, safety, interpretability and high resolution. Our contributions are as follows:

- We introduce Path-Wise Shapley (PWSHAP) effects, an extension of on-manifold Shapley values for locally explaining treatment effect under a causal DAG. Robustness to adversarial attacks (and thus safety) is guaranteed by the adoption of a conditional reference distribution. Reliability is ensured by the acknowledgment of the causal structure. As such, PWSHAP reconciles both safety and reliability.
- We show how our method can be used as a non-parametric alternative to mediation and bias analysis.We further show how PWSHAP can be used for fairness studies when the causal graph involves a mixture of fair/unfair paths, and under randomised treatment to assess effect modification (also referred to as moderation). We further show that Causal Shapley (Heskes et al., 2020), the closest method to ours, does not acknowledge moderation.
- We establish error bounds (i) from the outcome model to the Shapley values and PWSHAP effects (ii) from the Shapley values and treatment model (referred to as propensity score) to the PWSHAP effects.

To the best of our knowledge, we are the first to interrogate the link between the Shapley feature importance of a treatment, and the standard notion of the treatment effect as defined within the causal inference literature, as the expected difference between potential outcomes under the two treatments.

2. Shapley Values

Shapley values are a local feature attribution method. They quantify the importances of the features $\{1, \dots, m\}$ of a complex machine learning model $f : \mathbb{R}^m \to \mathbb{R}^l$ at an instance $x \in \mathbb{R}^m$, given only black-box access to the model. The local prediction $f(x)$ is formulated as a sum of individual feature contributions: $f(x) = \phi_0^f(x) + \sum_{i=1}^M \phi_i^f(x)$, where $\phi_j^f(x)$ is the contribution of feature j to $f(x)$ and $\phi_0^f(x) = \mathbb{E}[f(X))]$ is the averaged prediction with the expectation over the observed data distribution. The Shapley value of a feature j captures the change in model outcome comparing the prediction when the feature value x_j is included to when it's removed from the input. This change is computed from the difference in value function v when setting feature j equal to the instance feature value x_j, averaged over all possible coalitions S of features excluding feature j. If a feature is included in the coalition its value is set to the observed instance value x_j. To model feature removal, the value function takes the expectation of the black-box algorithm at observation x over the non-included features $\overline{S}$ using a reference distribution $r(X \mid x_S)$ such that $v_f(S, x) = \mathbb{E}_{r(X \mid x_S)}[f(x_S, X_{\overline{S}})]$ for $\overline{S} := \{1, \dots, m\} \backslash S$ and the operation $(x_S, x_{\overline{S}})$ denoting the concatenation of its two arguments. Binomial weights $|S|!(m - |S| - 1)!/(m - 1)!$ take account of all possible orderings. The Shapley value of feature j is thus:

$$\phi_j^f(x) = \sum_{i=0}^{m-1} \frac{1}{m\binom{m-1}{i}} \sum_{\substack{S \not\ni j \\ |S|=i}} [v_f(S \cup \{j\}, x) - v_f(S, x)],$$

i.e. $\phi_j^f(x) = \mathbb{E}_{p(S|j \notin S)}[\phi_{j,S}^f(x)]$ where $\phi_{j,S}^f(x) := v_f(S \cup \{j\}, x) - v_f(S, x)$ and $\forall j,\ p(S \mid j \notin S) = 1/m\binom{m-1}{|S|}$. Shapley values have become a gold standard amongst explanation models due to their desirable properties (model agnostic, additive) and axioms (*Symmetry*, *Efficiency*, *Linearity* and *Dummy*—see Supplement D.1 for details). However, the method has not been adopted in critical settings due to the considerable limitations of both possible reference distributions (Janzing et al., 2020; Chen et al., 2020; Sundararajan & Najmi, 2020).

Limitations of Shapley values On the one hand, *on-manifold* Shapley values (Aas et al., 2021) use a conditional reference distribution, conditioning on x_S to better account for correlations between features $r(X \mid x_S) := p(X \mid X_S = x_S)$. Sampling from a conditional distribution forces the model to be evaluated on plausible instances that lie on the data manifold. It thus improves the adversarial robustness and thus the safety of the method (Slack et al., 2020). However, on-manifold Shapley values have been shown to be unreliable as they can generate misleading explanations (Janzing et al., 2020; Sundararajan & Najmi, 2020). On the other hand, *off-manifold* Shapley values use

a marginal reference distribution, that is $r(X \mid x_S) := p(X)$ (Lundberg and Lee, 2017). The resulting explanations reveal the functional dependence better, also known as being "true to the model" (Chen et al., 2020). However, sampling from the marginal distribution breaks the dependence between features. Consequently, off-manifold Shapley values are sensitive to adversarial robustness and thus deemed unsafe (Slack et al., 2020). Note that adversarial robustness is key for fairness studies. If an unfair model undergoes an adversarial attack, it may "counterbalance" its potential prejudice on real-world data by forming predictions favourable to disadvantaged groups on implausible inputs. Since only the marginal distribution is used, the resulting Shapley value of a sensitive attribute might look fair even though the model would predict unfairly on real-world data (Slack et al., 2020). Ultimately, Shapley values can't provide both reliable and safe explanations, which may hinder their adoption in safety-critical settings (see further details on Shapley values in Supplements A).

Shapley values also have limited interpretability. The attribution of a target feature j is the result of model evaluations averaged over all coalitions excluding j. The goal of this procedure is to acknowledge all the correlations amongst features. However, if some features are assumed to be independent, this assumption fails. Averaging over coalitions with/without independent features may generate redundancies and unbalance the resulting attribution. Also, the *interpretation* of on-manifold and off-manifold Shapley values is agnostic to the assumed *causal structure*, if any. When a specific treatment is of interest, causal interpretation of its Shapley values should be done in light of the relative roles of other features: confounder, moderator or mediator; see Supplement B for a definition of these notions. Interpreting Shapley values causally would be a case of the "Table 2 fallacy" (Westreich & Greenland, 2013), where all coefficients of a model are misleadingly interpreted as adjusted causal effects. Thus we claim that under a posited DAG, the Shapley value of a feature should be computed according to the assumed statistical dependencies, i.e. the *edges* in the DAG, and interpreted in light of its causal links with other variables, i.e. the *directions* of the arrows in the DAG.

In PWSHAP, we use a conditional reference distribution to ensure the robustness to adversarial attacks and safety of our method. Meanwhile, we are able to generate feature attributions that are both reliable and interpretable, thanks to the tailored causal interpretations of the effects we compute.

3. Path-Wise SHAP (PWSHAP)

The intuition behind the introduced method is two-fold. First, we decompose the Shapley value as a weighted sum of quantities that can be interpreted causally as treatment effects along coalitions. Second, by only considering relevant coalitions, we are able to deduce quantities that can be interpreted causally along paths. Since the treatment T is of special interest, we separate it from the other variables, that we call covariates C, such that $X = (C, T)$.

3.1. Problem Setup

Let C denote covariates, T a binary treatment and Y an outcome of interest. We assume that $Y = f^*(C, T) + \epsilon$, with $\mathbb{E}[\epsilon|C, T] = 0$. Our black-box f is an arbitrary function of $X = (C, T)$ which aims at predicting f^*. We aim at explaining the specific effect of the treatment variable T on the predictions made by the black-box f for an individual with values c of covariates C. To do so, we first decompose the Shapley value of T into a weighted sum of "Shapley effects" which are inspired by conditional average treatment effects, commonly used in the causal literature. We refer to a coalition S excluding treatment T as a *subset of covariates* and note the value function as $v_f(S \cup \{T\}, c_S, t)$ when it is taken over the coalition $S \cup \{T\}$ and $v_f(S, c_S)$ when taken over S. Notations are summarised in Section C, with a running example to illustrate them all in Supplement I.

PWSHAP relies on two assumptions: (i) the treatment of interest is a causal ancestor of the outcome (no anti-causal learning) and (ii) the DAG is compatible with the observed data i.e. all conditional independence constraints implied by graphical d-separation relations hold in the data. The user-supplied DAG thus only encodes the conditional dependences and is not assumed to be identical to the underlying model behavior. The "direction" of the arrows in the DAG is only used for causally *interpreting* the PWSHAP values.

3.2. Decomposition into Shapley Effects

First, we notice a connection between value functions v_f of the black-box f and conditional average treatment effects using coalition-wise Shapley values.

Definition 3.1 (Coalition-wise Shapley effect)**.** We define the coalition-wise Shapley effect[2] of T on Y along the covariates C_S indexed by the subset of covariates S as:

$$\Psi^f_{T\to Y|C_S}(c_S) = v_f(S\cup\{T\}, c_S, 1) - v_f(S\cup\{T\}, c_S, 0)$$

The coalition-wise Shapley effect can be understood as a generalisation of conditional average treatment effects. Indeed, for the true model f^*, the RHS is equal to $\mathbb{E}[Y|C_S = c_S, T = 1] - \mathbb{E}[Y|C_S = c_S, T = 0]$. Under the typical causal treatment effect identification assumptions, i.e. no interference, consistency, and conditional exchangeability given C (Imbens & Rubin, 2010), this is the conditional average treatment effect (CATE) (Rubin, 2005) (definition in Supplement B) when S is the complete coalition, i.e.

[2]Note that our Shapley effects are orthogonal to those introduced by (Iooss & Prieur, 2017) for numerical models .

containing all covariates. In addition, $\Psi^f_{T\to Y|\emptyset}$ is the "base" treatment effect, i.e. a population-wide estimate of treatment effect. Its exact causal interpretation depends on the structure of the DAG, but in some cases it equates to the Average Treatment Effect (ATE) as defined by (Rubin, 2005) (definition in Supplement B). The coalition-specific Shapley effect can be linked to the original Shapley values as follows.

Property 3.1 (Decomposing Shapley values into Shapley effects). The *coalition-wise Shapley value* $\phi^f_{T,S}(c,t)$ is equal to a weighted estimate of a local treatment effect,

$$\phi^f_{T,S}(c,t) = w^*_S(c_S,t)\cdot\Psi^f_{T\to Y|C_S}(c_S), \qquad (1)$$

where $w^*_S(c,t)$ denotes what we call the "propensity weights" defined by $w^*_S(c,t) = t - p(T=1|C_S=c_S)$. This name follows the fact that these weights are related to whether the sample is an outlier or not.

The proof can be found in Supplement J.1. Property 3.1 shows that each coalition-specific term in the original on-manifold Shapley value is equal to the product of two terms. The first is a weight that depends on the propensity score. The second is a measure of the treatment effect, namely the coalition-specific Shapley effect. As a result, the overall Shapley value $\phi^f_T(c,t)$ can be decomposed as

$$\phi^f_T(c,t) = \mathbb{E}_{p(S|T\notin S)}[w^*_S(c_S,t)\cdot\Psi^f_{T\to Y|C_S}(c_S)].$$

3.3. Path-Wise Shapley (PWSHAP) Effects

Although we connected Shapley values to coalition-wise Shapley effects the latter still only apply to *coalitions* and not specific *paths*. However, the coalition-wise Shapley effect $\Psi^f_{T\to Y|C_S}(c_S)$ can be understood as the causal flow from T to Y through a set of covariates S. Thereby, we define the causal flow along the (undirected) path from T to Y through C_i as the difference between the causal flow through all covariates and the causal flow through all covariates but C_i. See Supplement G for a generalisation of paths of length 3 or more.

Definition 3.2 (Path-wise Shapley effect)**.** Let S^* be the coalition with all covariates. We refer to the following quantity as the path-wise Shapley effect of T on Y along the path from T to Y through C_i:

$$\Psi^f_{C_i}(c) = \Psi^f_{T\to Y|C_{S^*}}(c) - \Psi^f_{T\to Y|C_{S^*\setminus\{i\}}}(c_{S^*\setminus\{i\}}).$$

For instance in the fairness example from Section 1, the path-wise Shapley effect of sex on admission (Adm) mediated by the chosen department (Dpt) $\Psi^f_{Sex\to Dpt\to Adm}$ is $\Psi^f_{Sex\to Adm|Dpt,Exam} - \Psi^f_{Sex\to Adm|Exam}$.

Path-wise Shapley effects thus quantify the change in model outcome when specifying the feature values along a specific path, compared to when all features are specified but the ones on the path of interest. As such, PWSHAP measures the effect of the treatment on the outcome through a causal pathway. Ultimately, conditioning on all other features reinforces the locality of our result. It can also be seen as a contribution to the shift from a global estimated "base" treatment effect to an individual estimated treatment effect. However, note that PWSHAP violates the efficiency property i.e. they do not sum up to an interpretable quantity like the original Shapley feature attributions do. Moreover, Property 3.2 shows that integrating PWSHAP effects can help isolate covariates that are conditionally independent on the treatment given other covariates (the Supplement J.2 for the proof). As shown in Section 6, the actual causal meaning of this conditional independence depends on the posited DAG of the data, however.

Property 3.2 (Integration of the PWSHAP effects). Let C_i be a covariate such that $C_i \perp\!\!\!\perp T|C_{-i}$ where $C_{-i} := C_{S^*\setminus\{i\}}$. Then for any function f and any value c_{-i} of C_{-i},

$$\mathbb{E}[\Psi^f_{C_i}(C_i,c_{-i})|C_{-i}=c_{-i}] = 0$$

3.4. Estimation of Shapley Effects from Shapley Values

Using Property 3.1, we can express the coalition-wise Shapley effects $\Psi^f_{T\to Y|C_S}$ from the coalition-wise Shapley values $\phi^f_{T,S}(c,t)$ as $\Psi^f_{T\to Y|C_S}(c_S) = {}^{\phi^f_{T,S}(c,t)}/_{w^*_S(c,t)}$. Therefore, the path-wise Shapley effects $\Psi^f_{C_i}$ are computed as:

$$\Psi^f_{C_i}(c) = \frac{\phi^f_{T,S^*}(c,t)}{w^*_{S^*}(c,t)} - \frac{\phi^f_{T,S^*\setminus\{i\}}(c,t)}{w^*_{S^*\setminus\{i\}}(c,t)}.$$

In practice, path-wise Shapley effects are computed by replacing the true propensity weights with weights that use an estimate of the propensity score. For this, we further need to assume positivity holds. The path-wise Shapley effect of T on Y through C_i is thus estimated in three steps: (i) computing the coalition-wise Shapley values for S^* the entire set of covariates and $S^*\setminus\{i\}$; (ii) dividing each of these terms by an estimate of their corresponding propensity weight; (iii) taking the difference between the two resulting quantities (also known as coalition-specific Shapley effects) to isolate the effect along the path through C_i. Note that division by weights requires overlap, that is $\forall c, 0 < p(T=1|C=c) < 1$.

4. Related Work

4.1. Conceptual Distinction Between Local Explanations Models and Causal Inference

Causal inference aims at assessing the effect of a *feature* at a global scale (e.g. ATE) or within subgroups (e.g. CATE). Contrastingly, Shapley values assess the *local* effect of a

feature value compared to the values taken by that feature in the reference distribution. Therefore, to identify path-wise local effects, we consider a path to be "deactivated" when the covariate value gets sampled from the reference distribution, i.e. the covariate is *not* in the coalition. Conversely, specifying a covariate value, i.e. when the covariate *is* in the coalition, "activates" a path. Thereby, coalition-wise effects are conditional treatment effects marginalised over covariates that aren't in the coalition. In the admission example from Section 1, the coalition-specific Shapley effect for $\{Exam\}$, $\Psi^f_{Sex \to Adm|\{Exam\}}$ corresponds to the treatment effect along two paths: the direct path $Sex \to Adm$ and the path from Sex to Adm through $Exam$.

4.2. Comparison of PWSHAP with Existing Methods

We compare our method to two baseline explanation methods. Our first baseline is Causal Shapley (CS) (Heskes et al., 2020), another method aiming to explain a model under an assumed causal DAG. Like PWSHAP, Causal Shapley splits Shapley attributions, although the split is binary (direct/indirect effect). In Causal Shapley, the indirect effect of a feature j, the distribution of the 'out-of-coalition' features changes due to the do-operator (see Suppl. A and D.2 for further details). Our second baseline is on-manifold Shapley, a natural choice given that PWSHAP augments the original method. Section D.3 details other graph based Shapley methods (Wang et al., 2021; Singal et al., 2021), which are not appropriate baselines here due to structural differences.

Higher model fidelity, lower reliance on causal assumptions than Causal Shapley We claim that PWSHAP has higher model fidelity and relies less on the assumed causal structure than Causal Shapley. As the direct/indirect effect split is based on *do*-calculus in Causal Shapley, the computation of the attributions depends on the assumed DAG (both the edges and their directions). In contrast, PWSHAP computations only depend on the hypothesised feature *dependencies* i.e. the edges in the DAG. Only the causal *interpretation* of PWSHAP depends on the direction of the edges. We view the fact that our approach is agnostic to the choice of a (compatible) DAG as a strength, as it allows different experts to explain the black-box model output according to their own causal beliefs about the data or phenomenon being studied (see D.5 for a detailed discussion on this). Ultimately, by applying *do*-calculus, Causal Shapley computes feature attributions according to preconceptions of how the model should reason, and as such is "forcing" explanations to fit to a presumed causal structure. To further illustrate the limitations of relying on the causal assumptions and show that PWSHAP has higher fidelity to the model, let us consider a black-box with a single covariate C, and a treatment T. If we wrongly assume C to be a confounder instead of a mediator, the indirect effect of treatment i.e. the mediation of treatment through C would be null according to Causal Shapley (see Property D.1 in Supplement D.2). By contrast, only the causal interpretation of the PWSHAP effect through C would be incorrect, but its value would remain unaltered.

Increased resolution, better interpretability Compared to both Causal and on-manifold Shapley, PWSHAP has higher resolution—as it is path-specific instead of feature-specific—and improved interpretability. In Causal Shapley and on-manifold Shapley values, feature attributions result from taking an average over coalitions, whereas PWSHAP only considers coalitions used to compute effects. Ultimately, evidence has shown that on-manifold Shapley values and Causal Shapley values can generate misleading interpretations (Sundararajan & Najmi, 2020). In on-manifold Shapley values the attribution of a feature that does not appear in the algebraic formulation of the model can be non-zero, depending on how the data is distributed. This is induced by both the conditional reference distribution, the average taken over multiple coalitions. By providing an exact interpretation for the computed quantities, PWSHAP overcomes this unreliability issue. If a PWSHAP effect $\Psi^f_{C_i}$ is null, it means that specifying the covariate $C_i = c_i$ has had no impact on the treatment effect compared to marginalising it, *according to our black-box* (see Lemma 6.2 and Property 6.1). Meanwhile, PWSHAP remains robust to adversarial attacks, as it samples from a conditional reference distribution (Slack et al., 2020). PWSHAP thus reconciles *safety* and *reliability*. However, a limitation of PWSHAP compared to both baselines is that it violates the efficiency property (see Suppl. D).

5. Error Bounds

We now show how to obtain error bounds for quantities like path-wise Shapley effects from other quantities like the outcome model, according to Figure 1. To the best of our knowledge, these are the first results regarding error bounds for on-manifold Shapley values. In the following, $(\hat{f}_N)$ denotes a sequence of estimators of f^*. The proofs can be found in Supplements J.3 and J.4

Property 5.1 (Convergence of the outcome model implies convergence of Shapley values and PWSHAP effects). If $\forall c, t, N, |\hat{f}_N(c,t) - f^*(c,t)| \leq e_N^{\text{outcome}}$ then:

1. Convergence of the coalition-specific Shapley terms: $\forall c, t, N, \quad |\phi^{\hat{f}_N}_{T,S}(c,t) - \phi^{f^*}_{T,S}(c,t)| \leq 2e_N^{\text{outcome}}$. which implies the convergence of the Shapley value of the estimated model to that of the true model.
2. Convergence of the path-wise Shapley effects: $\forall i, c, N, \quad |\Psi^{\hat{f}_N}_{C_i}(c) - \Psi^{f^*}_{C_i}(c)| \leq 4e_N^{\text{outcome}}$.

Property 5.2 (Convergence of estimated coalition-specific Shapley values and propensity score implies con-

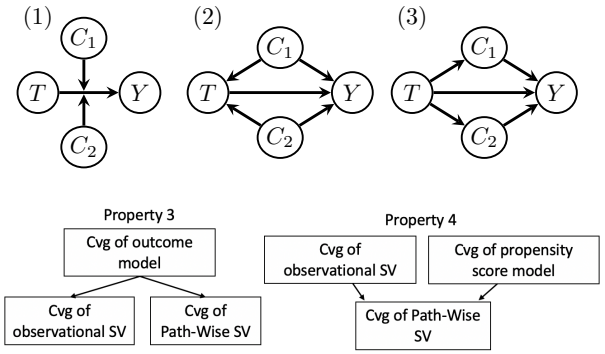

Figure 1: DAGs for Building Blocks (Up) and Error Bound (Down, Cvg=Convergence, SV=Shapley value)

vergence of estimated PWSHAP effects). Assuming that: (1) the arbitrary propensity score model π^N and the true propensity score model π^* verify ϵ-strong overlap,
(2) $\forall c, N | \pi^N(c) - \pi^*(c)| \leq e_N^{\text{propensity}}$
(3) $\forall S$ s.t. $T \notin S, c, N, |\hat{\phi}_{T,S}^{N,\hat{f}_N}(c,t) - \phi_{T,S}^{f^*}(c,t)| \leq e_N^{\text{Shap}}$,

with $w_S^N(c,t) = t - \mathbb{E}_p(C_{\bar{S}}|C_S = c_S)[\pi^N(c_S, C_{\bar{S}})]$ we show the convergence of the estimated PWSHAP effects to the true PWSHAP effects, $\forall i, c, t, N$,

$$|\hat{\Psi}_{C_i}^{N,\hat{f}^N}(c) - \Psi_{C_i}^{f^*}(c)| \leq \frac{4e_N^{\text{Shap}}}{\epsilon} + \frac{4||f^*||_\infty \cdot e_N^{\text{propensity}}}{\epsilon^2}.$$

6. Causal Interpretations

PWSHAP effects are interpreted by revisiting causal inference concepts of confounding, moderation and mediation at a local scale. As our method stands on theoretical grounding, we first provide objective evidence using explicit equations.

6.1. Local Bias Analysis

Under DAG (2) of Figure 1, PWSHAP effects are causally interpreted as follows:

$$\begin{aligned}\phi_T^{f^*} &= {}^{w_{12}^*}/_3 \cdot \text{CATE}(c_1, c_2) + {}^{w_1^*}/_6 \cdot \text{ "CATE"}_{C_1}(c_1) \\ &+ {}^{w_2^*}/_6 \cdot \text{"CATE"}_{C_2}(c_2) + {}^{w^*}/_3 \cdot \text{Diff. in means}\end{aligned}$$

where $\text{"CATE"}_{C_S}(c_s) = \mathbb{E}[Y|T=1, C_S = c_S] - \mathbb{E}[Y|T = 0, C_S = c_S]$. We refer to these terms as "CATE"s, in an abuse of notation, but note that they are not true causal conditional average treatment effects, as they include confounded paths. The term "Diff. in means" stands for $E[Y|T=1] - E[Y|T=0]$. To isolate the spurious effect of the confounders, we further assume that the two confounders are not effect moderators.

Definition 6.1 (Local confounding effect)**.** In this example, we call the PWSHAP effect of C_2 the "local confounding effect of C_2" and note it $\Psi_{T \leftarrow C_2 \rightarrow Y}^f$. In other words, $\Psi_{T \leftarrow C_2 \rightarrow Y}^f := \Psi_{C_2}^f$

Notably, for the true model f^*, $\Psi_{T \leftarrow C_2 \rightarrow Y}^{f^*}(c_1, c_2) = \text{CATE}(c_1, c_2) - \text{"CATE"}_{C_1}(c_1)$. This quantity has been referred to as the bias due to unmeasured confounding (assuming we observe C_1 but not C_2) in a segment of the sensitivity analysis literature (Veitch and Zaveri, 2020). Therefore, our measure of confounding effect is a local equivalent of this bias. Indeed, integrating out this difference over C_1, C_2 yields $\text{ATE} - \mathbb{E}[\mathbb{E}[Y|T=1, C_1] - \mathbb{E}[Y|T=0, C_1]]$ where ATE is the Average Treatment Effect. However, if a covariate is both a confounder and an effect modifier, its path-wise attribution will cover both phenomena and the two effects will be indiscernible. Ultimately, PWSHAP contrasts with sensitivity analysis methods which are meant for quantifying unobserved confounding, whereas our method measures the impact of an observed confounder. Further details on such techniques can be found in the Supplement D.6.

Lemma 6.2 (Integration of the local confounding effect, true model)**.** *Let C_1, C_2 be two pre-treatment covariates such that ignorability given C_1, C_2 holds, i.e. $\forall t, Y(t) \perp\!\!\!\perp T | C_1, C_2$. If, additionally, C_2 is not a confounder, i.e. C_1 alone guarantees ignorability or $\forall t, Y(t) \perp\!\!\!\perp T | C_1$, then the integral of the local confounding effect of f^* w.r.t. C_2 on the joint distribution of covariates is null:*

$$\mathbb{E}[\Psi_{T \leftarrow C_2 \rightarrow Y}^{f^*}(C_1, C_2)] = 0.$$

The proof can be found in Supplement J.5. For a variable that is not actually a confounder, the integration of the local confounding effect thus yields zero. This can be generalised to any number of confounding pre-treatment covariates, by grouping all of them in C_1. For any blackbox f, a stricter condition yields the same result as a corollary of Proposition 3.2. We give that result and an example of local bias analysis in Supplement E.2. Further, if the local confounding effect is zero for all individuals in the training set, then we can hypothesise that the model did not learn to predict through the confounding path $T \leftarrow C_2 \rightarrow Y$.

6.2. Local Moderation Analysis Under Randomised Treatment

Here, "moderation" refers to an effect modification as in (Boruvka et al., 2018). In the setting represented in Figure 1, causal graph (1), where treatment is assumed to be unconfounded, we interpret the PWSHAP decomposition as follows:

$$\begin{aligned}\phi_T^{f^*} &= {}^1/_3 \cdot w_{12}^* \cdot \text{CATE}(c_1, c_2) + {}^1/_6 \cdot w_1^* \cdot \text{CATE}_{C_1}(c_1) \\ &+ {}^1/_6 \cdot w_2^* \cdot \text{CATE}_{C_2}(c_2) + {}^1/_3 \cdot w^* \cdot \text{ATE}\end{aligned}$$

Definition 6.3 (Local moderating effect)**.** In this example, we call the PWSHAP effect of C_2 the "local moderating effect of C_2" and denote it $\Psi_{C_2:T \rightarrow Y}^f$. In other words, $\Psi_{C_2:T \rightarrow Y}^f := \Psi_{C_2}^f$.

PWSHAP assesses the local effect modification induced

by C_2 by "unspecifying" this feature. Having null local moderating effect would mean that C_2 did not act as a moderator for this specific subject, *according to our fitted black-box*. Unlike previous methods (Imai and Ratkovic, 2013; Athey and Imbens, 2016; Wang and Rudin, 2017), our PWSHAP approach to moderation analysis does not involve subgroup finding—a technique known to be underpowered (Holmes and Watson, 2018)—and is nonparametric (see Supplement D.6 for a review of moderation analysis). Ultimately, we show that in the presence of pre-treatment moderators, Causal Shapley compounds the main effect of treatment and its effect via moderation into a single "direct" effect, whereas PWSHAP explanations are able to distinguish the added treatment effect due to moderation from the main effect. We compare PWSHAP with Causal Shapley on an example as shown in DAG (1) of Figure 1 assuming $Y = \beta T + \gamma_1 C_1 + \gamma_2 C_2 + \alpha_1 T C_1 + \alpha_2 T C_2 + \epsilon$ with $\mathbb{E}[\epsilon|T, C_1, C_2] = 0$ and where C_1, C_2 are two independent moderators with $C_1, C_2 \sim \text{Uniform}(0, 1)$. Treatment is randomised: $T \sim \text{Bernoulli}(p)$. Details about the following are given in Supplement E.1. PWSHAP yields:

$$\begin{aligned}
\Psi^{f^*}_{T\to Y|C_1,C_2} &= \beta + \alpha_1 c_1 + \alpha_2 c_2 \\
\Psi^{f^*}_{C_1} &:= \Psi^{f^*}_{C_1:T\to Y} = \alpha_1(c_1 - 1/2) \\
\Psi^{f^*}_{T\to Y|\emptyset} &= \beta + \alpha_1/2 + \alpha_2/2 \\
\Psi^{f^*}_{C_2} &:= \Psi^{f^*}_{C_2:T\to Y} = \alpha_2(c_2 - 1/2)
\end{aligned}$$

where $\mathbb{E}[C_1] = \mathbb{E}[C_2] = 1/2$. PWSHAP effects through C_1 and C_2 are null if $C_1 = C_2 = 1/2$. The PWSHAP approach thus matches the default behaviour of local explanation methods: paths through effect moderators are given zero attribution if the moderator value is equal to the population average. This highlights the true locality of our method. Furthermore, one can check that the moderating effects integrate to 0. Again, this is coherent with the overall definition of randomised treatment in causal inference. By contrast, moderation by C_1 and C_2 is overlooked in Causal Shapley as $\phi^{f^*,\text{CS}}_{T,\text{indirect}} = 0$. Further, $\phi^{f^*,\text{CS}}_{T,\text{direct}} = (t-p)\{\beta + \frac{\alpha_1}{2}\cdot(c_1 + \frac{1}{2}) + \frac{\alpha_2}{2}\cdot(c_2 + \frac{1}{2})\}$ which does not reflect the local behaviour of the model.

6.3. Local Mediation Analysis

Under DAG (3) of Figure 1, i.e. with unconfounded treatment and two mediators only depending on it, the causal interpretation of the PWSHAP approach to mediation is:

$$\begin{aligned}
\phi^{f^*}_T &= 1/3 \cdot w^*_{12}\text{CDE}_{C_1,C_2}(c_1, c_2) + 1/6 \cdot w^*_2\text{CDE}_{C_2}(c_2) \\
&\quad + 1/6 \cdot w^*_1\text{CDE}_{C_1}(c_1) + 1/3 \cdot w^*\text{ATE}
\end{aligned}$$

where CDE refers to the *Controlled Direct Effect* (definition in Supplement B), with $\text{CDE}_{C_S}(c_s) = \mathbb{E}[Y|T = 1, C_S = c_s] - \mathbb{E}[Y|T = 0, C_S = c_s]$. We claim that the difference in CDE is able to isolate the local effect of a given mediator and has a causal interpretation, as outlined by the local mediating effect introduced below.

Definition 6.4 (Local mediating effect)**.** Here, we call the PWSHAP effect of C_2 the "local mediating effect of C_2" and note it $\Psi^f_{T\leftarrow C_2\to Y}$. So, $\Psi^f_{T\to C_2\to Y} := \Psi^f_{C_2}$.

Property 6.1 (Ancestors of outcome). Let M_1, M_2 be two post-treatment and pre-outcome variables. Assuming that variables C include all confounders of the relationships between T, Y and (M_1, M_2) and that $M_2 \perp\!\!\!\perp T, M_1|C$, then for any value c of C and m_1 of M_1,

$$\mathbb{E}[\Psi_{T\to M_2\to Y}(c, m_1, M_2) \mid C = c] = 0.$$

In other words, if M_2 is not mediating the effect of T on Y because $M_2 \perp\!\!\!\perp T|C$, and M_2 is independent of M_1 conditionally on T, C, then integrating the local mediating effect of M_2 yields 0 which is coherent with our intuition. The proof can be found in Suppl. J.6. See Suppl. D.6 for further comparisons with the traditional Natural Effects approach (definition in Supplement B) and with Causal Shapley.

7. Experiments: Synthetic Data

In the following two experiments, we show PWSHAP's ability to capture confounding and mediation on synthetic datasets. We infer path-specific Shapley effects following the procedure from Section 3.4 and compute the absolute values of their averages $\bar{\Psi}$ across the testing set. We divide these values by the empirical standard deviation of outcome σ_Y on the training set to mitigate variation due to the scale of the outcome. We follow the same process for Causal Shapley's direct and indirect effects w.r.t. the treatment. Results are averaged over 25 randomly sampled datasets, with standard errors shown in parentheses. More details are given in Supplement H.

Local bias analysis. We consider the previous model with two pre-treatment covariates C_1 and C_2 described in DAG (2) of Figure 1, and with results derived in Section 6.1. We look at three scenarios: (i) neither C_1 nor C_2 are confounders, (ii) C_1 is a confounder but C_2 is not, (iii) both are confounders. Results are shown in Table 1. Local confounding effects are significantly higher for confounding variables compared to non-confounding variables. This shows how these effects can isolate individual confounders in pre-treatment covariates, in accordance with Lemma 6.2. Conversely, we do not notice any significant change in Causal Shapley's direct effect. Causal Shapley's indirect effect is even lower - as it is expected to be zero from Property D.1.

Local mediation analysis. We consider DAG (3) of Figure 1, applied to a college admission example where we

Table 1: Results on local bias analysis.

Scenario	$\frac{\lvert\bar{\Psi}^f_{T\leftarrow C_1\rightarrow Y}\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\Psi}^f_{T\leftarrow C_2\rightarrow Y}\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\phi}^{f,\text{direct, CS}}_T\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\phi}^{f,\text{indirect, CS}}_T\rvert}{\sigma_Y}$
C_1, C_2 non-conf.	0.057 (0.011)	0.064 (0.013)	0.069 (0.010)	0.006 (0.001)
C_1 conf., C_2 not	0.505 (0.057)	0.054 (0.009)	0.067 (0.008)	0.002 (0.000)
C_1, C_2 conf.	0.322 (0.037)	0.277 (0.034)	0.076 (0.010)	0.006 (0.003)

Table 2: Results on local mediation analysis.

Scenario	$\frac{\lvert\bar{\Psi}^f_{T\rightarrow Q\rightarrow Y}\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\Psi}^f_{T\rightarrow D\rightarrow Y}\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\phi}^{f,\text{direct, CS}}_T\rvert}{\sigma_Y}$	$\frac{\lvert\bar{\phi}^{f,\text{indirect, CS}}_T\rvert}{\sigma_Y}$
Q, D non-med.	0.056 (0.012)	0.041 (0.007)	0.072 (0.012)	0.033 (0.003)
D med., Q not	0.059 (0.010)	0.715 (0.073)	0.080 (0.011)	0.091 (0.019)
Q, D med.	0.948 (0.117)	0.354 (0.065)	0.112 (0.014)	0.089 (0.015)

investigate the effect of sex—noted T — on the *logit* of the probability of admission, mediated by exam results $C_1 = Q$ and department choice $C_2 = D$. Details are in Supplement E.3. We look at three scenarios : (i) neither is a mediator, (ii) the former is a mediator but the latter is not, (iii) both are mediators. Results are presented in Table 2. Local mediating effects are significantly higher for mediating variables compared to non-mediating variables. This shows that these effects can isolate individual mediators in post-treatment covariates, in accordance with Property 6.1. Conversely, Causal Shapley's indirect effect seems to capture the presence of mediators - but not which variables are mediators.

7.1. Robustness to Adversarial Attacks

We empirically show that on-manifold Shapley values are more robust to adversarial attacks than off-manifold Shapley values, as they are computed using a conditional reference distribution. Consequently, PWSHAP effects which marginalise over a conditional reference distribution are also more robust to adversarial attacks than off-manifold Shapley values. We compare the three explanation methods on a synthetic fairness study. Here, we consider three black-box models: a fair model, an unfair model, and an "attacker model" that returns fair predictions on instances classified as on the data manifold and unfair predictions on instances classified as not belonging to the data manifold. We generate a gender sensitive attribute as $T \sim U(0,1)$, a departmental difficulty indicator $D \sim \textit{Bernoulli}(\pi(1-T) + (1-\pi)T)$, with $\pi = 0.99$. We also define a continuous test result $Q \sim U(0,1)$. Notably, $D = 1-T$ with high probability, so we take a classifier classifying a given unit (q, d, t) as belonging to the manifold iff $t = 1 - d$. The fair model is defined $f^{\text{fair}}(t, q, d) = q$, the unfair model as $f^{\text{unfair}}(t, q, d) = \frac{t+td}{2}$, and the attacker model as $f^{\text{attacker}}(t, q, d) = 1_{\{t=1-d\}} f^{\text{unfair}}(t, q, d) + 1_{\{t=d\}} f^{\text{fair}}(t, q, d)$. Figure 2 shows the explanations given by off-manifold Shapley values w.r.t. T, on-manifold Shapley values w.r.t. T, base PWSHAP effects $\Psi_{T\rightarrow Y|\emptyset}$, and PWSHAP effets wrt the path $T \rightarrow D \rightarrow T$ $\Psi_{T\rightarrow D\rightarrow Y}$ on for the three types of black-box models in Figure 2. The on-manifold Shapley values and PWSHAP effects return very similar boxplots for the unfair and attacker models. Both methods also capture that the unfair model and the attacker model make predictions from the sensitive attribute T while generating a low attribution to T for the fair model. However, note that the boxplot of $\Psi_{T\rightarrow D\rightarrow Y}$ for the unfair model does not match the theoretical expectation (from Suppl. E.3), a limitation that is most likely due to use of a potentially imprecise iterative imputer to fit conditional distributions, combined with division by small weights. Fitting conditional probabilities well to impute missing values remains an active topic of research (Lin & Tsai, 2020). For the attacker model, the off-manifold Shapley values are between that for the unfair model and the fair model, as approximately half of the data with columns generated independently is classified as belonging to the manifold and half is not. This illustrates how off-manifold Shapley values are less robust to adversarial attacks than on-manifold Shapley values and derived methods like PWSHAP effects.

Table 3: Results on Census Income Data.

Causal SHAP	ϕ_{direct}	$< 0.001(0.003)$
	ϕ_{indirect}	$0.004(0.006)$
PWSHAP	$\Psi_{\text{Race}\xrightarrow{\text{total}}\text{Inc}}$	$-0.005(0.003)$
	$\Psi_{\text{Race}\rightarrow\text{Capg}\rightarrow\text{Inc}}$	$0.077(0.004)$
	$\Psi_{\text{Race}\rightarrow\text{M.Stat}\rightarrow\text{Inc}}$	$0.361(0.004)$

8. Experiments : Real-World Data

We present a local mediation analysis experiment on the Adult data set from UCI (Asuncion & Newman, 2007), using the causal graph from (Frye et al., 2019). Further results and experimental details can be found in the Supplement H. The binary outcome denotes whether an individual's income exceeds $50,000 per year. The causal structure of the data is described in the DAG in Figure 3. Race was dichotomised into white/non-white. Our individual of interest is a white 38-year-old born in the US, whose marital status (M.Stat) is divorced and Relationship (Rltnshp) is unmarried, and who has a managerial occupation and no capital gain (Capg).

Our aim is to check how (i.e through which paths) being white has influenced our model's prediction of a 0.53 probability of high income for this individual, which is approx-

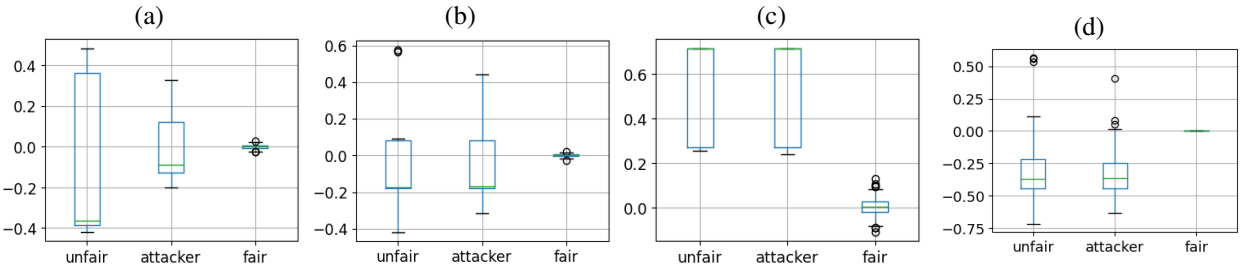

Figure 2: Boxplots of (a) off-manifold Shapley values w.r.t. T, (b) on-manifold Shapley values w.r.t. T, (c) base PWSHAP effects $\Psi_{T\to Y|\emptyset}$, (d) PWSHAP effects on the path through D $\Psi_{T\to D\to Y}$, each for the unfair, attacker and fair models.

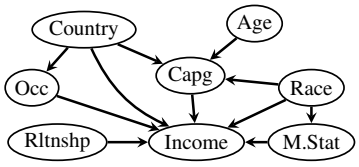

Figure 3: DAG of the Census Income Dataset.

imately 30 points higher than the average probability in the cohort. Results are shown in Table 3, with mean and standard deviation computed by subsampling. PWSHAP shows that the local mediating effect of race through marital status is predominant. Specifying marital status increases the *effect* of race by 0.361, as modeled by our black-box. Causal Shapley results in negligible direct and indirect effects, however we can't readily interpret these quantities obtained from averaging over many coalitions. This example illustrates three key attributes of PWSHAP compared with Causal Shapley, but also traditional Shapley methods overall: higher resolution, better interpretability of the resulting attributions, true locality of our explanations. The latter characteristic may explain why the effect of moderation by marital status is not captured by Causal Shapley. Here, our individual is an outlier as a large portion of white divorced individuals in the cohort have non-zero capital gain. Both the indirect and direct parts of Causal Shapley are a sum over all coalitions. This implies that for a majority of terms in Causal Shapley race and/or marital status are marginalised over, instead of being set to their feature values. The local effect of race via marital status is thus most likely "blurred" amongst all the coalitions that are "causally redundant" w.r.t. race i.e. where we add/drop features that are independent of race. By contrast, PWSHAP conditions on features that are independent of race (e.g. occupation), which ultimately increases the locality of our method.

9. Discussion

PWSHAP shows how Shapley values can generate granular causal explanations for local treatment effects under a posited causal graph. Our results on both simulated and real data show the applicability and locality of our method. The interpretability and safety of PWSHAP make it a strong candidate method for evaluating algorithms in sensitive environments, assuming treatment is binary and a known cause of the outcome. Practitioners could use PWSHAP explanations to identify inequity or unfairness, prior to deployment. However, our method heavily relies on the cumbersome estimation of the conditional distributions and on the assumed statistical dependencies. Further, dividing by propensity weights can bias the results if the weights are close to zero, and our approach is limited to binary treatments. Following (Chen et al., 2018), a potential extension of our work may acknowledge the proximity of features in the DAG, instead of only considering features with a direct edge to the target variable. Deriving weights to recover the efficiency property of Shapley is another perspective for future work. PWSHAP could also be used to help guide causal discovery as it can reveal possible conditional independence relationships between variables. Finally, although PWSHAP is a promising alternative to address rising ethical concerns in AI, note that fairness studies rely on the availability of sensitive data, which can be challenging for practical, ethical or legal reasons (Custers, 2010). We further caution against relying exclusively on XAI when causal knowledge is insufficient or for black-box without high-accuracy, as misleading interpretations may have negative social impact.

Acknowledgements

We thank Shahine Bouabid, Jake Fawkes, Siu Lun Chau, Robert Hu and our anonymous reviewers for their helpful feedback. LTM and OC are supported by the EPSRC Centre for Doctoral Training in Modern Statistics and Statistical Machine Learning (EP/S023151/1). LTM receives funding from EPSRC, OC from Novo Nordisk. KDO is funded by a Royal Society-Wellcome Trust Sir Henry Dale fellowship, grant number 218554/Z/19/Z. CH was supported by the EPSRC Bayes4Health programme grant and The Alan Turing Institute, UK.

References

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

Asuncion, A. and Newman, D. UCI machine learning repository, 2007.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Boruvka, A., Almirall, D., Witkiewitz, K., and Murphy, S. A. Assessing time-varying causal effect moderation in mobile health. *Journal of the American Statistical Association*, 113(523):1112–1121, 2018.

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

Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. L-shapley and c-shapley: Efficient model interpretation for structured data. *arXiv preprint arXiv:1808.02610*, 2018.

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

Custers, B. Data mining with discrimination sensitive and privacy sensitive attributes. *Custers BHM (2010), Data Mining with Discrimination Sensitive and Privacy Sensitive Attributes. In: Proceedings of ISP*, pp. 31–38, 2010.

Frye, C., Rowat, C., and Feige, I. Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*, 2019.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *arXiv preprint arXiv:2011.01625*, 2020.

Holmes, C. C. and Watson, J. A. Machine learning for randomised controlled trials: identifying treatment effect heterogeneity with strict control of type I error. *bioRxiv*, pp. 330795, 2018.

Imai, K. and Ratkovic, M. Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470, 2013.

Imbens, G. W. and Rubin, D. B. Rubin causal model. In *Microeconometrics*, pp. 229–241. Springer, 2010.

Iooss, B. and Prieur, C. Shapley effects for sensitivity analysis with dependent inputs: comparisons with sobol' indices, numerical estimation and applications. *International Journal for Uncertainty Quantification*, 9, 07 2017. doi: 10.1615/Int.J.UncertaintyQuantification.2019028372.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916. PMLR, 2020.

Lin, W.-C. and Tsai, C.-F. Missing value imputation: a review and analysis of the literature (2006–2017). *Artificial Intelligence Review*, 53:1487–1509, 2020.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777, 2017.

Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

Rubin, D. B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Singal, R., Michailidis, G., and Ng, H. Flow-based attribution in graphical models: A recursive Shapley approach. In *International Conference on Machine Learning*, pp. 9733–9743. PMLR, 2021.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Sundararajan, M. and Najmi, A. The many Shapley values for model explanation. In *International Conference on Machine Learning*, pp. 9269–9278. PMLR, 2020.

Veitch, V. and Zaveri, A. Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in Neural Information Processing Systems*, 33:10999–11009, 2020.

Wang, J., Wiens, J., and Lundberg, S. Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pp. 721–729. PMLR, 2021.

Wang, T. and Rudin, C. Causal rule sets for identifying subgroups with enhanced treatment effect. *arXiv preprint arXiv:1710.05426*, 2017.

Westreich, D. and Greenland, S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *American journal of epidemiology*, 177(4):292–298, 2013.

A. Shapley Values: Definitions

Off-manifold Shapley values

$$\phi_j^f(x) = \sum_{i=0}^{m-1} \frac{1}{m\binom{m-1}{i}} \sum_{\substack{S \not\ni j \\ |S|=i}} [\mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}})] - \mathbb{E}[f(x_S, X_{\overline{S}})]]$$

On-manifold Shapley values

$$\phi_j^f(x) = \sum_{i=0}^{m-1} \frac{1}{m\binom{m-1}{i}} \sum_{\substack{S \not\ni j \\ |S|=i}} [\mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}}) \mid X_{S\cup\{j\}} = x_{S\cup\{j\}}] - \mathbb{E}[f(x_S, X_{\overline{S}})] \mid X_S = x_S]$$

Causal Shapley values

$$\phi_j^f(x) = \sum_{i=0}^{m-1} \frac{1}{m\binom{m-1}{i}} \sum_{\substack{S \not\ni j \\ |S|=i}} [\mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}}) \mid \text{do}(X_{S\cup\{j\}} = x_{S\cup\{j\}})] - \mathbb{E}[f(x_S, X_{\overline{S}})] \mid \text{do}(X_S = x_S)]$$

where the contribution of feature j, $\phi_{j,S}^f(x)$, is decomposed into a direct and an indirect effect as follows:

$$\begin{aligned}
\phi_{j,S}^f(x) &= \mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}}) \mid \text{do}(X_{S\cup\{j\}} = x_{S\cup\{j\}}) - \mathbb{E}[f(x_S, X_{\overline{S}})] \mid \text{do}(X_S = x_S)] \\
&= \mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}}) \mid \text{do}(X_S = x_S)] - \mathbb{E}[f(x_S, X_{\overline{S}})] \mid \text{do}(X_S = x_S)] \\
&+ \mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}}) \mid \text{do}(X_{S\cup\{j\}} = x_{S\cup\{j\}})] - \mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}}) \mid \text{do}(X_S = x_S)]
\end{aligned}$$

where, in the last equality, the first line is the direct effect and the second line the indirect effect. See Section D.2 for details. Note that "interventional" Shapley values in Equation 3 of (Chen et al., 2020) are not causal Shapley values but actually off-manifold Shapley values : the do-operator is written in Equation 3, but it is misleading as the do-operator is said to intervene "by breaking the dependence between features in [the coalition] and the remaining features", i.e. making the expectation marginal as in off-manifold Shapley values.

B. Causal Inference Background

Confounder A confounder is a variable that is associated with both the exposure and the outcome, causing a spurious correlation. For instance, summer is associated with eating ice cream and getting sunburns, but there is no causal relationship between the two.

Mediator A mediator is a variable that is both an effect of the exposure and a cause of the outcome. In presence of a mediator, the total effect can be broken into two parts: the direct and indirect effect.

Moderator A mediator is a pre-exposure variable for which the causal effect is heterogeneous in subgroups.

Propensity score model A propensity score model is a function that predicts exposure from the observed covariates. We note it $\pi^*(c) = P(T = 1|C = c)$ and note π an estimate of π^*.

Potential outcome As defined by the Rubin causal model (Rubin, 2005), a potential outcome $Y(t)$ is the value that Y would take if T were set by (hypothetical) intervention to the value t.

Identification assumptions

- **No interference** For a given individual i, this assumption implies that $Y_i(t)$ represents the value that Y would have taken for individual i if T had been set to t for individual i, i.e the potential value of Y_i if T_i had been set to t.
- **Consistency** For a given individual i, $T_i = t \Rightarrow Y_i = Y_i(t)$. This means that for individuals who actually received exposure level t, their observed outcome is the same as what it would have been had they received exposure level t via an hypothetical intervention.
- **Conditional exchangeability** For a given individual i, we assume that conditional on C, the actual exposure level T is independent of each of the potential outcomes:
  $Y(t) \perp T \mid \mathbf{C}, \forall t$

Average Treatment Effect (ATE) The Average Treatment Effect for a binary treatment is the average difference in potential outcomes: $\mathbb{E}[Y(1) - Y(0)]$.

Conditional Average Treatment Effect (CATE) The Conditional Average Treatment Effect for a binary treatment, conditioned on C is the average difference in potential outcomes: $\mathbb{E}[Y(1) - Y(0)|C = c]$. If C is a sufficient adjustment set, i.e. conditional exchangeability w.r.t. C holds then the CATE can be identified as $\mathbb{E}[Y|T = 1, C = c] - \mathbb{E}[Y|T = 0, C = c]$.

Controlled Direct Effect Let $Y(t, m)$ be the potential outcome under exposure level $T = t$ and mediator level $M = m$. The controlled direct effect of T on outcome Y comparing $T = t$ with $T = t^*$ and setting M to m measures the effect of T on Y not mediated through M i.e. the effect of T on Y after intervening to fix the mediator to some value m. The controlled direct effect for individual i is then $\text{CDE}_i(t, t^*, m) = Y_i(t, m) - Y_i(t^*, m)$ (VanderWeele and Vansteelandt, 2009)

Natural Direct Effect The natural direct effect is defined as the difference between the value of the counterfactual outcome if the individual were exposed to $T = t$ and the value of the counterfactual outcome if the same individual were instead exposed to $T = t^*$, with the mediator M taking whatever value it would have taken at the reference value of the exposure $T = t^*$: $Y(t, M(t^*))\breve{}Y(t^*, M(t^*))$ (VanderWeele and Vansteelandt, 2009)

Natural Indirect Effect The natural indirect effect is the difference, having set the exposure to a fixed level $T = t$, between the value of the counterfactual outcome if the mediator M took whatever value it would have taken at a level of the exposure $T = t$ and the value of the counterfactual outcome if the mediator assumed whatever value it would have taken at a reference level of the exposure $T = t^*$: $Y(t, M(t))\breve{}Y(t, M(t^*))$ (VanderWeele and Vansteelandt, 2009)

C. Notations

- Coalition-wise Shapley *values* $\phi^f_{S,T}$ are the individual terms for a coalition in the weighted sum of the original definition of Shapley value, hence the term, *value*. It is common to use ϕ -even if specific to a coalition S- in reference to Shapley values.
- Coalition-wise Shapley *effects* $\Psi_{T \to Y|C_S}$ are the causal *effects* identified in coalition-wise Shapley *values* after dividing by the propensity weights. We use Ψ for effects, and describe the effect of T on Y along the multiple paths through the covariates C_S. We symbolise this by using the subscript $T \to Y|C_S$.
- Path-wise Shapley *effects* Ψ_{C_i} are similar to coalition-wise *effects* since they are also obtained after dividing by the weights. However, the conditioning is only on the features on a single causal pathway. We thus still use Ψ as it is an effect, but show that the conditioning only bears upon a path using the subscript C_i.

D. Further Related Work

D.1. Shapley Values: Axioms

In this section, in line with Section 2, the Shapley value $\phi^f_j(x)$ is always taken with respect to value function v, unless specified otherwise. In the latter case, if the value function is v', then the Shapley value is noted $\phi^f_j(x; v')$ instead. Shapley values have been shown to satisfy the four following axioms.

Dummy: A feature j receives a zero attribution if it has no possible contribution, i.e. $v_f(S \cup \{j\}, x) = v_f(S, x)$ for all $S \subseteq \{1, \ldots, m\}$.

Symmetry: Two features that always have the same contribution receive equal attribution, i.e. $v_f(S \cup \{i\}, x) = v(S \cup \{j\}, x)$ for all S not containing i or j then $\phi_i^f(x) = \phi_j^f(x)$.

Efficiency: The attributions of all features sum to the total value of all features. Formally, $\sum_j \phi_j^f(x) = v_f(\{1, .., m\}, x)$.

Linearity: For any value function v that is a linear combination of two other value functions u and w (i.e. $v(S) = \alpha u(S) + \beta w(S)$), the Shapley values of v are equal to the corresponding linear combination of the Shapley values of u and w (i.e. $\phi_i^f(x; v) = \alpha\phi_i^f(x; u) + \beta\phi_i^f(x; w)\Big)$.

D.2. Causal Shapley Values

Heskes et. al introduced the Causal Shapley values in 2020 (Heskes et al., 2020). For a coalition S, the contribution of feature j $\phi_{j,S}^f(x)$ is decomposed into a direct and an indirect effect:

$$\begin{aligned}\phi_{j,S}^f(x) &= \mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}}) \mid \text{do}(X_{S\cup\{j\}} = x_{S\cup\{j\}}) - \mathbb{E}[f(x_S, X_{\overline{S}})] \mid \text{do}(X_S = x_S)] \\ &= \mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}}) \mid \text{do}(X_S = x_S)] - \mathbb{E}[f(x_S, X_{\overline{S}})] \mid \text{do}(X_S = x_S)] \\ &+ \mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}}) \mid \text{do}(X_{S\cup\{j\}} = x_{S\cup\{j\}})] - \mathbb{E}[f(x_{S\cup\{j\}}, X_{\overline{S\cup\{j\}}}) \mid \text{do}(X_S = x_S)]\end{aligned}$$

where, in the last equality, the first line is the direct effect and the second line the indirect effect. The direct effect measures the expected change in prediction when the stochastic feature X_j is replaced by its feature value x_j, 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 $do\,(X_j = x_j)$. The direct and indirect parts of Shapley values are then be computed by taking a, possibly weighted, average over all coalitions.

We note that in the problem setup of Section 3.1 , if all covariates are pre-treatment then under mild assumptions the indirect effect of the treatment will be zero, as outlined in the following Proposition.

Property D.1 (Indirect part of Causal Shapley). Let S be a coalition containing pre-treatment covariates only. We assume that an unobserved (latent) variable generates all pre-treatment covariates. Then the indirect part of the Causal Shapley values of an exposure is null, i.e. we have

$$\mathbb{E}\left[f\left(C_{\bar{S}}, c_S, t\right) \mid \text{do}(C_S = c_S, T = t)\right] - \mathbb{E}\left[f\left(C_{\bar{S}}, c_S, t\right) \mid \text{do}(C_S = c_S)\right] = 0. \tag{2}$$

The proof can be found in Supplement J.7.

D.3. Edge-Based/Flow-Based Approaches to Shapley Values

Given that the proposed approach is model-agnostic, in this paragraph we will not review model-specific approaches that are considered to be "bespoke in nature and do not solve the problem of explainability in general" (Frye et al., 2019). Similarly, we do not review methods that violate *implementation invariance*[3]. Pan et. al (Pan et al., 2021) leverage Shapley values computation to define a new quantity that distributes credit for model disparity amongst the paths in a causal graph. However, the resulting quantity isn't a Shapley value itself. Shapley Flow (SF) assigns credits to "sink-to-node" paths. To do so, SF only considers orderings that are consistent with a depth first search. Furthermore, SF modifies the original definition of Shapley values by only explaining within successive cuts of the graphs or "explanation boundaries". Such cuts are considered as alternative models to be explained. Given this modification, it is unclear what connection the explanations generated by SF exhibit with the overall model. Recursive Shapley (Singal et al., 2021) is an edge-based approach which only considers active edges. Although it provides useful insights for mediation analysis, this method overlooks the impact of confounders. Ultimately, unlike our approach both Shapley Flow and Recursive Shapley aren't additive methods, but instead hold the property of "flow conservation" which allows a parent node to split its credit amongst its children. In contrast, the *efficiency* axiom of Shapley values ensure that the attributions of all features sum up to the model outcome $f(x)$. We argue that Shapley efficiency is more relevant in a regression setting, whereas flow conservation should be used for analysis of data with intrinsic ordering.

[3]Implementation invariance imposes that two black-box models that compute the same mathematical function have identical attributions for all features, regardless of how being implemented differently.

D.4. Other Causal Approaches to Interpreting Black-Box Models

Explaining black-box models in a causal manner remains challenging to this day. Zhao et. al (Zhao and Hastie, 2021) expand on the use of partial dependence plot (PDP), where the dependence on a set of covariates is computed by taking the expectation of the model over the marginal distribution of all other covariates. They note that the PDP formula is similar to Pearl's back-door adjustment. More specifically, marrying Shapley values with causal reasoning has been an active research question. Janzing et. al (Janzing et al., 2020) considers the model's prediction process itself to be a causal process: from features to model inputs and ultimately model output. The authors claim that marginal Shapley values can be apprehended in terms of *do*-calculus, if we consider that setting a feature to a given value is equivalent to intervening on it. As such, marginal or so-called off-manifold Shapley values may be sufficient to explain that specific causal process but this approach is contrived as it does not consider the real-world causal relationships between features. This approach however does not acknowledge any underlying causal structure from the real world. New causality based formulations of Shapley values have been proposed to compute feature attributions from a hypothesised causal structure of the data. Asymmetric Shapley values (Frye et al., 2019) use conditioning by observation but only consider causally-consistent coalitions i.e coalitions such that known causal ancestors precede their descendants. The resulting explanations quantify the impact a given feature has on model prediction while its descendants remain unspecified. As a result, they ignore downstream effects in favour of root causes (Wang et al., 2021).

Below is a table showing a summary comparison of existing causality-based or graph-based Shapley approaches with PWSHAP. Node efficiency refers to the original efficiency property of Shapley values: $f(x) = \phi_0^f(x) + \sum_{i=1}^{M} \phi_i^f(x)$, where $\phi_j^f(x)$ is the contribution of feature j to $f(x)$ and $\phi_0^f(x) = \mathbb{E}[f(X))]$ is the averaged prediction with the expectation over the observed data distribution. By game at each node/within each boundary we describe the fact that Shapley Flow and Recursive Shapley consider successive cuts from the graphs. Flow conservation or cut efficiency is the equivalent efficiency property, within such a cut. We refer the reader to the corresponding papers for further details.

Table 4: Comparison between PWSAP and existing causality-based or graph-based Shapley approaches

	Shapley Flow	Recursive Shapley	Asymmetric Shapley	Causal Shapley	PWSHAP
Flow conservation or cut efficiency	X	X			
Node efficiency			X	X	X
Node-based			X	X	X
Edge-based		X			
Source-to-sink path-based	X				
Path-based					X
Game at each node or within each boundary of explanation	X	X			
Ignores direct effects			X		
Fidelity to original Shapley			X (but violates symmetry axiom)	X	X

Ultimately, in a previous work (Sani et al., 2020), Sani et. al (2020) introduce an XAI method which uses auxiliary interpretable labels that are assumed to be readily available. Although their method has shown to perform well, it relies on the latter strong assumption which limits its applicability. Note that, like in PWSHAP, this approach further requires external information (besides the model and input/output data). We believe Sani et. al's method can be complementary to ours. For instance, in settings where the causal quantity of interest is identified from the obtained Partial Ancestral Graph, one may train a model regressing predictions on the interpretable labels and apply PWSHAP to the resulting model. We thank our anonymous reviewers for this remark.

D.5. Further discussion on an XAI model's dependency to the DAG

We view the fact that our approach is agnostic to the choice of (compatible) DAG as a strength, as it allows different experts to explain the black-box model output according to their own causal beliefs about the data or phenomenon being studied. For example, take the classical DAG with confounders, a treatment and an outcome (in Figure 1, graph 2). The black-box model takes all confounders and the treatment as inputs. However it does not "know" whether the covariates are actually confounders or mediators, and generates its output regardless. In PWSHAP, given that we have the additional (external) knowledge of the DAG representing our beliefs, we would interpret the output as a conditional average treatment effect. More generally, a model only gives indications about causal quantities that are already identified with statistical quantities according to prior causal assumptions, as in the causal roadmap by Petersen and van der Laan (2014) (Petersen and van der Laan, 2014)). Although causal discovery using both the model outputs *and the data* can yield a set of candidate DAGs that share the same edges, in the form of a Partial Ancestral Graph (PAG), some directions may be missing and one may still need to choose a DAG amongst multiple alternatives.

D.6. Bias, Mediation and Moderation Analysis

Our approach has added value compared to existing methods for sensitivity, mediation and moderation analysis. In the following paragraph, we review the state of the art with regards to each of these objectives.

Moderation analysis A common approach to assess moderation is to (i) fit an Heterogeneous Treatment Effect (HTE) model that predicts an individual's treatment effect from a set of covariates (ii) find subgroups with similar treatment effects. Subgroups can be infered directly from the data (Imai and Ratkovic, 2013; Wang and Rudin, 2017), from the individual predicted treatment effects (Foster et al., 2011) or using statistical hypothesis tests (Athey and Imbens, 2016; Song and Chi, 2007; Holmes and Watson, 2018). The main drawback of subgroup findings methods is that they are prone under-powered and time-consuming. Holmes et. al (Holmes and Watson, 2018) introduce a partitioning method which controls the type I error, however it is still limited to comparing subgroups two by two. Another approach to moderation analysis is to use interpretable models to predict HTEs. Nilsson et. al (Nilsson et al., 2019) build two potential outcomes models (treated/untreated) and fit a regression model to predict their difference from the covariates of interest. Regression coefficients are ultimately used as a measure of moderation, but this solution is prone to model misspecification. Explanation methods, and in particular feature attribution models allow for a finer-grained understanding of the sources of heterogeneity. More recently, Wu et. al suggested to use Distillation to generate explanations of HTE models and assess moderation induced by each covariate (Wu et al., 2021). However, explanation models such as Distillation that involve building a simpler surrogate model have received criticism for *approximating* the target black-box function instead of explaining it (Rudin, 2019). Ultimately, HTE models are built without taking the causal structure into consideration the causal structure e.g. not conditioning on post-treatment features. To the best of our knowledge, our approach is the first method that can assign attribution to moderators directly from the outcome regression model whilst acknowledging the posited causal structure and the rules of *do*-calculus by Pearl.

Sensitivity analysis In most treatment effect estimation studies, it is assumed that all confounders of treatment and outcome are observed. This is a strong assumption and one might wonder whether results will be greatly perturbed or not by the presence of an unobserved confounder. Sensitivity analysis generally aims at determining what the impact of a given amount of unobserved confounding would be on causal conclusions of the study. In particular, as we have no access to the unobserved confounder, we make assumptions about its relationship with treatment and outcome. One line of work is to assume parameters for this relationship and infer the rest of the model when values of these parameters are fixed, e.g. via maximum likelihood (Veitch and Zaveri, 2020; Rosenbaum and Rubin, 1983; Imbens, 2003). As a result, one can check the change in treatment effects with fixed values of the unobserved confounder (Rosenbaum and Rubin, 1983) or draw contour plots showing the bias depending on the parameters (Veitch and Zaveri, 2020; Imbens, 2003). Another line of work assumes a fixed ratio between the propensity score with only observed covariates and a variation of the propensity score that also includes unobserved confounders (Rosenbaum, 2005; Tan, 2006; Jesson et al., 2022). This ratio quantifies how much hidden confounding is present (with a ratio of 1 being no hidden confounding) and is set by the user. As a result, one can deduce intervals for inference quantities like p-values or treatment effects from a given ratio. This can be leveraged to find the lowest ratio that makes the interval reach thresholds invalidating causal conclusions, e.g. 0.05 for a p-value or zero for the treatment effects. The higher the ratio has to be, the more robust to unobserved confounding the study. This idea is close to the E-value, a scalar metric representing the minimal amount of unobserved confounding needed to fully explain

away the treatment-outcone relationship (VanderWeele and Ding, 2017). Although sensitivity analysis can also be applied to assess the role of a given observed confounder (Veitch and Zaveri, 2020; Imbens, 2003), it remains different from our local bias analysis approach that only applies to observed confounders. However, unlike most sensitivity analysis methods, our approach does not rely on parameters or assumptions other than the joint distribution of the data, and it summarises the confounding of a given pre-treatment covariate in a single bias scalar.

Mediation analysis The state-of-the-art approach to mediation analysis is based on Natural Direct and Indirect Effects (VanderWeele and Vansteelandt, 2009). Computation of Natural Direct/Indirect Effects requires four assumptions (VanderWeele and Vansteelandt, 2009) : 1) no unmeasured confounding for the exposure-outcome relationship, 2) no unmeasured confounding for the mediator-outcome relationship, 3) no unmeasured confounding for the exposure-mediator relationship, 4) no mediator-outcome confounding that is itself affected by the exposure. These are strong hypotheses, with the latter typically being considered to be unrealistic. PWSHAP also implicitly relies on these assumptions. However, it is common to assume unconfoundedness of the exposure-outcome relationship. Indeed, the exposure is naturally or experimentally randomised in many mediation analysis settings, such as fairness studies (e.g. when sex or race are the exposure). Thus, without confounders, PWSHAP effects are causal in these settings. Further, our approach to mediation analysis is more faithful to causal inference than Causal Shapley. By comparing CDEs, we assess the effect of setting the given mediator to its value. By contrast, Causal Shapley breaks the relationship between treatment and mediator when intervening on the mediator in the indirect effect. A local mediation analysis example is given in Supplement E.3. Supplement F details the application of PWSHAP to dependent mediators or in presence of both confounders and mediators.

E. Detailed Results for the "Building Blocks" Examples

E.1. Local Moderation Analysis

In the following, we prove the result of the first example where treatment is randomised
We assume the following model

$$C_1, C_2 \sim \text{Uniform}(0,1), \quad C_1 \perp\!\!\!\perp C_2$$
$$Y = \beta T + \gamma_1 C_1 + \gamma_2 C_2 + \alpha_1 T C_1 + \alpha_2 T C_2 + \epsilon$$

with $\mathbb{E}[\epsilon|T, C_1, C_2] = 0$. Assuming the true outcome model is known, our aim is to explain the following black-box: $f^*(c_1, c_2, t) = \beta t + \gamma_1 c_1 + \gamma_2 c_2 + \alpha_1 t c_1 + \alpha_2 t c_2$. We further assume that treatment is randomised by taking $T \sim \text{Bernoulli}(p)$.

We note that, from Property 3.1,

$$\phi^{f^*,\text{obs}}_{T,\{C_1,C_2\}}(c_1,c_2,t) = \phi^{f^*,\text{causal}}_{T,\{C_1,C_2\}}(c_1,c_2,t) = (t-p)(\beta + \alpha_1 c_1 + \alpha_2 c_2)$$
$$\phi^{f^*,\text{obs}}_{T,\{C_1\}}(c_1,t) = \phi^{f^*,\text{causal}}_{T,\{C_1\}}(c_1,t) = (t-p)(\beta + \alpha_1 c_1 + \alpha_2/2)$$
$$\phi^{f^*,\text{obs}}_{T,\{C_2\}}(c_2,t) = \phi^{f^*,\text{causal}}_{T,\{C_2\}}(c_2,t) = (t-p)(\beta + \alpha_1/2 + \alpha_2 c_2)$$
$$\phi^{f^*,\text{obs}}_{T,\emptyset}(t) = \phi^{f^*,\text{causal}}_{T,\emptyset}(t) = (t-p)(\beta + \alpha_1/2 + \alpha_2/2)$$

All of these Causal Shapley values only correspond to direct effects, as the indirect effect is zero from Property D.1.

E.2. Local Bias Analysis

We compare PWSHAP with Causal Shapley on a bias analysis example where C_1 and C_2 are distributed as before, but we assume instead that treatment allocation depends only on one covariate C_1: $\mathbb{E}[T|C_1, C_2] = C_1^\alpha$ which implies that C_1 is both a confounder and a moderator whereas C_2 only acts as a moderator. PWSHAP effects are then given as:

$$\Psi^{f^*}_{\substack{T\leftarrow C_1\to Y,\\ C_1:T\to Y}} = \alpha_1(c_1 - \frac{\alpha+1}{\alpha+2}) - \gamma_1 \frac{\alpha+1}{2(\alpha+2)} \qquad \Psi^{f^*}_{C_2:T\to Y} = \alpha_2(c_2 - \frac{1}{2})$$

$$\Psi^{f^*}_{T\to Y|\emptyset} = \beta + \gamma_1\frac{\alpha+1}{2(\alpha+2)} + \alpha_1\frac{\alpha+1}{\alpha+2} + \frac{\alpha_2}{2}.$$

We note that $\mathbb{E}[\Psi^{f^*}_{C_2:T\to Y}(C_1,C_2)] = 0$ but $\mathbb{E}[\Psi^{f^*}_{\substack{T\leftarrow C_1\to Y,\\ C_1:T\to Y}}(C_1,C_2)] \neq 0$. This illustrates not only Lemma 6.2, but also to the following corollary where we do not assume the model is true.

Corollary E.1 (Integration of the local confounding effect, black-box model)**.** *Let C_1, C_2 be two pre-treatment covariates such that $C_2 \perp\!\!\!\perp T|C_1$. Then the integral of the local confounding effect w.r.t. C_2 on the joint distribution of covariates is null*

$$\mathbb{E}[\Psi^{f}_{T\leftarrow C_2\to Y}(C_1,C_2)] = 0.$$

The proof can be found in Supplement J.8. By comparison, in Causal Shapley values, only the direct part is non null:

$$\begin{aligned}\phi^{f^*,\text{causal}}_{T,\text{direct}} = \beta\big(t - \frac{c_1^\alpha}{2} - \frac{1}{2(\alpha+1)}\big) + \alpha_1\big(\frac{c_1}{2}(t-c_1^\alpha) + \frac{1}{2}(\frac{t}{2} - \frac{1}{\alpha+2})\big)\\ + \alpha_2\big((\frac{c_2}{3}+\frac{1}{12})(t-c_1^\alpha) + (\frac{c_2}{6}+\frac{1}{6}(t-\frac{1}{\alpha+1})\big).\end{aligned}$$

Proof : First let's note that

$$\mathbb{E}[Y|T=1,C_1=c_1,C_2=c_2] - \mathbb{E}[Y|T=0,C_1=c_1,C_2=c_2] = \beta + \alpha_1 c_1 + \alpha_2 a_2$$

Then we show that :

$$\begin{aligned}&\mathbb{E}[Y|T=1,C_1=c_1] - \mathbb{E}[Y|T=0,C_1=c_1] = \beta + \alpha_1 c_1 + \frac{a_2}{2}\\ &\mathbb{E}[Y|T=1,C_2=c_2] - \mathbb{E}[Y|T=0,C_2=c_2] = \beta + \gamma_1\frac{\alpha+1}{2(\alpha+2)} + \alpha_1\frac{\alpha+1}{\alpha+2} + \alpha_2 c_2\\ &\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] = \beta + \gamma_1\frac{\alpha+1}{2(\alpha+2)} + \alpha_1\frac{\alpha+1}{\alpha+2} + \frac{\alpha_2}{2}\end{aligned}$$

First, let us note that, using independence of C_1 and C_2,

$$\begin{aligned}&\mathbb{E}[T|C_1=c_1] = \mathbb{E}[\mathbb{E}[T|C_1=c_1,C_2]|C_1=c_1] = \mathbb{E}[c_1^\alpha|C_1=c_1] = c_1^\alpha\\ &\mathbb{E}[T|C_2=c_2] = \mathbb{E}[\mathbb{E}[T|C_2=c_2,C_1]|C_2=c_2] = \mathbb{E}[C_1^\alpha|C_2=c_2] = \frac{1}{\alpha+1}\\ &\mathbb{E}[T] = \frac{1}{\alpha+1}\end{aligned}$$

By Bayes's rule and independence of C_1 and C_2,

$$p(c_2|c_1,t=1) = \frac{p(t=1|c_1,c_2)p(c_1)p(c_2)}{p(t=1|c_1)p(c_1)} = \frac{p(t=1|c_1,c_2)}{p(t=1|c_1)} = \frac{c_1^\alpha}{c_1^\alpha} = 1$$

and, similarly,

$$p(c_2|c_1,t=0) = 1$$

As a result,

$$\mathbb{E}[C_2|c_1,t=1] = \mathbb{E}[C_2|c_1,t=0] = \frac{1}{2}$$

$$\mathbb{E}[C_2|c_1, t=1] - \mathbb{E}[C_2|c_1, t=0] = 0$$

and

$$\begin{aligned}\mathbb{E}[Y|T=1, C_1=c_1] - \mathbb{E}[Y|T=0, C_1=c_1] &= \beta + \gamma_2(\mathbb{E}[C_2|c_1, t=1] - \mathbb{E}[C_2|c_1, t=0]) + \alpha_1 c_1 + \alpha_2 \mathbb{E}[C_2|c_1, t=1] \\ &= \beta + \alpha_1 c_1 + \frac{a_2}{2}\end{aligned}$$

which proves the first equality. For the second equality, we have

$$p(c_1|c_2, t=1) = \frac{p(t=1|c_1, c_2)p(c_1)p(c_2)}{p(t=1|c_2)p(c_2)} = \frac{c_1^\alpha}{\frac{1}{\alpha+1}} = (\alpha+1)c_1^\alpha$$

and, similarly,

$$p(c_1|c_2, t=0) = \frac{1 - c_1^\alpha}{1 - \frac{1}{\alpha+1}}$$

Thereby, we obtain

$$\begin{aligned}\mathbb{E}[C_1|c_2, t=1] &= \frac{\alpha+1}{\alpha+2} \\ \mathbb{E}[C_1|c_2, t=0] &= \frac{\alpha+1}{2(\alpha+2)} \\ \mathbb{E}[C_1|c_2, t=1] - \mathbb{E}[C_1|c_2, t=0] &= \frac{\alpha+1}{2(\alpha+2)}\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[Y|T=1, C_2=c_2] - \mathbb{E}[Y|T=0, C_1=c_1] &= \beta + \gamma_1(\mathbb{E}[C_1|c_2, t=1] - \mathbb{E}[C_1|c_2, t=0]) \\ &\quad + \alpha_2 c_2 + \alpha_1 \mathbb{E}[C_1|c_2, t=1] \\ &= \beta + \gamma_1 \frac{\alpha+1}{2(\alpha+2)} + \alpha_2 c_2 + \alpha_1 \frac{\alpha+1}{\alpha+2}\end{aligned}$$

Similarly, for the third equality, we note that, as before,

$$\begin{aligned}p(c_2|t=1) &= 1 \\ p(c_2|t=0) &= 1 \\ p(c_1|t=1) &= (\alpha+1)c_1^\alpha \\ p(c_1|t=0) &= \frac{1 - c_1^\alpha}{1 - \frac{1}{\alpha+1}}\end{aligned}$$

which leads to, as before,

$$\begin{aligned}\mathbb{E}[C_2|t=1] &= \frac{1}{2} \\ \mathbb{E}[C_2|t=0] &= \frac{1}{2} \\ \mathbb{E}[C_2|t=1] - \mathbb{E}[C_2|c_1, t=0] &= 0 \\ \mathbb{E}[C_1|t=1] &= \frac{\alpha+1}{\alpha+2} \\ \mathbb{E}[C_1|t=0] &= \frac{\alpha+1}{2(\alpha+2)}\end{aligned}$$

$$\mathbb{E}[C_1|t=1] - \mathbb{E}[C_1|c_2, t=0] = \frac{\alpha+1}{2(\alpha+2)}$$

thereby

$$\begin{aligned}\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] &= \beta + \gamma_1(\mathbb{E}[C_1|T=1] - \mathbb{E}[C_1|T=0]) \\ &+ \gamma_2(\mathbb{E}[C_2|T=1] - \mathbb{E}[C_2|T=0]) \\ &+ \alpha_1\mathbb{E}[C_1|T=1] + \alpha_2\mathbb{E}[C_2|T=1] \\ &= \beta + \gamma_1\frac{\alpha+1}{2(\alpha+2)} + \alpha_1\frac{\alpha+1}{\alpha+2} + \frac{a_2}{2}\end{aligned}$$

Now, for Causal Shapley values, we can show that

$$\begin{aligned}\mathbb{E}[Y|\,\mathrm{do}(t,c_1,c_2)] - \mathbb{E}[Y|\,\mathrm{do}(c_1,c_2)] &= (t-c_1^\alpha)(\beta+\alpha_1c_1+\alpha_2c_2) \\ \mathbb{E}[Y|\,\mathrm{do}(t,c_1)] - \mathbb{E}[Y|\,\mathrm{do}(c_1)] &= \beta(t-c_1^\alpha) + \alpha_1c_1(t-c_1^\alpha) + \alpha_2(\frac{t}{2} - \frac{c_1^\alpha}{2}) \\ \mathbb{E}[Y|\,\mathrm{do}(t,c_2)] - \mathbb{E}[Y|\,\mathrm{do}(c_2)] &= \beta(t-\frac{1}{\alpha+1}) + \alpha_2c_2(t-\frac{1}{\alpha+1}) + \alpha_1(\frac{t}{2} - \frac{1}{\alpha+2}) \\ \mathbb{E}[Y|\,\mathrm{do}(t)] - \mathbb{E}[Y] &= \beta(t-\frac{1}{\alpha+1}) + \alpha_1(\frac{t}{2} - \frac{1}{\alpha+2}) + \alpha_2(\frac{t}{2} - \frac{1}{2(\alpha+1)})\end{aligned}$$

as $\mathbb{E}[TC_1|c_1] = c_1^{\alpha+1}$, $\mathbb{E}[TC_1|c_2] = \mathbb{E}[TC_1] = \frac{1}{\alpha+2}$, $\mathbb{E}[TC_2|c_1] = \frac{c_1^\alpha}{2}$, $\mathbb{E}[TC_2|c_2] = \frac{c_2}{\alpha+1}$, $\mathbb{E}[TC_2] = \frac{1}{2(\alpha+1)}$

E.3. Local Mediation Analysis

We compare PWSHAP with Causal Shapley on a mediation analysis example inspired by the Berkeley dataset (Bickel et al., 1975). An algorithm predicts the probability of success of an applicant to a college. In this example, $X = (T, Q, D)$ where T is the gender of the applicant Q is an exam result and D is the department. We assume $Q \sim \text{Uniform}(0,1)$, $D|T=0 \sim \text{Bernoulli}(0.8)$ and $D|T=1 \sim \text{Bernoulli}(0.2)$. D is a mediator of gender however Q is only an ancestor of the outcome, and not a mediator. Our black-box is the true outcome model, and $Y = \alpha_Q Q + \alpha_D D + \alpha_T T + \alpha_{DT} DT + \alpha_{QT} QT + \epsilon$, with $\mathbb{E}[\epsilon|D,Q,T] = 0$.

$$\Psi^{f^*}_{T\to D\to Y} = 0.6\alpha_D + \alpha_{DT}(d - \frac{1}{5}) \qquad \Psi^{f^*}_{T\to Q\to Y} = \alpha_{QT}(q - \frac{1}{2})$$

$$\begin{aligned}\Psi^{f^*}_{T\to Y|\emptyset} &= \alpha_T - 0.6\alpha_D + \frac{\alpha_{DT}}{5} + \frac{\alpha_{QT}}{2} \\ \phi^{f^*,\text{causal}}_{T,\text{direct}} &= \alpha_T(t-\frac{1}{2}) + \alpha_{DT}[\frac{d}{2}(t-\frac{1}{2}) + \frac{1}{2}(\frac{t}{2} - \frac{1}{10})] + \frac{\alpha_{QT}}{2}(t-\frac{1}{2})(q+\frac{1}{2}) \\ \phi^{f^*,\text{causal}}_{T,\text{indirect}} &= \frac{\alpha_D}{2}(\frac{3}{10} - \frac{3t}{5})\end{aligned}$$

We note that $\int_q \Psi^{f^*}_{T\to Q\to Y}(p,q)dp(q) = 0$ but $\int_d \Psi^{f^*}_{T\to D\to Y}(p,q)dp(d) \neq 0$.

Proof : Coalition-wise Shapley effects are :

$$\begin{aligned}\Psi^{f^*}_{T\to Y|D,Q}(d,q) &= \text{CDE}(d,q) \\ &= \mathbb{E}[Y|T=1,D=d,Q=q] - \mathbb{E}[Y|T=0,D=d,Q=q] \\ &= \alpha_T + \alpha_{DT}d + \alpha_{QT}q \\ \Psi^{f^*}_{T\to Y|D}(d) &= \text{CDE}(d) \\ &= \mathbb{E}[Y|T=1,D=d] - \mathbb{E}[Y|T=0,D=d]\end{aligned}$$

$$
\begin{aligned}
&= \alpha_T + \alpha_{DT} d + \alpha_{QT} \mathbb{E}[Q|T=1] \\
&= \alpha_T + \alpha_{DT} d + \alpha_{QT} \mathbb{E}[Q] \\
&= \alpha_T + \alpha_{DT} d + \alpha_{QT} \frac{1}{2} \\
\Psi^{f^*}_{T\to Y|Q}(q) &= \text{CDE}(q) \\
&= \mathbb{E}[Y|T=1,Q=q] - \mathbb{E}[Y|T=0,Q=q] \\
&= \alpha_T + \alpha_{DT}\mathbb{E}[D|T=1] + \alpha_{QT} q + \alpha_D(\mathbb{E}[D|T=1] - \mathbb{E}[D|T=0]) \\
&= \alpha_T + \frac{\alpha_{DT}}{5} + \alpha_{QT} q - \alpha_D \frac{3}{5} \\
\Psi^{f^*}_{T\to Y|\emptyset}(q) &= \text{ATE} \\
&= \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] \\
&= \alpha_T + \alpha_{DT}\mathbb{E}[D|T=1] + \alpha_{QT}\mathbb{E}[Q|T=1] + \alpha_D(\mathbb{E}[D|T=1] - \mathbb{E}[D|T=0]) \\
&= \alpha_T + \frac{\alpha_{DT}}{5} + \frac{\alpha_{QT}}{2} - \alpha_D \frac{3}{5}
\end{aligned}
$$

As a result, we deduce path-wise effects

$$
\begin{aligned}
\Psi^{f^*}_{T\to D\to Y} &= \Psi^{f^*}_{T\to Y|D,Q} - \Psi^{f^*}_{T\to Y|Q} = 0.6\alpha_D + \alpha_{DT}(d - \frac{1}{5}) \\
\Psi^{f^*}_{T\to Q\to Y} &= \Psi^{f^*}_{T\to Y|D,Q} - \Psi^{f^*}_{T\to Y|D} = \alpha_{QT}(q - \frac{1}{2})
\end{aligned}
$$

Causal Shapley values are :

$$
\begin{aligned}
\phi^{f^*,\text{causal}}_{T,\{D,Q\},\text{direct}}(d,q,t) &= (\alpha_T + \alpha_{DT} d + \alpha_{QT} q)(t - \frac{1}{2}) \\
&= \mathbb{E}[f(d,q,t)|\,\text{do}(d,q)] - \mathbb{E}[f(d,q,T)|\,\text{do}(d,q)] \\
&= \alpha_T(t - \mathbb{E}[T]) + \alpha_{DT}(dt - d\mathbb{E}[T]) + \alpha_{QT}(qt - q\mathbb{E}[T]) \\
&= (t - \frac{1}{2})(\alpha_T + \alpha_{DT} d + \alpha_{QT} q) \\
\phi^{f^*,\text{causal}}_{T,\{D,Q\},\text{indirect}}(d,q,t) &= \mathbb{E}[f(d,q,t)|\,\text{do}(d,q,t)] - \mathbb{E}[f(d,q,t)|\,\text{do}(d,q)] \\
&= 0 \\
\phi^{f^*,\text{causal}}_{T,\{D\},\text{direct}}(d,t) &= \mathbb{E}[f(d,Q,t)|\,\text{do}(d)] - \mathbb{E}[f(d,Q,T)|\,\text{do}(d)] \\
&= \alpha_T(t - \mathbb{E}[T]) + \alpha_{DT}(dt - d\mathbb{E}[T]) + \alpha_{QT}(\mathbb{E}[Q]t - \mathbb{E}[QT]) \\
&= (\alpha_T + \alpha_{DT} d + \frac{\alpha_{QT}}{2})(t - \frac{1}{2}) \\
\phi^{f^*,\text{causal}}_{T,\{D\},\text{indirect}}(d,t) &= \mathbb{E}[f(d,Q,t)|\,\text{do}(d,t)] - \mathbb{E}[f(d,Q,t)|\,\text{do}(d)] \\
&= 0 \\
\phi^{f^*,\text{causal}}_{T,\{Q\},\text{direct}}(q,t) &= \mathbb{E}[f(D,q,t)|\,\text{do}(q)] - \mathbb{E}[f(D,q,T)|\,\text{do}(q)] \\
&= \alpha_T(t - \mathbb{E}[T]) + \alpha_{DT}(\mathbb{E}[D]t - \mathbb{E}[DT]) + \alpha_{QT}(qt - q\mathbb{E}[T]) \\
&= \alpha_T(t - \frac{1}{2}) + \alpha_{DT}(\frac{t}{2} - \frac{1}{10}) + \alpha_{QT} q(t - \frac{1}{2}) \\
\phi^{f^*,\text{causal}}_{T,\{Q\},\text{indirect}}(q,t) & \\
&= \mathbb{E}[f(D,q,t)|\,\text{do}(q,t)] - \mathbb{E}[f(D,q,t)|\,\text{do}(q)]
\end{aligned}
$$

$$
\begin{aligned}
&= \alpha_D(\mathbb{E}[D|t] - \mathbb{E}[D]) \\
&= \alpha_D(\frac{3}{10} - \frac{3t}{5}) \\
\phi^{f^*,\text{causal}}_{T,\emptyset,\text{direct}}(t) &= \mathbb{E}[f(D,Q,t)] - \mathbb{E}[f(D,Q,T)] \\
&= \alpha_T(t - \mathbb{E}[T]) + \alpha_{DT}(\mathbb{E}[D]t - \mathbb{E}[DT]) + \alpha_{QT}(\mathbb{E}[Q]t - q\mathbb{E}[QT]) \\
&= \alpha_T(t - \frac{1}{2}) + \alpha_{DT}(\frac{t}{2} - \frac{1}{10}) + \frac{\alpha_{QT}}{2}(t - \frac{1}{2}) \\
\phi^{f^*,\text{causal}}_{T,\emptyset,\text{indirect}}(t) & \\
&= \mathbb{E}[f(D,Q,t)|\,\text{do}(t)] - \mathbb{E}[f(D,Q,t)] \\
&= \alpha_D(\mathbb{E}[D|t] - \mathbb{E}[D]) \\
&= \alpha_D(\frac{3}{10} - \frac{3t}{5})
\end{aligned}
$$

Summing them altogether with appropriate binomial weights, we have

$$
\begin{aligned}
\phi^{f^*,\text{causal}}_{T,\text{direct}} &= \alpha_T(t - \frac{1}{2}) + \alpha_{DT}[\frac{d}{2}(t - \frac{1}{2}) + \frac{1}{2}(\frac{t}{2} - \frac{1}{10})] + \frac{\alpha_{QT}}{2}(t - \frac{1}{2})(q + \frac{1}{2}) \\
\phi^{f^*,\text{causal}}_{T,\text{indirect}} &= \frac{\alpha_D}{2}(\frac{3}{10} - \frac{3t}{5})
\end{aligned}
$$

F. Further Results for Complex "Building Blocks" DAGS

F.1. A Mix of Confounders and Mediators

We now assume the model :

$$
\begin{aligned}
C_1 &\sim \text{Bernoulli}(\frac{1}{2}) \\
C_2 &\sim \text{Bernoulli}(\frac{1}{2}) \\
Q|C_1 &\sim \text{Bernoulli}(1 - C_1) \\
T|C_1 &\sim \text{Bernoulli}(C_1) \\
D|T,C_1 &\sim \text{Bernoulli}(\frac{4}{5} - \frac{3}{5}\frac{T + C_1}{2}) \\
Y &= \alpha_Q Q + \alpha_D D + \alpha_T T + \alpha_1 C_1 + \alpha_2 C_2 + \alpha_{DT} DT + \alpha_{QT} QT \\
&\quad + \alpha_{1T} C_1 T + \alpha_{2T} C_2 T + \epsilon, \text{ with } \mathbb{E}[\epsilon|D,Q,T] = 0.
\end{aligned}
$$

We have,

$$
\begin{aligned}
\Psi^{f^*}_{T\to Y|C_1,C_2,D,Q}(c_1,c_2,d,q) &= \alpha_T + \alpha_{DT}d + \alpha_{QT}q + \alpha_{1T}c_1 + \alpha_{2T}c_2 \\
\Psi^{f^*}_{T\to Y|C_1,C_2,Q}(c_1,c_2,q) &= \alpha_T + \alpha_{DT}\mathbb{E}[D|T=1,c_1,c_2,q] + \alpha_{QT}q + \alpha_{1T}c_1 + \alpha_{2T}c_2 \\
&\quad + \alpha_D(\mathbb{E}[D|T=1,c_1,c_2,q] - \mathbb{E}[D|T=0,c_1,c_2,q]) \\
&= \alpha_T + \alpha_{DT}(\frac{1}{2} - \frac{3c_1}{10}) + \alpha_{QT}q + \alpha_{1T}c_1 + \alpha_{2T}c_2 - \frac{3\alpha_D}{10} \\
\Psi^{f^*}_{T\to Y|C_1,C_2,D}(c_1,c_2,d) &= \alpha_T + \alpha_{DT}d + \alpha_{QT}\mathbb{E}[D|T=1,c_1,c_2,d] + \alpha_{1T}c_1 + \alpha_{2T}c_2 \\
&\quad + \alpha_Q(\mathbb{E}[Q|T=1,c_1,c_2,d] - \mathbb{E}[Q|T=0,c_1,c_2,d]) \\
&= \alpha_T + \alpha_{DT}d + \alpha_{QT}(1 - c_1) + \alpha_{1T}c_1 + \alpha_{2T}c_2
\end{aligned}
$$

$$\Psi^{f^*}_{T\to Y|C_1,D,Q}(c_1,d,q) = \alpha_T + \alpha_{DT}d + \alpha_{QT}q + \alpha_{1T}c_1 + \frac{\alpha_{2T}}{2}$$
$$\Psi^{f^*}_{T\to Y|C_2,D,Q}(c_2,d,q) = \alpha_T + \alpha_{DT}d + \alpha_{QT}q + \alpha_{2T}c_2 + \alpha_{1T}\mathbb{E}[C_1|T=1,d,q]$$
$$+ \alpha_1(\mathbb{E}[C_1|T=1,d,q] - \mathbb{E}[C_1|T=0,d,q])$$
$$\text{with, in the general case, } \mathbb{E}[C_1|T=t,d,q] \neq 0 \ \ \forall t$$
$$\text{and } \mathbb{E}[C_1|T=1,d,q] - \mathbb{E}[C_1|T=0,d,q] \neq 0$$

Thereby,

1. Local mediating effects are given as

$$\Psi^{f^*}_Q(c_1,c_2,d,q) = \alpha_{QT}(q-(1-c_1)) \text{ and } \Psi^{f^*}_D(c_1,c_2,d,q) = \frac{3\alpha_D}{10} + \alpha_{DT}(\frac{3c_1}{10} - \frac{1}{2})$$

   so $\mathbb{E}[\Psi^{f^*}_Q(c_1,c_2,d,Q)|c_1,c_2] = 0$ but $\mathbb{E}[\Psi^{f^*}_D(c_1,c_2,D,q)|c_1,c_2] \neq 0$. This illustrates the relevance of Property 6.1 to isolate the fact that Q is not an actual mediator conditionally on confounders.
2. Local confounding effects are given as

$$\Psi^{f^*}_{C_2}(c_1,c_2,d,q) = \alpha_{2T}(c_2 - \frac{1}{2})$$
$$\Psi^{f^*}_{C_1}(c_1,c_2,d,q) = \alpha_1(\mathbb{E}[C_1|T=0,d,q] - \mathbb{E}[C_1|T=1,d,q]) - \alpha_{1T}\mathbb{E}[C_1|T=1,d,q]$$

   so $\mathbb{E}[\Psi^{f^*}_{C_2}(C_1,C_2,d,q)] = 0$ but $\mathbb{E}[\Psi^{f^*}_{C_1}(C_1,C_2,D,q)] \neq 0$. This illustrates the relevance of the two following results, which themselves generalise results of Section 6.1, to isolate the fact that C_2 is not an actual confounder of the relationship between treatment-outcome, treatment-mediator and mediator-outcome relationships.

Lemma F.1 (Integration of the local confounding effect with mediators, true model)**.** *Let M denote post-treatment and pre-outcome variables. Define $\mathcal{H}(C)$ as follows :*

$$\mathcal{H}(C): \ \forall t,m,\ Y(t,m) \perp\!\!\!\perp T|C \text{ and } Y(t,m) \perp\!\!\!\perp M|T,C,$$

or, in other words, C includes all confounders of the treatment-outcome and mediator-outcome relationships. We further assume consistency of the potential outcome, i.e. $Y(T,M) = Y$. Let C_1, C_2 be two pre-treatment covariates such that $\mathcal{H}(C_1,C_2)$ holds. If, additionally, C_2 is not a confounder, i.e. $\mathcal{H}(C_1)$ holds, then the integral of the local confounding effect of f^* w.r.t. C_2 on the joint distribution of covariates for fixed values of mediators is null, i.e.

$$\forall m,\ \ \mathbb{E}[\Psi^{f^*}_{C_2}(C_1,C_2,m)] = 0.$$

Corollary F.2 (Integration of the local confounding effect with mediators, black-box model)**.** *Let C_1, C_2 be two pre-treatment covariates and M post-treatment and pre-outcome variables such that $C_2 \perp\!\!\!\perp T, M|C_1$. Then the integral of the local confounding effect w.r.t. C_2 on the joint distribution of covariates is null, i.e.*

$$\forall m,\ \ \mathbb{E}[\Psi^{f}_{C_2}(C_1,C_2,m)] = 0.$$

The proofs can be found in Supplements J.9 and J.10

F.2. Dependent Mediators

We now assume the model :

$$Q \sim \text{Uniform}(0,1)$$
$$T \sim \text{Bernoulli}(0.5)$$
$$D|T,Q \sim \text{Bernoulli}(\frac{4}{5} - \frac{3}{5}\frac{T+Q}{2})$$
$$Y = \alpha_Q Q + \alpha_D D + \alpha_T T + \alpha_{DT}DT + \alpha_{QT}QT + \epsilon, \quad \mathbb{E}[\epsilon|D,Q,T] = 0.$$

We have

$$\begin{aligned}
\Psi^{f^*}_{T\to Y|D,Q}(d,q) &= \alpha_T + \alpha_{DT}d + \alpha_{QT}q \\
\Psi^{f^*}_{T\to Y|Q}(q) &= \alpha_T + \alpha_{DT}\mathbb{E}[D|T=1,q] + \alpha_{QT}q + \alpha_D(\mathbb{E}[D|T=1,q] - \mathbb{E}[D|T=0,q]) \\
&= \alpha_T + \alpha_{DT}(\frac{1}{2} - \frac{3q}{10}) + \alpha_{QT}q - \frac{3\alpha_D}{10} \\
\Psi^{f^*}_{T\to Y|D}(d) &= \alpha_T + \alpha_Q(\mathbb{E}[Q|T=1,d] - \mathbb{E}[Q|T=0,d]) + \alpha_{DT}d + \alpha_{QT}\mathbb{E}[Q|T=1,d] \\
&= \alpha_T - \frac{3\alpha_Q}{(13-6d)(7+6d)} + \alpha_{DT}d + \alpha_{QT}\frac{7-4d}{13-6d} \\
\Psi^{f^*}_{T\to Y|\emptyset} &= \alpha_T + \alpha_{DT}\mathbb{E}[D|T=1] + \alpha_{QT}q + \alpha_D(\mathbb{E}[D|T=1] - \mathbb{E}[D|T=0]) \\
&= \alpha_T + \frac{7\alpha_{DT}}{20} + \alpha_{QT}q - \frac{3\alpha_D}{10}
\end{aligned}$$

where we used

$$\begin{aligned}
\mathbb{E}[Q|T=1,d] &= \frac{7-4d}{13-6d} \\
\mathbb{E}[Q|T=0,d] &= \frac{4+2d}{7+6d} \\
\mathbb{E}[Q|T=1,d] - \mathbb{E}[Q|T=0,d] &= \frac{-3}{(13-6d)(7+6d)}
\end{aligned}$$

which we prove from

$$\begin{aligned}
p(d|q,t) &= dp(d=1|q,t) + (1-d)p(d=0|q,t) \\
&= d(\frac{4}{5} - \frac{3}{10}(t+q)) + (1-d)(\frac{1}{5} + \frac{3}{10}(t+q)) \\
p(d|t) &= \mathbb{E}[p(d|Q,t)|t] \\
&= \mathbb{E}[p(d|Q,t)] \text{ as } Q \perp\!\!\!\perp T \\
&= \frac{7}{20} + \frac{3t}{10} + \frac{3d}{5}(\frac{1}{2} - t) \\
p(q|d,t) &= \frac{p(d|q,t)p(q|t)p(t)}{p(d|t)p(t)} \\
&= \frac{p(d|q,t)}{p(d|t)} \\
&= \frac{1 + \frac{3}{2}(t+q) + 3d(1-t-q)}{\frac{7}{4} + \frac{3t}{2} + 3d(\frac{1}{2} - t)} \\
\mathbb{E}[Q|T=1,d] &= \int \frac{1 + \frac{3}{2} + \frac{3q}{2} - 3dq}{\frac{7}{4} + \frac{3}{2} - \frac{3d}{2}} q dq \\
&= \frac{\frac{1}{2}\frac{5}{2} + \frac{1}{3}(\frac{3}{2} - 3d)}{\frac{13}{4} - \frac{3d}{2}} \\
&= \frac{7-4d}{13-6d} \\
\mathbb{E}[Q|T=0,d] &= \int \frac{1 + 3d + q(\frac{3}{2} - 3d)}{\frac{7}{4} + \frac{3d}{2}} q dq \\
&= \frac{\frac{1}{2}(1+3d) + \frac{1}{3}(\frac{3}{2} - 3d)}{\frac{7}{4} + \frac{3d}{2}} \\
&= \frac{4+2d}{7+6d}
\end{aligned}$$

Thereby,

$$\begin{aligned}\Psi_Q^{f^*}(d,q) &= \Psi^{f^*}_{T\to Y|D,Q}(d,q) - \Psi^{f^*}_{T\to Y|D}(D) \\ &= \frac{3\alpha_Q}{(13-6d)(7+6d)} - \alpha_{QT}\frac{7-4d}{13-6d}\end{aligned}$$

Notably, we note that $\mathbb{E}[\Psi_Q^{f^*}(d,Q)] \neq 0$. Thereby, the local mediating effect as defined in Definition 6.4 is not able to isolate the absence of mediation from Q. However, defining

$$\Psi^f_{C_i,\text{alternative}}(q) := \Psi^f_{T\to Y|C_i}(c_i) - \Psi^f_{T\to Y|\emptyset} \tag{3}$$

we note that $\mathbb{E}[\Psi^{f^*}_{Q,\text{alternative}}(Q)] = 0$ and $\mathbb{E}[\Psi^{f^*}_{D,\text{alternative}}(D,q)] \neq 0$. Thereby, an alternative definition of the local mediating effect, as given in 3, is able to isolate the absence of mediation from Q. This actually holds in a more general setting.

Property F.1 (Ancestor of outcome). Let M be a post-treatment and pre-outcome variable. Assume that $M \perp\!\!\!\perp T$, or in other words M is not really a mediator. Then,

$$\mathbb{E}[\Psi^f_{T\to M\to Y}(M)] = 0$$

The proof can be found in Supplement J.11. Thereby, the original definition of the local mediating effect is not always suited for mediation analysis. However, picking up the alternative definition given above will mean we will not be able to use the same quantity for mediation, bias analysis and mediation analysis. Solving this dilemma is left for future work.

G. Generalisation to a Path of Length 3 or More

Assume the path p of length $L(p)+1$ is defined as $T \rightleftarrows C_{p(1)} \rightleftarrows C_{p(2)} \rightleftarrows \cdots \rightleftarrows C_{p(L(p))} \to Y$ where $\rightleftarrows$ is either $\leftarrow$ or $\to$ and there are no other paths from T into any of $C_{p(1)}, \cdots, C_{p(L(p))}$ or from any of $C_{p(1)}, \cdots, C_{p(L(p))}$ into Y. Then the path-wise Shapley effect with respect to p is

$$\Psi^f_p(c) = \Psi^f_{T\to Y|C_{S^*}}(c) - \Psi^f_{T\to Y|C_{S^*\setminus\{p(1),\ldots,p(L(p))\}}}(c_{S^*\setminus\{p(1),\ldots,p(L(p))\}}) \tag{4}$$

where the second term takes the coalition with all covariates in the path removed. Local confounding and moderating effects (Definitions 6.1 and 6.3, respectively) can be generalized to paths p such that $T \leftarrow C_{p(1)}$, and local mediating effects (Definition 6.4) to paths p such that $T \to C_{p(1)}$. Lemma 6.2 and Property 6.1 can be generalized with these effects, by replacing covariates in the conditional independence statements with all covariates in either path.

If there are other paths of the form $T \to C_{p(k)}$ or $C_{p(k)} \to Y$, then the effect from Equation 4 represents the effect of both p and the paths $T \rightleftarrows C_{p(1)} \rightleftarrows C_{p(2)} \rightleftarrows \cdots \rightleftarrows C_{p(k)} \to Y$ for all such k, as by grouping $C_{p(1)}, \cdots, C_{p(L(p))}$ into $C_{\{p(1),\ldots,p(L(p))\}}$, all these paths would be merged into a single path $T \rightleftarrows C_{\{p(1),\ldots,p(L(p))\}} \to Y$.

H. Further Experimental Results and Details

The code used for experiments is available at `https://github.com/oscarclivio/pwshap`.

H.1. Details of Experiments on Synthetic Datasets

Models : We model the outcome and propensity models using linear and logistic regression, respectively. Conditional distributions for PWSHAP are inferred by training an iterative imputer. Linear regressions with second order polynomial features are used to infer outcome models and logistic regressions for propensity models.

Causal Shapley : Regarding Causal Shapley, we model the $\mathbb{E}\left[f\left(X_{\bar{S}}, x_{S\cup\{j\}}\right) \mid do\left(X_S = x_S\right)\right]$ term that is added and subtracted to obtain the direct and indirect effects (see Supplement D.2) by using an iterative imputer on the dataset obtained by removing the treatment column from the original dataset. do−distributions are modelled differently depending on the situation. They involve making variables independent from others. This is made by reshuffling the columns of these variables in the train set, then training the imputer on this modified dataset again.

Datasets : datasets are generated according to the models in Supplement E.2 for local bias analysis and in Supplement E.3 for local mediation analysis. 200 samples are generated and between training and testing sets as a 50/50 split. We change the links between treatment and covariates to model confounding and mediation, in the following way :

Table 5: Results for additional experiment on Census Income dataset where treatment is the individual's occupation. Note that $\Psi^f_{\substack{\text{Occ}\leftarrow\text{Cntr}\rightarrow\text{Capg}\rightarrow\text{Inc}\\ \text{Occ}\leftarrow\text{Cntr}\rightarrow\text{Inc}}}$ is defined according to the generalization from Section G.

Causal SHAP		PWSHAP			
ϕ_{direct}	ϕ_{indirect}	$\Psi^f_{\text{Occ}\rightarrow\text{Inc}}$	$\Psi^f_{\text{Occ}\leftarrow\text{Cntr}\rightarrow\text{Inc}}$	$\Psi^f_{\substack{\text{Occ}\leftarrow\text{Cntr}\rightarrow\text{Capg}\rightarrow\text{Inc}\\ \text{Occ}\leftarrow\text{Cntr}\rightarrow\text{Inc}}}$	$\Psi^f_{\text{Relation:Occ}\rightarrow\text{Inc}}$
0.132 (0.22)	0 (0.027)	0.152 (1.98)	0.083	0.147 (1.13)	0.217 (0.831)

- Local bias analysis :
  - No confounders : $p(T = 1|C_1, C_2) = 0.5$
  - C_1 is a confounder, C_2 is not : $p(T = 1|C_1, C_2) = C_1$
  - C_1 and C_2 are confounders : $p(T = 1|C_1, C_2) = C_1C_2$
- Local mediation analysis :
  - No mediators : $Q|T \sim \text{Uniform}(0, 1), D|T \sim \text{Binomial}(0.5)$.
  - D is a mediator but not Q : $Q|T \sim \text{Uniform}(0, 1), D|T \sim \text{Binomial}(\frac{4}{5} - \frac{3}{5}T)$
  - Q and D are mediators : $D|T \sim \text{Binomial}(\frac{4}{5} - \frac{3}{5}T)$ and $Q = \frac{3}{5} \cdot T \cdot U + (1 - T) \cdot (\frac{3}{5} \cdot U + \frac{2}{5})$ where $U \sim \text{Uniform}(0, 1)$

H.2. Further Results on the UCI Dataset

In an additional experiment, we consider the treatment of interest to the occupation of the individual. We focus on the question: *"Under what mechanisms did having a managerial occupation impact the model prediction?"*. Results comparing PWSHAP with Causal Shapley are shown in Table 5. Occupation seems to have a considerable impact throughout the cohort, with the base treatment effect showing that having a managerial job increases the predicted probability of high income by 15.2 points. Moderation by relationship status had a predominant effect in the model prediction for our individual. Being unmarried has increased the positive effect of having a managerial occupation by another 21.7 points. CausalSHAP does not capture this phenomenon as all the effect of occupation is deemed direct by definition. This further shows the high resolution of PWSHAP and impossibility of explaining. The local confounding effect of the pathway through country and capital gain also had an important impact.

H.3. Experimental Details on UCI

UCI dataset The UCI dataset –also known as "Census Income" dataset– predicts whether income exceeds \$50K/yr based on census data. It includes 11 features and 32,561 individuals. We select a random subsample of 5000 individuals and only consider the following 7 features: age, capital gain, native country, income, marital status, race and relationship. The occupation (Occ) and race variables were dichotomised, respectively into managerial/non-managerial and white/non-white. Other categorical features, namely native-country, marital status and relationship were encoded into numerical values. The URL for this dataset is `https://archive.ics.uci.edu/ml/datasets/adult`.

Pre-processing and model performance We use a Random Forest with 500 estimators and balanced class weights for all three models: the outcome model, the race propensity score model, the occupation propensity score model. We further used a Bayesian Ridge to learn the joint distribution on all covariates. Note that we could have inferred a propensity score model from the Bayesian Ridge but decided to fit a separate model so both weights and outcome models would be modelled with a classification tree.
The 5000 observations in the set were subsampled again 10 times, by taking out 20% of the sample. Each time, we use the subsample as both a training and testing set for our models. We further use it as a reference population. The goal is to explain how the model learned on this set, therefore using the same training set for testing isn't problematic. We ultimately want a model that has high accuracy whilst limiting the compute time. Means and standard deviations over all 10 subsamples are reported.

The accuracy of our outcome model is 0.827 and the AUC 0.944.
The accuracy of our race propensity score model is 0.894 and the AUC 0.950.
The accuracy of our occupation propensity score model is 0.757 and the AUC 0.863.

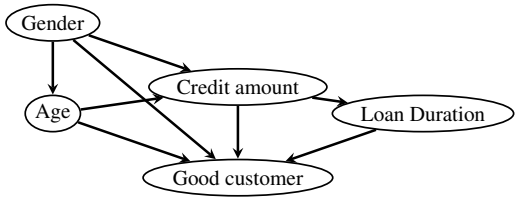

Figure 4: DAG of the German Credit Dataset.

Table 6: Results on the German Credit Dataset. Note that the last PWSHAP effect follows the generalisation from Appendix G.

Causal SHAP	ϕ_{direct}	0.025
	ϕ_{indirect}	0.004
PWSHAP	$\Psi_{\text{Gender} \xrightarrow{\text{total}} \text{Good}}$	0.069
	$\Psi_{\text{Gender}\to\text{Amount}\to\text{Good}}$	-0.1819
	$\Psi_{\substack{\text{Gender}\to\text{Amount}\to\text{Duration}\to\text{Good} \\ \text{Gender}\to\text{Amount}\to\text{Good}}}$	-0.2371

Computation and packages Random Forest and Bayesian Ridge were implemented using the `sklearn` package. We use the `dataset_fetcher` function from the Explanation GAME (Merrick and Taly, 2020), available at `https://github.com/fiddler-labs/the-explanation-game-supplemental` (no license). Experiments were ran using a 2,6 GHz 6-Core Intel Core i7. The amount of compute time was approximately 2,500 clock-time seconds.

H.4. Experiments on the German Credit Dataset

Here, we apply PWSHAP to a local mediation analysis example on the German Credit Dataset (Karimi et al., 2021). This dataset assigns people described by age and gender (1 for male, 0 for female) and seeking a loan with a certain credit amount and a certain duration to a label about whether they are good customers (1 for yes, 0 for no). The DAG is given in Figure 4.

We consider a male candidate, aged 42, asking for a loan with credit amount 5507 euros and duration 24 years. This individual's loan application gets a probability of 0.713 from the black-box model, which is below the average probability of 0.728 in the training set. Results are shown in Figure 6. PWSHAP suggests that although the total effect of gender on decisions is slightly positive, the effect of gender on decisions through credit amount, or through credit amount and loan duration, is negative. This suggests that although being male might give a slight advantage to this individual, it is compensated by high credit amount and loan duration, which are in the 65-th and 81-th percentiles, respectively, producing a relatively bad prediction compared to the average. Causal Shapley notes both a positive direct effect and a small positive indirect effect of gender on credit amount, which is not consistent with the finding that the predicted probability of being a good customer is lower than average for this individual.

H.5. Computational Complexity

Complexity of PWSHAP effects : For ease of notation, assume that the size of the data and the number of Monte-Carlo samples for expectations are all of size $\mathcal{O}(n)$. We denote by p the number of covariates. We also assume the black-box model and the propensity score model require $\mathcal{O}(p)$ steps to be evaluated, as e.g. in linear/logistic regression, that we have access to all conditional distributions and that sampling any feature from other features can be done in $\mathcal{O}(p)$ too (e.g. if, again, all conditional distributions are GLMs).

For any sample i and any path of length $l+1$ (with the treatment, l covariates and the outcome), computing the path-wise PWSHAP effect mostly requires evaluating expectations of the black-box or propensity score model, where :

- we take $\mathcal{O}(n)$ Monte-Carlo samples of l (when the treatment is observed) or $l+1$ (when the treatment is missing) variables considered to be missing from other covariates (complexity $\mathcal{O}(nlp)$) ;
- we evaluate and aggregate the black-box for all these imputed points (complexity $\mathcal{O}(np)$).

Thus, the PWSHAP effect on a path of size $l+1$ for one sample has a complexity $\mathcal{O}(nlp)$. Aggregated over all samples,

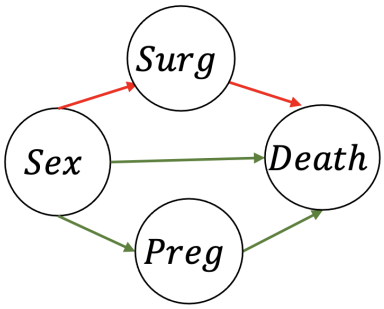

Figure 5: DAG for the running example

the PWSHAP of this path is of complexity $\mathcal{O}(n^2lp)$. Aggregated over all paths for all l, we have a complexity of at most $\mathcal{O}(n^2p^22^p)$. However, this can be reduced to $\mathcal{O}(n^2p^2)$ if the number of paths is small.

All of this assumes that we have access to conditional distributions. In practice, we do not and resort to an iterative imputer that relies on the MICE algorithm. Although, we could not find details on its complexity, scikit-learn documentation[4] suggests that imputation and thus sampling from the conditional distribution can be done in polynomial factors w.r.t. n and p too. If training the imputer is polynomial, then the overall complexity of computing all PWSHAP effects will remain polynomial in n and p.

Thus, the method might not scale well with p, but with few paths in the DAG it would scale better than exact classical observational/marginal Shapley values whose complexities have a 2^p factor.

Complexity of other Shapley value methods : with the same assumptions, exact on-manifold and off-manifold Shapley values have a $\mathcal{O}(n^2p^22^p)$ or $\mathcal{O}(n^2p2^p)$ complexity, respectively. Although the general computational complexity of the approximation made by KernelSHAP (Lundberg and Lee, 2017) is not explicited in their work, we expect it to give better complexity at the cost of further approximating Shapley values compared to direct evaluation of expectations. TreeSHAP (an explanation model for tree ensemble models only) (Lundberg et al., 2018) reduces the 2^p factor by a square factor in the maximal depth of trees, however this method isn't model agnostic like most Shapley approaches, including PWSHAP.

I. Running Example

Using the causal structure described in the DAG in Figure 5 as a running example, we illustrate our concepts. Here, we consider three variables: *Sex* (binary, denoting the female sex), *Surg* (binary, denoting the execution of a surgery) and *Preg* (binary, denoting whether the subject is pregnant). We denote *Death* the logit of the probability of death as a result of the surgery, of its absence. We analyse the effect of the sensitive attribute *Sex* on the *Death* outcome, with regards to model fairness. The model is as follows:

$$\begin{aligned}
\mathbb{E}[\text{Sex}] &= 0.5 \\
\mathbb{E}[\text{Preg}|\text{Sex}] &= p_{\text{Preg}} \cdot \text{Sex} \\
\mathbb{E}[\text{Surg}|\text{Sex}] &= p_{\text{Surg}} \cdot \text{Sex} \\
\mathbb{E}[\text{Death}|\text{Sex, Surg, Preg}] &= f(\text{Sex, Surg, Preg}) := \alpha_{\text{Sex}}\text{Sex} + \alpha_{\text{Surg}}\text{Surg} + \alpha_{\text{Preg}}\text{Preg}
\end{aligned}$$

We outline a few definitions and concepts applied to this example :

- On-manifold Shapley value for Sex on the empty coalition :

$$\begin{aligned}
\phi^{f,\text{obs}}_{\text{Sex},\emptyset}(\text{sex}) = &(\alpha_{\text{Sex}}\text{sex} + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}|\text{sex}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}|\text{sex}]) \\
&- (\alpha_{\text{Sex}}\mathbb{E}[\text{Sex}] + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}])
\end{aligned}$$

[4] https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html

- Off-manifold Shapley value for Sex on the empty coalition :

$$\begin{aligned}\phi^{f,\text{off-manifold}}_{\text{Sex},\emptyset}(\text{sex}) &= (\alpha_{\text{Sex}}\text{sex} + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}]) \\ &\quad - (\alpha_{\text{Sex}}\mathbb{E}[\text{Sex}] + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}]) \\ &= \alpha_{\text{Sex}}(\text{Sex} - \mathbb{E}[\text{Sex}])\end{aligned}$$

- Causal Shapley value for Sex on the empty coalition :

$$\begin{aligned}\phi^{f,\text{causal}}_{\text{Sex},\emptyset}(\text{sex}) &= (\alpha_{\text{Sex}}\text{sex} + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}|\text{do(Sex)}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}|\text{do(Sex)}]) \\ &\quad - (\alpha_{\text{Sex}}\mathbb{E}[\text{Sex}] + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}]) \\ &= (\alpha_{\text{Sex}}\text{sex} + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}|\text{sex}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}|\text{sex}]) \\ &\quad - (\alpha_{\text{Sex}}\mathbb{E}[\text{Sex}] + \alpha_{\text{Surg}}\mathbb{E}[\text{Surg}] + \alpha_{\text{Preg}}\mathbb{E}[\text{Preg}]) \\ &= \phi^{f,\text{obs}}_{\text{Sex},\emptyset}(\text{sex}) \quad \text{for this value in this specific example.}\end{aligned}$$

- Coalition-specific Shapley effect (for Sex, our treatment of interest) on the empty coalition : one can note that the on-manifold Shapley value can be decomposed as

$$\phi^{f,\text{obs}}_{\text{Sex},\emptyset}(\text{sex}) = (\text{sex} - \mathbb{E}[\text{Sex}])(\alpha_{\text{Sex}} + \alpha_{\text{Preg}} \cdot p_{\text{Preg}} + \alpha_{\text{Surg}} \cdot p_{\text{Surg}})$$

  where the second factor on the right-hand side is the coalition-specific Shapley effect, which here is written as

$$\Psi^{f}_{\text{Sex}\to\text{Death}|\emptyset} = \mathbb{E}[f(\text{sex} = 1, \text{Surg,Preg})] - \mathbb{E}[f(\text{sex} = 0, \text{Surg,Preg})]$$

- Path-wise Shapley effect of Surg : here it is expressed as

$$\Psi^{f}_{\text{Surg}} = \Psi^{f}_{\text{Sex}\to\text{Death | Surg, Preg}}(\text{surg, preg}) - \Psi^{f}_{\text{Sex}\to\text{Death| Preg}}(\text{preg})$$

  We can interpret as an effect through the path Sex $\to$ Surg $\to$ Death, re-noting it as $\Psi^{f}_{\text{Sex}\to\text{Surg}\to\text{Death}}$.

J. Proofs of Properties and Lemmas

J.1. Proof of Property 3.1

It suffices to show that

$$\phi^{f}_{T,S}(c,t) = w^{*}_{S} \times \big(v_f(S \cup \{T\}, c, 1) - v_f(S \cup \{T\}, c, 0)\big)$$

where

$$\begin{aligned}v_f(S \cup \{T\}, c, t) = v_f(S \cup \{T\}, c_S, t) &= \mathbb{E}_{p(C_{\bar{S}}|c_S,t)}[f(c_S, C_{\bar{s}}, t)] \\ v_f(S, c) &= \mathbb{E}_{p(C_{\bar{S}},T|c_S)}[f(c_S, C_{\bar{s}}, T)] \\ w^{*}_{S} = w^{*}(c_S, t) &= t - P(T = 1|C_S = c_S)\end{aligned}$$

In other words we want to show that

$$\phi^{f}_{T,S}(c,t) = (t - p(T = 1|C_S = c_S)) \times \big(\mathbb{E}_{p(C_{\bar{S}}|c_S,1)}[f(c_S, C_{\bar{s}}, 1)] - \mathbb{E}_{p(C_{\bar{S}}|c_S,0)}[f(c_S, C_{\bar{s}}, 0)]\big)$$

We note that

$$\begin{aligned}v_f(S,c) &= \mathbb{E}[f(c_S, C_{\bar{S}}, T)|c_S = c_S] \\ &= \mathbb{E}[\mathbb{E}[f(c_S, C_{\bar{S}}, T)|c_S = c_S, T]|c_S = c_S] \\ &= p(T = t|c_S) \times \mathbb{E}[f(c_S, C_{\bar{S}}, t)|c_S = c_S, T = t] \\ &\quad + p(T = 1 - t|c_S) \times \mathbb{E}[f(c_S, C_{\bar{S}}, 1 - t)|c_S = c_S, T = 1 - t] \\ &= p(T = t|c_S) \times v_f(S \cup \{T\}, c, t)\end{aligned}$$

$$+ p(T = 1 - t|c_S) \times v_f(S \cup \{T\}, c, 1 - t)$$

and, from $1 = p(T = t|c_S) + p(T = 1 - t|c_S)$,

$$\begin{aligned}
&v_f(S \cup \{T\}, c, t) - v_f(S, c) \\
&= p(T = t|c_S)v_f(S \cup \{T\}, c, t) + p(T = 1 - t|c_S)v_f(S \cup \{T\}, c, t) \\
&\quad - p(T = t|c_S)v_f(S \cup \{T\}, c, t) - p(T = 1 - t|c_S)v_f(S \cup \{T\}, c, 1 - t).
\end{aligned}$$

So terms in $p(T = t|c_S)$ cancel out and we get:

$$\begin{aligned}
v_f(S \cup \{T\}, c, t) - v_f(S, c) &= p(T = 1 - t|c_S) \times [v_f(S \cup \{T\}, c, t) - v_f(S \cup \{T\}, c, 1 - t)] \\
&= (t - \pi^*_S(c_S)) \times [v_f(S \cup \{T\}, c, 1) - v_f(S \cup \{T\}, c, 0)]
\end{aligned}$$

where $\pi^*_S(c_S) = P(T = 1|C_s = c_S)$ and the second equality comes from $t \in \{0, 1\}$.

J.2. Proof of Property 3.2

Let c_{-i} be a value of C_{-i}. It suffices to show that $\mathbb{E}[\Psi^f_{T\to Y|C_i,C_{-i}}(C_i, c_{-i})|C_{-i} = c_{-i}] = \Psi^f_{T\to Y|C_{-i}}(c_{-i})$, which is true as

$$\begin{aligned}
&\mathbb{E}[\Psi^f_{T\to Y|C_i,C_{-i}}(C_i, c_{-i})|C_{-i} = c_{-i}] \\
&= \int_{c_i} (v_f(S^* \cup \{T\}, c_i, c_{-i}, 1) - v_f(S^* \cup \{T\}, c_i, c_{-i}, 0))\mathrm{d}p(c_i|c_{-i}) \\
&= \int_{c_i} (f(c_i, c_{-i}, 1) - f(c_i, c_{-i}, 0))\mathrm{d}p(c_i|c_{-i}) \\
&= \int_{c_i} f(c_i, c_{-i}, 1)\mathrm{d}p(c_i|c_{-i}) - \int_{c_i} f(c_i, c_{-i}, 0)\mathrm{d}p(c_i|c_{-i}) \\
&= \int_{c_i} f(c_i, c_{-i}, 1)\mathrm{d}p(c_i|c_{-i}, t = 1) - \int_{c_i} f(c_i, c_{-i}, 0)\mathrm{d}p(c_i|c_{-i}, t = 0) \text{ as } T \perp\!\!\!\perp C_i|C_{-i} \\
&= \mathbb{E}[f(C_i, c_{-i}, 1)|C_{-i} = c_{-i}, T = 1] - \mathbb{E}[f(C_i, c_{-i}, 0)|C_{-i} = c_{-i}, T = 0] \\
&= v_f((S^*\backslash\{i\}) \cup \{T\}, c_{-i}, 1) - v_f((S^*\backslash\{i\}) \cup \{T\}, c_{-i}, 0) \\
&= \Psi^f_{T\to Y|C_{-i}}(c_{-i})
\end{aligned}$$

which completes the proof.

J.3. Proof of Property 5.1

We assume that

$$\forall c, t, N, |\hat{f}_N(c, t) - f^*(c, t)| \le e^{\text{outcome}}_N$$

Then, for any coalition S without T, c and N,

$$\begin{aligned}
&|\Psi^{\hat{f}_N}_{T\to Y|C_S}(c) - \Psi^{f^*}_{T\to Y|C_S}(c)| \\
&\le |(\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[\hat{f}_N(c_S, C_{\bar{S}}, 1)] - \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[\hat{f}_N(c_S, C_{\bar{S}}, 0)]) \\
&\quad - (\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[f^*(c_S, C_{\bar{S}}, 1)] - \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[f^*(c_S, C_{\bar{S}}, 0)])| \\
&= |(\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[\hat{f}_N(c_S, C_{\bar{S}}, 1) - f^*(c_S, C_{\bar{S}}, 1)] \\
&\quad - \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[\hat{f}_N(c_S, C_{\bar{S}}, 0) - f^*(c_S, C_{\bar{S}}, 0)])| \\
&\le |\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[\hat{f}_N(c_S, C_{\bar{S}}, 1) - f^*(c_S, C_{\bar{S}}, 1)]| \\
&\quad + |\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[\hat{f}_N(c_S, C_{\bar{S}}, 0) - f^*(c_S, C_{\bar{S}}, 0)]| \text{ from the triangle inequality}
\end{aligned}$$

$$
\begin{aligned}
&\leq \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[|\hat{f}_N(c_S, C_{\bar{S}}, 1) - f^*(c_S, C_{\bar{S}}, 1)|] \\
&\quad + \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[|\hat{f}_N(c_S, C_{\bar{S}}, 0) - f^*(c_S, C_{\bar{S}}, 0)|] \text{ from Jensen's inequality} \\
&\leq \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[e_N^{\text{outcome}}] + \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[e_N^{\text{outcome}}] \text{ by assumption} \\
&= 2e_N^{\text{outcome}}
\end{aligned}
$$

which shows bound 1. Now, let S as before, c, t, N,

$$
\begin{aligned}
&|\phi_{T,S}^{\hat{f}_N}(c,t) - \phi_{T,S}^{f^*}(c,t)| \\
&= |w_S^*(t, c_S) \cdot (\Psi_{T\to Y|C_S}^{\hat{f}_N}(c) - \Psi_{T\to Y|C_S}^{f^*}(c))| \\
&= |w_S^*(t, c_S)||\Psi_{T\to Y|C_S}^{\hat{f}_N}(c) - \Psi_{T\to Y|C_S}^{f^*}(c)| \\
&\leq |\Psi_{T\to Y|C_S}^{\hat{f}_N}(c) - \Psi_{T\to Y|C_S}^{f^*}(c)| \text{ as } |w_S^*(t, c_S)| \leq 1 \\
&\leq 2e_N^{\text{outcome}} \text{ from the previous bound}
\end{aligned}
$$

which proves bound 2. Now, for any covariate feature i as before, c, N

$$
\begin{aligned}
&|\Psi_{C_i}^{\hat{f}_N}(c) - \Psi_{C_i}^{f^*}(c)| \\
&= |\Psi_{T\to Y|C_{S^*}}^{\hat{f}_N}(c) - \Psi_{T\to Y|C_{S^*\setminus\{i\}}}^{\hat{f}_N}(c_{S^*\setminus\{i\}}) - (\Psi_{T\to Y|C_{S^*}}^{f^*}(c) - \Psi_{T\to Y|C_{S^*\setminus\{i\}}}^{f^*}(c_{S^*\setminus\{i\}}))| \\
&\leq |\Psi_{T\to Y|C_{S^*}}^{\hat{f}_N}(c) - \Psi_{T\to Y|C_{S^*}}^{f^*}(c)| \\
&\quad + |\Psi_{T\to Y|C_{S^*\setminus\{i\}}}^{\hat{f}_N}(c_{S^*\setminus\{i\}}) - \Psi_{T\to Y|C_{S^*\setminus\{i\}}}^{f^*}(c_{S^*\setminus\{i\}})| \\
&\quad \text{from the triangle inequality} \\
&\leq 2e_N^{\text{outcome}} + 2e_N^{\text{outcome}} \quad \text{from the bound on the coalition-wise Shapley effect}
\end{aligned}
$$

which proves bound 3.

J.4. Proof of Property 5.2

We assume that

$$
\forall S \text{ s.t. } T \notin S, c, N, |\hat{\phi}_{T,S}^{N,\hat{f}_N}(c,t) - \phi_{T,S}^{f^*}(c,t)| \leq e_N^{\text{Shap}},
$$

that the arbitrary propensity score model π^N and π^* verify ϵ-strong overlap, ie $\epsilon \leq \pi^N \leq 1-\epsilon$, and $\epsilon \leq \pi^* \leq 1-\epsilon$, and that we have $\forall c, N, |\pi^N(c) - \pi^*(c)| \leq e_N^{\text{propensity}}$.

Then, for any S s.t. $T \notin S$, we note that ϵ-strong overlap for π^N and π^* implies ϵ-strong overlap for $\pi_S^N(c_S) := \mathbb{E}_{p(C_{\bar{S}}|C_S=c_S)}[\pi^N(c_S, C_{\bar{S}})]$ and $P(T = 1|C_S = c_S) = \mathbb{E}_{p(C_{\bar{S}}|C_S=c_S)}[\pi^*(c_S, C_{\bar{S}})]$ (by taking the expectation w.r.t. $p(C_{\bar{S}}|c_S)$) and also Jensen's inequality yields

$$
\forall c, N, |\mathbb{E}_{p(C_{\bar{S}}|C_S=c_S)}[\pi^N(c_S, C_{\bar{S}})] - P(T = 1|C_S = c_S)| \leq e_N^{\text{propensity}},
$$

which further gives $\forall t, c, N, |w_S^N(c,t) - w_S^*(c,t)| \leq e_N^{\text{propensity}}$.

So for any S s.t. $T \notin S$, c, N,

$$
\begin{aligned}
|w_S^N(c,0)| &= |-\pi_S^N(c)| = \pi_S^N(c) \geq \epsilon \\
|w_S^N(c,1)| &= |1 - \pi_S^N(c)| = 1 - \pi_S^N(c) \geq \epsilon
\end{aligned}
$$

Thereby, for any S s.t. $T \notin S$, t, c, N,

$$\frac{1}{|w_S^N(c,t)|} \leq \frac{1}{\epsilon}$$

and, similarly,

$$\frac{1}{|w_S^*(c,t)|} \leq \frac{1}{\epsilon}.$$

We also show that $|\phi_{T,S}^{f^*}(c,t)| \leq 2||f^*||_\infty$: indeed,

$$\begin{aligned}
|\phi_{T,S}^{f^*}(c,t)| &= |w_S^*(c,t)||\mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[f^*(c_S,C_{\bar{S}},1)] - \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[f^*(c_S,C_{\bar{S}},0)]| \\
&\leq \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[|f^*(c_S,C_{\bar{S}},1)|] + \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[|f^*(c_S,C_{\bar{S}},0)|] \\
&\quad \text{from } |w_S^*(c,t)| \leq 1 \text{ and the triangle inequality and Jensen's inequality} \\
&\leq \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=1)}[||f^*||_\infty] + \mathbb{E}_{p(C_{\bar{S}}|C_S=c_s,T=0)}[||f^*||_\infty] \\
&= 2||f^*||_\infty
\end{aligned}$$

In the end, we have

$$\begin{aligned}
&|\hat{\Psi}_{T\to Y|C_S}^{N,\hat{f}_N}(c) - \Psi_{T\to Y|C_S}^{f^*}(c)| \\
&= |\frac{\hat{\phi}_{T,S}^{N,\hat{f}_N}(c,t)}{w_S^N(c,t)} - \frac{\phi_{T,S}^{f^*}(c,t)}{w_S^*(c,t)}| \\
&= |\frac{\hat{\phi}_{T,S}^{N,\hat{f}_N}(c,t) - \phi_{T,S}^{f^*}(c,t)}{w_S^N(c,t)} + \phi_{T,S}^{f^*}(c,t)(\frac{1}{w_S^N(c,t)} - \frac{1}{w_S^*(c,t)})| \\
&= |\frac{\hat{\phi}_{T,S}^{N,\hat{f}_N}(c,t) - \phi_{T,S}^{f^*}(c,t)}{w_S^N(c,t)} + \phi_{T,S}^{f^*}(c,t)\frac{w_S^*(c,t) - w_S^N(c,t)}{w_S^*(c,t)w_S^N(c,t)}| \\
&\leq \frac{|\hat{\phi}_{T,S}^{N,\hat{f}_N}(c,t) - \phi_{T,S}^{f^*}(c,t)|}{|w_S^N(c,t)|} + |\phi_{T,S}^{f^*}(c,t)|\frac{|w_S^*(c,t) - w_S^N(c,t)|}{|w_S^*(c,t)||w_S^N(c,t)|} \quad \text{from Jensen's inequality} \\
&\leq \frac{2e_N^{\text{Shap}}}{\epsilon} + 2||f^*||_\infty \cdot \frac{e_N^{\text{propensity}}}{\epsilon^2}
\end{aligned}$$

which proves bound 4. Bound 5 is proven similarly to bound 3 above.

J.5. Proof of Lemma 6.2.

If unconfoundess w.r.t. C_1, C_2 holds then

$$\mathbb{E}[\Psi_{T\to Y|C_1,C_2}^{f^*}(C_1,C_2)] = \mathbb{E}[\mathbb{E}[Y|T=1,C_1,C_2] - \mathbb{E}[Y|T=0,C_1,C_2]] = \text{ATE}.$$

If unconfoundess w.r.t. C_1 also holds then

$$\mathbb{E}[\Psi_{T\to Y|C_1}^{f^*}(C_1)] = \mathbb{E}[\mathbb{E}[Y|T=1,C_1] - \mathbb{E}[Y|T=0,C_1]] = \text{ATE}$$

In the end,

$$\mathbb{E}[\Psi_{T\leftarrow C_2\to Y}^{f^*}(C_1,C_2)] = \mathbb{E}[\Psi_{T\to Y|C_1,C_2}^{f^*}(C_1,C_2) - \Psi_{T\to Y|C_1}^{f^*}(C_1)] = \text{ATE} - \text{ATE} = 0.$$

J.6. Proof of Property 6.1

Let c and m_1 be values of C and M_1, respectively. We note that

$$\mathbb{E}[\Psi_{T\to M_2\to Y}^{f}(c,m_1,M_2)|C=c]$$

$$
\begin{aligned}
&= \mathbb{E}[\Psi^f_{T\to M_2\to Y}(c, m_1, M_2)|C = c, M_1 = m_1] \text{ as } M_1 \perp\!\!\!\perp M_2|C \\
&= 0 \qquad \text{from Property 3.2 as } M_1 \perp\!\!\!\perp T|M_2, C
\end{aligned}
$$

which completes the proof.

J.7. Proof of Property D.1

If a latent variable generates all pre-treatment covariates of T, then we can factorise the distribution of $(T, C_S, C_{\bar{S}}, f\left(C_{\bar{S}}, c_S, t\right))$ in the ADMG with those variables as nodes and edges $C_S \leftrightarrow C_{\bar{S}} \to f\left(C_{\bar{S}}, c_S, t\right)$ and $C_{\bar{S}} \to T \leftarrow C_S$. We aim to apply rule 3 of Pearl's do-calculus. If we remove edges pointing into C_S and T, we obtain an ADMG with only the edge $C_{\bar{S}} \to f\left(C_{\bar{S}}, c_S, t\right)$. In this graph T and $f\left(C_{\bar{S}}, c_S, t\right)$ are m-separated by $C_{\bar{S}}$. Therefore, rule of 3 of do-calculus applies and we can remove the $T = t$ term in the do-operator of the left-hand term, yielding the right-hand term. Hence the indirect effect is zero.

J.8. Proof of Corollary E.1

We note that

$$
\begin{aligned}
\mathbb{E}[\Psi^f_{T\leftarrow C_2\to Y}(C_1, C_2)] &= \mathbb{E}[\mathbb{E}[\Psi^f_{T\leftarrow C_2\to Y}(C_1, C_2)|C_1]] \text{ from the tower property} \\
&= \mathbb{E}[0] \text{ from Property 3.2 as } C_2 \perp\!\!\!\perp T|C_1 \\
&= 0
\end{aligned}
$$

J.9. Proof of Lemma F.1.

Let m be a value of M. If $\mathcal{H}(C)$ holds then for any $t = 0, 1$

$$
\begin{aligned}
\mathbb{E}[Y(t, m)] &= \mathbb{E}[\mathbb{E}[Y(t, m)|C]] \\
&= \mathbb{E}[\mathbb{E}[Y(t, m)|C, t]] \text{ from } Y(t, m) \perp\!\!\!\perp T|C \\
&= \mathbb{E}[\mathbb{E}[Y(t, m)|C, t, m]] \text{ from } Y(t, m) \perp\!\!\!\perp M|T, C \\
&= \mathbb{E}[\mathbb{E}[Y|C, t, m]] \text{ from consistency.}
\end{aligned}
$$

so

$$
\begin{aligned}
\mathbb{E}[\Psi^{f^*}_{T\to Y|C,M}(C, m)] &= \mathbb{E}[\mathbb{E}[Y|C, t = 1, m]] - \mathbb{E}[\mathbb{E}[Y|C, t = 0, m]] \\
&= \mathbb{E}[Y(1, m)] - \mathbb{E}[Y(0, m)] \text{ from the above} \\
&= \text{CDE}(m).
\end{aligned}
$$

So if both $\mathcal{H}(C_1, C_2)$ and $\mathcal{H}(C_1)$ hold then

$$
\begin{aligned}
\mathbb{E}[\Psi^{f^*}_{C_2}(C_1, C_2, m)] &= \mathbb{E}[\Psi^{f^*}_{T\to Y|C_1,C_2,M}(C_1, C_2, m) - \Psi^{f^*}_{T\to Y|C_1,M}(C_1, m)] \\
&= \text{CDE}(m) - \text{CDE}(m) \\
&= 0.
\end{aligned}
$$

J.10. Proof of Corollary F.2

First, let's note that $C_2 \perp\!\!\!\perp T, M|C_1$ implies $C_2 \perp\!\!\!\perp M|C_1$ and $C_2 \perp\!\!\!\perp T|C_1, M$. Let m be a value of M. We note that

$$
\mathbb{E}[\Psi^f_{C_2}(C_1, C_2)] = \mathbb{E}[\mathbb{E}[\Psi^f_{C_2}(C_1, C_2)|C_1]] \text{ from the tower property}
$$

$$
\begin{aligned}
&= \mathbb{E}[\mathbb{E}[\Psi^f_{C_2}(C_1, C_2)|C_1, M]] \text{ from the property } C_2 \perp\!\!\!\perp M|C_1 \\
&= \mathbb{E}[0] \text{ from Property 3.2 as } C_2 \perp\!\!\!\perp T|C_1, M \\
&= 0
\end{aligned}
$$

which completes the proof.

J.11. Proof of Property F.1

It suffices to show that $\mathbb{E}[\Psi^f_{T\to Y|M}(M)] = \Psi^f_{T\to Y|\emptyset}$, which is true as

$$
\begin{aligned}
&\mathbb{E}[\Psi^f_{T\to Y|M}(M)] \\
&= \int_m (v_f(\{M, T\}, m, t = 1) - v_f(\{M, T\}, m, t = 0))\mathrm{d}p(m) \\
&= \int_m (\mathbb{E}[f(C, m, t = 1)|m, t = 1] - \mathbb{E}[f(C, m, t = 0)|m, t = 0)\mathrm{d}p(m) \\
&= \int_m \mathbb{E}[f(C, m, t = 1)|m, t = 1]\mathrm{d}p(m) - \int_m \mathbb{E}[f(C, m, t = 0)|m, t = 0]\mathrm{d}p(m) \\
&= \int_m \mathbb{E}[f(C, m, t = 1)|m, t = 1]\mathrm{d}p(m|t = 1) - \mathbb{E}[f(C, m, t = 1)|m, t = 0]\mathrm{d}p(m|t = 0) \quad \text{as } M \perp\!\!\!\perp T \\
&= \mathbb{E}[f(C, M, t = 1)|t = 1] - \mathbb{E}[f(C, M, t = 1)|t = 0] \\
&= v_f(T, t = 1) - v_f(T, t = 0) \\
&= \Psi^f_{T\to Y|\emptyset}
\end{aligned}
$$

which completes the proof.

References

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Bickel, P. J., Hammel, E. A., and O'Connell, J. W. (1975). Sex bias in graduate admissions: Data from Berkeley. *Science*, 187(4175):398–404.

Foster, J. C., Taylor, J. M., and Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880.

Frye, C., Rowat, C., and Feige, I. (2019). Asymmetric Shapley values: incorporating causal knowledge into model-agnostic explainability. *arXiv preprint arXiv:1910.06358*.

Heskes, T., Sijben, E., Bucur, I. G., and Claassen, T. (2020). Causal Shapley values: Exploiting causal knowledge to explain individual predictions of complex models. *arXiv preprint arXiv:2011.01625*.

Holmes, C. C. and Watson, J. A. (2018). Machine learning for randomised controlled trials: identifying treatment effect heterogeneity with strict control of type I error. *bioRxiv*, page 330795.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Imbens, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132.

Janzing, D., Minorics, L., and Blöbaum, P. (2020). Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pages 2907–2916. PMLR.

Jesson, A., Douglas, A., Manshausen, P., Meinshausen, N., Stier, P., Gal, Y., and Shalit, U. (2022). Scalable sensitivity and uncertainty analysis for causal-effect estimates of continuous-valued interventions. *arXiv preprint arXiv:2204.10022*.

Karimi, A.-H., Schölkopf, B., and Valera, I. (2021). Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 353–362.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2018). Consistent individualized feature attribution for tree ensembles. *ArXiv*, abs/1802.03888.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

Merrick, L. and Taly, A. (2020). The explanation game: Explaining machine learning models using shapley values. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 17–38. Springer.

Nilsson, A., Bonander, C., Strömberg, U., and Björk, J. (2019). Assessing heterogeneous effects and their determinants via estimation of potential outcomes. *European journal of epidemiology*, 34(9):823–835.

Pan, W., Cui, S., Bian, J., Zhang, C., and Wang, F. (2021). Explaining algorithmic fairness through fairness-aware causal path decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1287–1297.

Petersen, M. L. and van der Laan, M. J. (2014). Causal models and learning from data: integrating causal modeling and statistical estimation. *Epidemiology (Cambridge, Mass.)*, 25(3):418.

Rosenbaum, P. R. (2005). Sensitivity analysis in observational studies. *Encyclopedia of statistics in behavioral science*.

Rosenbaum, P. R. and Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Sani, N., Malinsky, D., and Shpitser, I. (2020). Explaining the behavior of black-box prediction algorithms with causal learning. *arXiv preprint arXiv:2006.02482*.

Singal, R., Michailidis, G., and Ng, H. (2021). Flow-based attribution in graphical models: A recursive Shapley approach. In *International Conference on Machine Learning*, pages 9733–9743. PMLR.

Song, Y. and Chi, G. Y. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in medicine*, 26(19):3535–3549.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274.

VanderWeele, T. J. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, 2:457–468.

Veitch, V. and Zaveri, A. (2020). Sense and sensitivity analysis: Simple post-hoc analysis of bias due to unobserved confounding. *Advances in Neural Information Processing Systems*, 33:10999–11009.

Wang, J., Wiens, J., and Lundberg, S. (2021). Shapley flow: A graph-based approach to interpreting model predictions. In *International Conference on Artificial Intelligence and Statistics*, pages 721–729. PMLR.

Wang, T. and Rudin, C. (2017). Causal rule sets for identifying subgroups with enhanced treatment effect. *arXiv preprint arXiv:1710.05426*.

Wu, H., Tan, S., Li, W., Garrard, M., Obeng, A., Dimmery, D., Singh, S., Wang, H., Jiang, D., and Bakshy, E. (2021). Distilling heterogeneity: From explanations of heterogeneous treatment effect models to interpretable policies. *arXiv preprint arXiv:2111.03267*.

Zhao, Q. and Hastie, T. (2021). Causal interpretations of black-box models. *Journal of Business & Economic Statistics*, 39(1):272–281.