Theoretical Evaluation of Asymmetric Shapley Values for Root-Cause Analysis

Domokos Kelen, Mihàly Petreczky, Péter Kersch, Andràs Benczúr

Theoretical Evaluation of Asymmetric Shapley Values for Root-Cause Analysis

Domokos M. Kelen
*HUN-REN SZTAKI**,
Budapest, Hungary
kdomokos@sztaki.hu

Mihály Petreczky
CNRS, University Lille
Lille, France
mpetrec@gmail.com

Péter Kersch
Ericsson Hungary
Budapest, Hungary
peter.kersch@ericsson.com

András A. Benczúr
*HUN-REN SZTAKI**,
Budapest, Hungary
benczur@sztaki.hu

***Abstract*—In this work, we examine Asymmetric Shapley Values (ASV), a variant of the popular SHAP additive local explanation method. ASV proposes a way to improve model explanations incorporating known causal relations between variables, and is also considered as a way to test for unfair discrimination in model predictions. Unexplored in previous literature, relaxing symmetry in Shapley values can have counter-intuitive consequences for model explanation. To better understand the method, we first show how local contributions correspond to global contributions of variance reduction. Using variance, we demonstrate multiple cases where ASV yields counter-intuitive attributions, arguably producing incorrect results for root-cause analysis. Second, we identify generalized additive models (GAM) as a restricted class for which ASV exhibits desirable properties. We support our arguments by proving multiple theoretical results about the method. Finally, we demonstrate the use of asymmetric attributions on multiple real-world datasets, comparing the results with and without restricted model families using gradient boosting and deep learning models.**

***Index Terms*—explainability, SHAP, causality**

I. Introduction

Removal-based feature attribution methods have gained huge popularity in recent years. To gain insight into machine learning models, they calculate expectations of model predictions with and without revealing certain feature values and examine the change in expectation. Methods like SHAP [1] yield easy-to-interpret explanations for model predictions by attributing parts of the prediction to individual features.

In some cases, the prediction can be made from either of multiple predictors with overlapping information content, such as when multiple variables in a causal chain are present in the dataset. The Asymmetric Shapley Values (ASV) [2] framework promises to solve this problem by relaxing the symmetry property in the original Shapley formula. However, as we show in this paper, relaxing symmetry can yield unexpected results in case of complex interactions between variables.

A method to handle overlapping information content is essential in the case of root cause analysis, where the explanation method is used for identifying the underlying issues for a certain prediction value, with multiple complex causal relationships among different levels of observations. For example, in radio telecommunications, base station configuration affects load and interference, which then affect the radio channel quality and eventually the user experience. We can try to assert the cause of user experience degradation through local explanations using SHAP; however, the degradation can most likely be predicted from either of multiple variables in the causal chain. In this case, we prefer to prioritize the root (or *distal*) causes over more immediate ones in the explanation, and ASV provides us with a convenient way to do so.

In other cases, such as detecting unfair discrimination, the same framework can be used to instead prefer the immediate (or *proximate*) causes. In social sciences, family attributes affect education, which determines income and poverty, which eventually predict crime rate. One might prefer the explanations to attribute shared contribution to resolving variables that mediate the effects of sensitive attributes [2].

The previous two use cases primarily consider the effect of shared contributions, where multiple predictors provide redundant information about a target variable through correlation or other nonlinear means. However, interacting variables can also result in additional information about the target variable that cannot be inferred from either of the predictor variables alone but is revealed by the variables when observed together. Such interactions, which we call *complex interactions*, can lead to arguably incorrect results in the case of distal attribution schemes and disproportionate contributions in the case of proximate attribution schemes. It is important to consider the consequences of different kinds of interaction in various use cases, as the attributions produced by the method can be misleading without a proper understanding of the different ways that variables can interact.

In this work, we conduct a theoretical analysis of ASV, focusing on regression, and formalizing the effects of relaxing symmetry in terms of variance reduction of the target variable. We seek answers to the following research questions:

RQ1 Can we reason about the attributions of ASV formally?
RQ2 Is ASV a reliable method for root-cause analysis?
RQ3 Can we avoid undesired behavior by using GAMs?

In our analysis, we use the framework of variance reduction to show how ASV can yield counter-intuitive results in the presence of nonlinearity or even simple non-additive variable interactions. As a solution, we propose applying ASV in conjunction with generalized additive models (GAMs). In

Supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory Program.
*Hungarian Research Network, Institute for Computer Science and Control

Theorem B, our main result proves a sufficient condition for correct behavior when ASV is applied to GAMs.

Our paper is structured as follows. In Section II, we summarize related literature. In Section III, we recall the definitions of SHAP and ASV. In Section IV, we describe how the behavior of local explanations can be studied through the framework of variance reduction. In Section V, we characterize the behavior of ASV, and present theoretical examples where ASV yields counter-intuitive results. In Section VI, we show that many of the listed counter-intuitive behaviors vanish when restricting the usage of ASV to GAMs. In Section VII, we discuss generalization to classification. Finally, in Section IX, we demonstrate our findings on real-world datasets.

II. Related literature

SHAP, introduced in [1], has sparked a surge of research in recent years into related explainability methods. In this paper, we study Asymmetric Shapley Values [2], which provides a way to account for known causal relationships in the dataset when calculating contributions. We will enumerate a number of other works that build upon SHAP to account for causality in attributions; however, we are not aware of works on formally assessing the suitability of ASV for root-cause analysis.

Results on measuring causal contributions [3]–[6] enforce restricting explanation to causal effects by relying on Pearl's do-calculus [7]. In contrast, our pathological examples (Section V) stem from the combination of asymmetry and non-linearity, which are orthogonal to the question of direct and indirect causal effects. Do-calculus-based methods combine well with ours by detecting confounders while our results detect complex interactions. Shapley flow [4] extends the do-calculus idea by assigning contributions to causal edges instead of variables. Causal Shapley Values [3] decompose the causal effects into direct and indirect effects, which can then be combined both with symmetric SHAP and ASV [3, Fig. 1], hence can leverage on our results. Finally, the main contribution in [5] consists of criteria to efficiently infer do-calculus values from fixed measurement data, which is non-trivial in our use cases with no intervention opportunity when training the model.

In a completely different approach to model explanation, in [8] the use of generalized additive models (GAM) [9] is proposed to combine the expressivity of a general model with the inherent intelligibility of GAMs but without providing local post-hoc explanations. The relationship of SHAP and GAMs are studied in [10], which introduces n-Shapley Values to explain n-wise symmetric interactions. As a step beyond, in Section VI we formalize the behavior ASV when applied in conjunction with GAMs vs. arbitrary models.

We study ASV by explaining the variance reduction of the target variable in the SHAP framework. Our variance-reduction-based approach is equivalent in expectation to SAGE [11], [12] with the MSE loss function. In general, analysis of variance (ANOVA) and variance-based sensitivity analysis are popular approaches [13]–[16], which we use here to study ASV. Note that our ASV variance reduction explanation does not modify the validity of the other SHAP axioms [1] beyond symmetry.

III. Background

A. Shapley Additive Explanations (SHAP)

In [1], SHAP is defined as a way to explain a specific prediction of a machine learning model by assigning contribution values to different input features of the model. More specifically, given model f, feature set $\{X_s \mid s \in \mathbb{S}\}$, and a specific point x from the dataset, the contribution ϕ_j of feature $j \in \mathbb{S}$ can be defined as

$$\phi_j(f,x) = \sum_{S \subseteq \mathbb{S}\setminus\{j\}} \frac{|S|!(|\mathbb{S}|-|S|-1)!}{|\mathbb{S}|!}\, \varphi_j(f,x,S), \quad (1)$$

$$\text{where } \varphi_j(f,x,S) = v_{f(x)}(S\cup\{j\}) - v_{f(x)}(S). \quad (2)$$

Here $v_{f(x)}(S)$ is called the *value function* and represents the function of evaluating the model f at point x while only using features from the coalition S. The evaluation is usually done by taking the expectation of $f(x)$ conditional on the in-coalition feature values S:

$$v_{f(x)}(S) = E[f(X)\,|\,X_s = x_s, s \in S]. \quad (3)$$

With $\phi_0 = E(f(X))$, the feature contributions sum up to the prediction value:

$$\sum_{j=0}^{|\mathbb{S}|} \phi_j(f,x) = f(x). \quad (4)$$

We note that two commonly used variants of SHAP are conditional and marginal SHAP [17]. Similar to [2], we focus on *conditional* SHAP, where the dependencies between features are accounted for in the expected value in (3).

B. Asymmetric Shapley Values

When the prediction power of two variables together is not equal to the sum of their individual prediction powers, then these variables are said to interact. The original weighting scheme of Shapley values, described in (1), guarantees such interactions to be distributed equally between variables. In [2], ASV is introduced as a way to prefer certain variables over others by assigning interactions asymmetrically. This can be useful for example in the presence of causal relationships, where one might prefer assigning shared contribution to either root causes or immediate causes. Formally, Asymmetric Shapley Values are defined as

$$\phi_j^{\omega}(f(x)) = \sum_{\pi\in\Pi} \omega(\pi)\, \varphi_j(f,x,R_j), \quad (5)$$

where Π is the set of all possible permutations of the model features, $R_j = \{i : \pi(i) < \pi(j)\}$, ω is a weighting over permutations, and $\varphi_j(f,x,R_i)$ is defined in (2). The definition allows any weighting function, for example, the distal (or root cause) function [2]

$$\omega_{distal}(\pi) \propto \begin{cases} 1 & \text{if } \pi(i) < \pi(j) \text{ for all } i,j \text{ where } i \\ & \text{is a causal ancestor of } j \\ 0 & \text{otherwise,} \end{cases} \quad (6)$$

which essentially means that causal ancestors are always revealed before variables affected by them. The intuition given behind this approach is exemplified by assuming two identical X_1, X_2, however with X_1 being a causal ancestor of X_2. In this case, one might want to attribute all importance to X_1 instead of sharing it evenly.

The difference between ASV and SHAP is how variable interactions are handled. Two features i and j are said to interact when

$$v_{f(x)}(\{i\}) + v_{f(x)}(\{j\}) \neq v_{f(x)}(\{i,j\}), \tag{7}$$

i.e. the effect of revealing the value of both at the same time to the model can not be additively predicted from revealing them separately. This can be the case, e.g., when they contain redundant information. In such cases, the difference between the left and right-hand side of (7) needs to be added to $\phi_i(f,x)$ and $\phi_j(f,x)$ such that (4) holds. SHAP distributes the interaction value uniformly between the interacting variables, while ASV assigns it to the variable revealed last in each specific permutation π of (5).

IV. Variance Reduction in the SHAP framework

The correctness of explanation methods is hard to quantify or benchmark, as there is rarely any ground truth to compare against. In the case of SHAP and ASV, even general behavior is hard to reason about, as these methods explain model predictions locally, i.e., each $f(x)$ prediction separately in the input point x. A practical approach is to study them based on aggregate behavior, e.g., in [2], the authors observe the behavior of ASV through the expectation of $\phi_j(f,x)$ over $x \sim X$. However, it is important to pay attention to how well the average reflects local behavior, as contribution values can have both positive and negative values.

In our case, the primary goal is to evaluate ASV for the regression task. Optimizing for mean squared error results in the model $f(x)$ approximating the conditional expectation of the target T, i.e., $f(X) \approx E[T \,|\, X]$, also implying

$$E[f(X) \,|\, X_S] \approx E[E[T \,|\, X] \,|\, X_S] = E[T \,|\, X_S] \tag{8}$$

when using conditional SHAP, where $X_S = \{X_s \,|\, s \in S\}$.

As a result, however, using the average of local contribution values is pointless, as it is guaranteed to approximate zero. To see this, notice first that $\phi_0 = E[f(X)]$ as in (4), and second that the expectation $E[\phi_j(f,X)] = 0$ for all $j \neq 0$, since for any S, f, and x, due to the law of total expectation,

$$E[v_{f(X)}(S)] = E[E[f(X) \,|\, X_S]] = E[f(X)], \tag{9}$$

implying $E[\varphi_j(f,x,S)] = 0$ in turn for any j, S in (2).

As a solution, some authors [18], [19] as well as the SHAP Python package consider features with large absolute Shapley values important. The average of the absolute contribution values $|\phi|$ is also used as feature importance in [20], [21]. However, the average absolute contribution value does not have a theoretic foundation, which again makes it hard to formally reason about the behavior of the method.

Instead, to answer **RQ1**, we propose observing the changes in the variance of the target variable T with different feature sets revealed to the model. We first show that this value is strongly tied to the local behavior of the model. Let us define a new value function w to replace v of (2):

$$w_{f(x),t}(S) \stackrel{def}{=} (t - v_{f(x)}(S))^2. \tag{10}$$

On the one hand, the behavior of w directly corresponds to the behavior of v, as it is just a simple transformation of the latter. On the other hand, averaging $w_{f(x),t}(S)$ over the dataset for a given S approximates the variance of T conditioned on features from S, called the residual sum of squares:

$$E[w_{f(X),T}(S)] = E[(T - v_{f(X)}(S))^2] \tag{11}$$

$$= E\big[(T - E(T \,|\, X_S))^2\big] \tag{12}$$

where $X_S = \{X_s\}_{s \in S}$. The value in (12) is also called conditional variance [22]. Notice that

$$\phi_0 = E[w_{f(X),T}(\emptyset)] = E\big[(T - E(T))^2\big] = \sigma^2(T), \tag{13}$$

i.e., the average contribution of the empty set ϕ_0 becomes the variance itself, and that

$$R^2 \cdot \sigma^2(T) = \sigma^2(T) - \sigma^2(T - E(T \,|\, X_S)) = \tag{14}$$

$$= E[w_{f(X),T}(\emptyset)] - E[w_{f(X),T}(S)], \tag{15}$$

where R^2 is the *coefficient of determination* from statistics.

Since (1) and (2) are linear transformations of the value function, all of the above means that when SHAP is used to explain the value function w of (10) locally, then the averages of contributions explain the variance reduction of the target variable globally in the SHAP framework. The contribution of the empty set, i.e., ϕ_0, is equal to the unconditional variance, while the addition of predictor features lowers this variance, such that in the end the sum of the contributions equals $E\big[(T - E(T \,|\, X_S))^2\big]$. The resulting contributions are equivalent in expectation to the global explanations provided by SAGE [11] for the squared error loss function.

In what follows, we denote R^2 unnormalized as

$$L_{T}(X) \stackrel{def}{=} \sigma^2(T) - \sigma^2(T - E[T \,|\, X]), \tag{16}$$

also allowing multiple variables in place of X, denoted as, e.g., $L_{T}(X,Y)$. The value of $L_{T}(X)$ can be interpreted as *the predictive power of* X when no other variables are present. Note that a positive amount of variance reduction is realized as a *negative contribution* of the variable in the SHAP framework: when averaged, summing ϕ_0 to the contribution values equals the reduced variance of the target.

We also propose a definition to characterize the relation of the variance reduction of two attributes X and Y separately as opposed to the pair X, Y together, similar to the classic Shapley interaction value (7). We define

$$W_{T}(X;Y) \stackrel{def}{=} L_{T}(X,Y) - L_{T}(X) - L_{T}(Y), \tag{17}$$

the *interaction of variance reduction*. The value $L_{T}(X)$ is guaranteed to be positive, while $W_{T}(X;Y)$ can be either positive or negative.

V. Analysis of ASV through variance reduction

Using the notation introduced in Section IV, we can now answer **RQ1**. We can express the contributions assigned by ASV in terms of variance reduction as follows.

Proposition A. *For a given permutation π, using w of* (10) *in place of v in* (2)*, and $R_j = \{i : \pi(i) < \pi(j)\}$,*

$$-E[\varphi_j(f, X, R_j)] = L_T(X_{\pi(j)}) + W_T(X_{\pi(j)}; R_j). \quad (18)$$

Proof. From the definitions in (2), (10), (16), (17) and observing (14) and (15), the proposition immediately follows. □

To illustrate the meaning of Proposition A, recall the distal weighting function ω of (6). Along with (5), it filters out any π in which the causal ordering is flipped for any two variables. For the remaining π, variable X_j gets contribution negatively proportional to the variance of T that it explains, plus the interaction value W between it and its predecessors. The total contribution of X_j is then an average over every such π.

Next, we show three examples where ASV can be argued to produce counter-intuitive results. ASV reveals variables to the model in the order specified by the permutation π in order to prefer the variables that are revealed earlier in the explanation. This approach seems intuitive, as variables earlier in the permutation get their full contribution, while later variables are assigned leftover prediction power. However, in some cases, the intuition fails. In the three examples below, we reveal two variables in the order (X_1, X_2), and show that ASV assigns disproportionate contributions to X_2.

Example A (Pairwise independence)**.** Let X_1 and X_2 be pairwise independent of T, however with $T = f(X_1, X_2)$ for some f. With the permutation (X_1, X_2), ASV assigns all contribution to X_2 for predicting T.

A pair of variables can be pairwise independent, but at the same time not mutually independent together with the target variable.

Example B (Nonlinearity)**.** For $X_1, X_2 \sim N(0, 1)$ independent Gaussians, let $T = (2X_1 + 2X_2)^2$. Both variables contribute the same amount, yet with the permutation (X_1, X_2), ASV assigns three times the importance of X_1 to X_2 when measured using variance reduction. For an exact calculation of contributions, see Appendix XI-C.

Example C (Non-additive effects)**.** Let X_1 and X_2 be as in Example B and $T = X_1 \cdot X_2$. Here $E[T \mid X_1] = 0$, therefore with the permutation (X_1, X_2), ASV assigns all contribution to X_2 for predicting T, even though both variables contributed equally to T.

In all three of these examples, the main source of counter-intuitive behavior is that X_2 gets assigned more contribution *because* it is introduced after X_1, more contribution than if it was the only predictor used, which goes directly against the intuition of the ordering of the predictors in π. In the examples, the effects are not the result of causal relationships between X_1 and X_2, rather they are artifacts of relaxing symmetry itself. However, real-world relationships can contain a combination of multiple effects, including causal and nonlinear ones. As an arbitrary causal example, we could have, e.g., $X_2 = X_1 + 2Y$ with $Y \sim N(0, 1)$ and $T = (X_1 + X_2)^2$, resulting in the same incorrect contributions as in Example B.

Thus, answering **RQ2**, we conclude that no, ASV is not always a reliable way of assigning contributions in root-cause analysis, as the effects demonstrated by Examples A, B, and C can result in proximate instead of root causes getting a large share of the contributions. The counter-intuitive behavior is due to the fact that the inequality

$$W_T(X, Y) = L_T(X, Y) - L_T(X) - L_T(Y) \overset{?}{\leq} 0 \quad (19)$$

does not hold in general: the prediction power of using two variables at the same time can be greater than the sum of their prediction powers separately, since the interaction between variables can be quite complex. Thus we can have two kinds of interaction between variables: interactions that result in the combined prediction power being less, and interactions that result in the combined prediction power being more than the sum of the individual prediction powers of the variables. We call the former kind of interaction *redundant information* and the latter *complex interaction*. The actual measured W_T value is the sum of interactions of different kinds.

VI. ASV on Generalized Additive Models

From an interpretability standpoint, complex interactions are undesirable: they cannot be attributed cleanly to any of the involved variables. Rather, complex interactions are inherently the property of the variables being used together. However, attributing redundant information in causal settings is straightforward both in theory and in practice, as described and demonstrated by ASV. In this section, we propose using generalized additive models [9] (GAMs) in an attempt to exclude undesirable complex interactions. We study the behavior of ASV when applied to GAMs, and prove that ASV exhibits certain desirable properties when applied on GAMs.

GAMs are a special class of prediction model, which can be written as a sum of smooth functions of different features:

$$g(E[T \mid X]) \approx f_0 + f_1(X_1) + ... + f_n(X_n), \quad (20)$$

where g is called the link function. Observe that f_0 is nonessential, as its effects can be assimilated within g or f_i. The explainability properties of GAMs have been studied before [8], [23], as well as various ways of approximating them using machine learning models. While these restricted models are strictly less powerful than their unrestricted counterparts, sacrificing some prediction power for interpretability can be reasonable in many domains and applications where the trustworthiness of the results is critical [8].

Note that GAMs themselves are additive, for which marginal SHAP [1] guarantees that contributions reflect the outputs of individual f_i functions. The same is, however, *not* true for conditional SHAP [17], where (8) remains true. Illustrating this point, assume two identical predictor variables. As

also observed in [10], a GAM is free to choose either of them for predicting T. Therefore, hiding one of the variables could result in the model losing information when using marginal SHAP. In contrast, in conditional SHAP, the expectation is forced to account for either of them being present through their joint distribution.

To reason about ASV and GAMs formally, we propose the following definitions. We define the *restricted* conditional expectation of T given $X_1, \dots, X_n$ as

$$E^r[T \mid X_1, \dots, X_n] \stackrel{def}{=} \operatorname*{argmin}_{\{f_i\}} \sigma^2\left(T - \sum_i f_i(X_i)\right), \quad (21)$$

where f_i are measurable and $E[f_i(X_i)^2] < \infty$. Equation (21) gives the minimum variance predictor of T that can be written as the sum of functions of X_i, i.e., as a GAM. We use the notation E^r as a contrast to regular conditional expectation, which is equivalent to using a single measurable function over all X_i to minimize variance [24]. The function E^r can be interpreted as the best possible GAM while using the identity link function. Notice that with only one predictor variable, the definitions of functions E and E^r are equivalent. For a discussion of the existence of E^r, see Appendix XI-A.

Further, from the definition of (21), we can derive matching definitions for restricted versions of L_T and W_T:

$$L^r_T(X) \stackrel{def}{=} \sigma^2(T) - \sigma^2(T - E^r[T \mid X]) \quad (22)$$

$$W^r_T(X;Y) \stackrel{def}{=} L^r_T(X,Y) - L^r_T(X) - L^r_T(Y). \quad (23)$$

Next, we prove that ASV exhibits a number of desirable properties when applied in combination with GAMs. Our first theorem refers back to Example A.

Theorem A (Additivity of contributions)**.** *If X, Y are independent, then $L^r_T(X) + L^r_T(Y) = L^r_T(X,Y)$.*

See Appendix XI-E for proof. Theorem A proves that under the GAM restriction, the predictive power of two independent variables together always equals the sum of their individual predictive powers. It also resolves Example A, as X_1 and X_2 must get the same contribution regardless of order of inclusion. Since in the example both predictor variables are also pairwise independent of T, both variables get a contribution score of zero: the full prediction power of the model is the result of a complex interaction, which is excluded from E^r, as desired.

Observe that Theorem A also applies to Examples B and C. Specifically, in Example B, both X_1 and X_2 can be used to predict T only to a certain degree, with half of its variance remaining, however, their contributions this time are equal. See Appendix XI-D for the exact calculation. In Example C, $E^r[T \mid X_1, X_2] = 0$, therefore, similar to Example A, both variables get zero contribution.

Finally, we turn to the question raised by (19). With the following theorem, we prove that if the prediction function F can be decomposed as $F_X(X) + F_Y(Y)$ (i.e., is a GAM), then ASV works as intended in most cases, shifting the contribution values to prefer variables earlier in the permutation π.

Theorem B (Upper bound of W^r_T)**.**

$$W^r_T(X_1, \dots, X_n; Y_1, \dots, Y_m) \le -2cov(F_X(X), F_Y(Y))$$

assuming $\exists F_X, F_Y, cov(F_X(X), F_Y(Y)) < \infty$ such that

$$F_X(X) + F_Y(Y) = E^r[T \mid X_1, \dots, X_n, Y_1, \dots, Y_m]. \quad (24)$$

For a proof of the theorem, see Appendix XI-F.

Theorem B means that as long as the covariance of $F_X(X)$ and $F_Y(Y)$ is positive, the interaction term W is negative, i.e., (19) is essentially true. Unfortunately, the theorem comes with the caveat that when the covariance is negative, the interaction *can* be positive, as illustrated by the next example.

Example D (Rank deficiency)**.** Let $A, B \sim N(0,1)$ be independent joint Gaussian variables, and let

$$X_1 = (0.1A + B),\ X_2 = (0.1A - B), \text{ and } T = A. \quad (25)$$

Then $E[T \mid X_1, X_2] = 5X_1 + 5X_2 = T$.

Example D is similar to the ones listed in Section V in that the target T cannot be efficiently predicted from X_1 or X_2 alone, but it can be predicted from X_1 and X_2 together. However, Example D is special in the sense that it persists even when using predictors according to E^r, i.e., the relations are all additive. In fact, the example remains valid even if we restrict the model family to linear regression. However, the example is quite unnatural, as the problematic behavior occurs because both variables X_1 and X_2 contain the same dominant noise variable, which can only be canceled out when observing both together. Such problems can be avoided on a theoretical level by assuming that noise terms are independent, but in practice, this might not always hold.

Answering **RQ3**, the GAM restriction filters out a large class of undesired complex interactions including those of Examples A, B, and C; however certain hard-to-interpret interactions can happen even with the restriction. Fortunately, as we see in Section IX, such relationships are uncommon in practice.

VII. Extension to classification

As stated before, our primary goal is to study ASV for the regression task. Aside from mutually exclusive events, probabilities are inherently non-additive, implying the same for the classification task. Practical modeling approaches usually involve having the model approximate the *logits* or log-odds of the probabilities, i.e., $\log \frac{p}{1-p}$, and applying the logistic sigmoid, its inverse, to the model output. Thus explaining the probabilities themselves using ASV may result in having to deal with nonlinear effects similar to Example B. Similarly, using GAMs for modeling probabilities directly is unlikely to result in accurate models. At the same time, logits are often treated as additive, e.g., in logistic regression. Therefore, we propose observing the output of the model before applying the sigmoid function, i.e., explaining the logits, as also done in [8]. Using GAMs to predict logits is straightforward, with the link function g of (20) chosen as the log-odds function.

Generalizing our results to classification seems a nontrivial task at first glance, as theorems A and B observe the behavior of ASV through variance. Moreover, the target variable in classification is the ground-truth probabilities, which are unobservable in general. However, notice the fact that the only actual assumption made in our analysis is that the model approximates a conditional expectation in (8). Further, since $f(X) = E[f(X) \mid X]$ in general, the model function can always be interpreted as a conditional expectation. Thus our approach to generalizing our results is by explaining the variance of the output of the full model when predicted using subsets of features, i.e., we set $T = f(X)$.

Ultimately, we consider the non-additivity of probabilities a limitation inherent to classification when being explained using additive explanation methods such as SHAP and ASV. However, explaining logits instead is a natural solution whenever applicable.

VIII. Approximation in practice

In this Section, we describe how we approximate the proposed definitions in practice. Regression modeling tasks, when using the mean squared error objective function, end up approximating the conditional expectation of the target variable E, as described in Section IV. To approximate E^r instead, we need to train GAMs, which we can achieve by using restricted versions of model families.

Gradient-boosted decision tree (GBDT) models [25] train an ensemble of decision tree models, where the final prediction is the sum of the predictions made by the individual trees. The training is done iteratively, always approximating the residual error. To achieve the GAM restriction, each individual tree needs to use variables of a single feature or feature group X_i. This way, the end result can be written as a sum of sets of trees, each depending on the value of a single X_i.

The function E^r can be approximated using other model families as well. For example with neural networks, the restriction can be achieved by modifying the architecture of the network to represent a sum of predictors based on different features or feature groups [8]. We primarily use GBDT-based models because of the tabular nature of our datasets. We run experiments implemented by LightGBM [25], where the necessary restriction can be met using the *interaction_constraints* flag. We include further experiments using neural network models [8] in our source code repository.

To evaluate ASV with conditional expectations, we follow the approach also used in [26], which means training a separate model for each coalition S to predict $E[T \mid X_S]$, and using these models as an approximation of $v_{f(x)}(S)$ for a given x with different feature subsets. In some experiments below, we also report contribution attributable to complex interactions, denoted by $\phi_{\mathcal{I}}$. The size $\phi_{\mathcal{I}}$ is determined by how much more powerful an unrestricted model is than its restricted counterpart. Further analysis could be applied to attribute parts of $\phi_{\mathcal{I}}$ to different features, however, this is out of scope for this work.

IX. Experiments

In this section, we demonstrate the effect of applying ASV to GAMs on real-world datasets. Our main goals in this section are thus to see:

- **Q1** how common complex interactions are in practice, both between variables and groups of variables;
- **Q2** whether Example D manifests in practice;
- **Q3** the extent to which the predictive power of models can be attributed to complex interactions, by comparing the performance of GAMs to that of unrestricted models.

We make the code of our experiments publicly available[1].

A. Datasets and methodology

We use the following datasets from the UCI Machine Learning Repository [27]. Other than unifying the target variable in the second dataset (PM2.5), no additional pre-processing is done on the UCI datasets.

Communities and Crime Unnormalized (CaCU). Combines socio-economic data from the '90 Census, law enforcement data from the 1990 Law Enforcement Management and Admin Stats survey, and crime data from the 1995 FBI UCR. Semantic groups are listed in Appendix XI-G.

PM2.5 Data of Five Chinese Cities (PM2.5). Contains air-quality data in Beijing, Shanghai, Guangzhou, Chengdu, and Shenyang with hour granularity, also including meteorological data for each city [28]. As a unified target, we average all measurement stations into a single variable in all cities.

Superconductivity. Contains features extracted from superconductors along with the critical temperature [29].

Productivity Prediction of Garment Employees (Garment). Includes attributes of the garment manufacturing process and the productivity of the employees [30], [31].

We also present evaluation on a large proprietary dataset:

Mobile Telecommunications (Telco). Real-world mobile telco. data, consisting of performance management (PM) data from radio access network cells with 15-minute granularity and configuration management (CM) data with daily granularity. We list feature groups in Appendix XI-G. Causality relations are displayed in Figure 5.

Relevant statistics of datasets are presented in Table I. Train/validation/test splits are done in ratios of $0.8/0.1/0.1$, and evaluations are presented as measured on the test set. GBDT experiments are conducted using LightGBM [25].

TABLE I
Datasets used.

Name	Features	Rows	Target variable
CaCU	124	2215	ViolentCrimesPerPop
PM2.5	14	262920	PM
Superconductivity	81	21263	critical_temp
Garment	14	1197	actual_productivity
Telco	78	19343000	downlink_throughput

[1] https://github.com/proto-n/shap-asv-icdm

B. Complex interaction example from the CaCU dataset

In this section, we demonstrate a real-world example of complex interactions on a concrete example from the CaCU dataset. Here the goal is to predict the relative number of violent crimes in communities (*ViolentCrimesPerPop*) from 124 features related to income, education, age, etc.

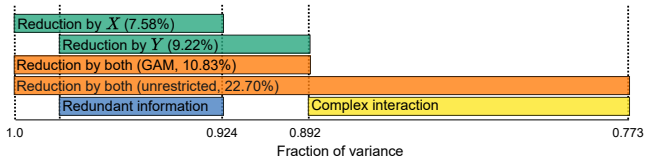

Fig. 1. Different kinds of interactions in the example of Section IX-B.

To start, we take two highly correlated variables: the percentage of the population that is 12-29 in age X (*agePct12t29*) and the percentage of males who have never married Y (*MalePctNevMarr*). Either of them can be used to reduce the target variance by a considerable amount: X reduces the variance by 7.58%, and Y by 9.22% when used on their own, see Figure 1. Because they are highly correlated ($r = 0.79$), one could expect them to contain redundant information on the target. However, their combined predictive power is actually larger than the sum of their separate predictive powers, together they reduce the variance by 22.70%. Because of this, the ASV contribution of the second variable ends up being almost double its individual predictive power.

When checking the same example using restricted models, we find that the combined predictive power of the two variables is much more in line with what we expect from two correlated variables, together they reduce the variance by 10.83% under the GAM restriction. The contribution of the second variable is now reduced compared to its individual predictive power by a significant margin.

C. Interactions between pairs of variables

Investigating **Q1**, we measure the interactions between pairs of variables in the datasets to see the prevalence of complex interactions. In the CaCU dataset, we measure the interactions between possible pairs of features and find that complex interactions occur in a large portion of the pairs. The distribution of W and W^r are shown in Figure 2 as a heatmap for the 7626 possible feature pairs. Figure 3 presents the interactions between pairs of features for all four datasets.

When using the restricted models, the large majority of such interactions disappear, see again Figures 2 and 3. The remaining positive interactions can be attributed to the small size of the dataset in both the CaCU and the Garment datasets: due to the low number of samples, measured MSE and training accuracy have relatively large variance themselves. However, in the PM2.5 dataset, some larger positive interactions remain despite the size of the dataset. Investigating these positive interaction values, we find that many of them are *not* accompanied by negative covariances between parts of the prediction model, see Figure 4, which should be impossible as per

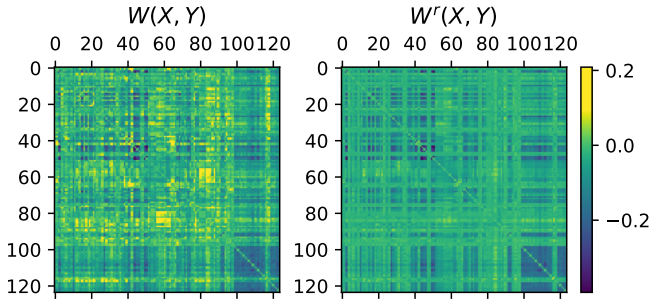

Fig. 2. Heat-map of interactions between features in the CaCU dataset. Values are given as percentages of the original variance of T. Notice the relative absence of positive values on the figure on the right side, which represent complex interactions.

Theorem B. Upon further investigation, we find that these anomalies are caused by the modeling process occasionally training sub-optimal models on the PM2.5 dataset when only a single variable is used, leading to incorrect W^r values. When optionally using the single-variable function F_Y of the two variable GAM models instead of individually trained single-variable models, the anomalies disappear, see Figure 4. For **Q2**, overall, we do not detect examples similar to Example D.

D. Interactions between pairs of feature groups

With larger datasets, it often makes sense to group variables before defining causal relationships. In such cases, we can compute the attribution for the groups instead of single variables while allowing complex interactions within each group. To test the behavior of interactions, we group both semantically and at random. For example, we use a manually compiled semantic grouping of CaCU variables categorizing the 124 features into 13 semantic groups presented in Appendix XI-G.

The interactions observed between pairs of variables or groups are always composed of redundant (negative) and complex (positive) interactions. A negative sum can still include positive interaction terms. In a restricted model, we expect no positive interactions and negative interactions to change to even more negative. We display the distribution of interactions between groups on the CaCU dataset in Figure 6.

For **Q3**, Table III presents the predictive power of GAMs with different groupings of variables along with unrestricted models. We observe that a varying amount of predictive power is based on complex interactions, depending on the dataset. However, in the two examples with semantic groups, the performance decrease in the restricted model is very low, especially compared to using a GAM over individual features.

E. Causality-aware attributions

We test causality-aware contributions on the Telco dataset based on a causality graph of the variables established with the help of domain experts, see Figure 5 and Appendix XI-G. Each topological ordering of the causality graph corresponds to a permutation π for ASV. There are 1134 possible topological orderings of the causality graph with 62 distinct starting subsets of feature groups. In practice, this means training 62 models, which is feasible in our case.

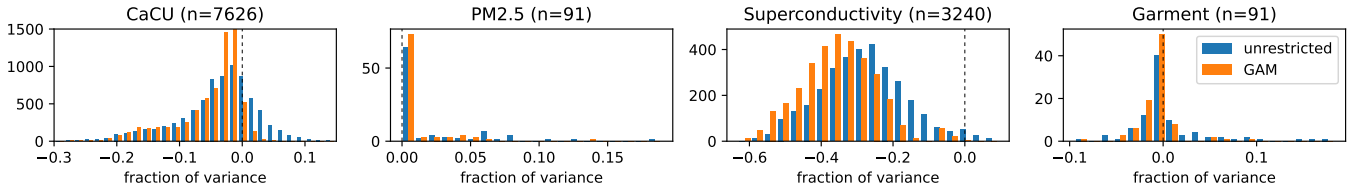

Fig. 3. Histogram of interactions between pairs of features on four datasets (n=number of pairs). Negative values represent redundant information.

TABLE II
CONTRIBUTIONS TO VARIANCE REDUCTION FOR CAUSAL-ATTRIBUTIONS IN THE MOBILE TELECOMMUNICATIONS DATASET

	ϕ_1	ϕ_2	ϕ_3	ϕ_4	ϕ_5	ϕ_6	ϕ_7	ϕ_8	ϕ_9	ϕ_{10}	$\phi_{\mathcal{I}}$
Contributions to σ^2 with GAMs	-14.10	0.00	-0.84	-15.08	-3.15	-5.16	-0.43	-3.47	-1.78	-5.26	-2.80
Contributions to σ^2 with unrestricted models	-14.10	0.00	-1.59	-15.38	-4.08	-5.55	-0.53	-3.76	-2.10	-4.99	0.00

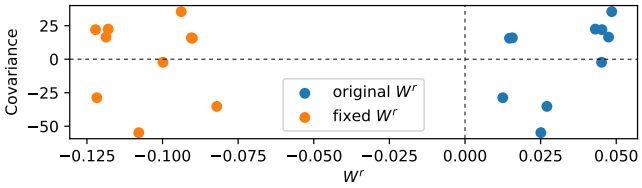

Fig. 4. Initial and fixed covariance values in the PM2.5 dataset.

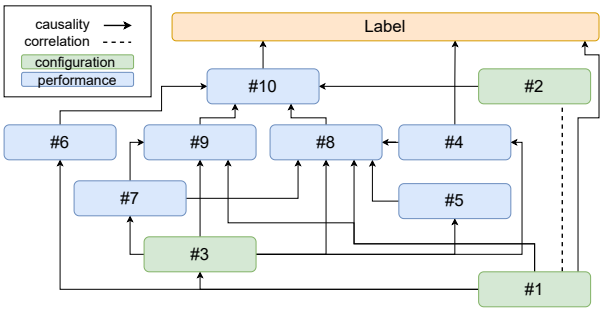

Fig. 5. Causal relations between different feature groups in the mobile telecommunications dataset. See Appendix XI-G for variable names.

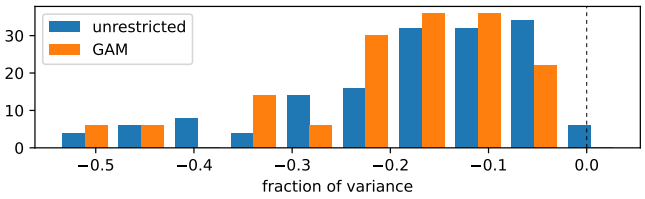

Fig. 6. Histogram of interactions of pairs of semantic groups in the CaCU dataset. Negative values represent redundant information.

Contributions to variance reduction are presented in Table II for ASV with and without the GAM restrictions. Compared to the overall variance reduction from 93.35 to 41.28, both the amount of complex interactions ($\phi_{\mathcal{I}} = -2.8$) and the difference between the two kinds of models are modest, which indicates a correct causal order and lack of confounders. However, we do observe relatively large changes in some of the contributions, indicating that complex interactions indeed occur. From the small but substantial differences, domain experts can identify unexpected interactions among components at different levels of the causal graph.

TABLE III
THE UNEXPLAINED FRACTION OF VARIANCE WITH DIFFERENT GROUPINGS OF VARIABLES USING GAMS VS. THE UNRESTRICTED BASELINE.

Grouping	**# of groups**	Remaining variance %
Communities and Crime Unnormalized		
Features as groups	124	0.4062
Random groups of size 6	20	0.3740
Semantic groups	13	0.3856
Unrestricted model	1	0.3803
PM2.5 Data of Five Chinese Cities		
Features as groups	14	0.8327
Random groups of size 3	4	0.7336
Unrestricted model	1	0.6251
Superconductivity		
Features as groups	81	0.1775
Random groups of size 6	13	0.1193
Unrestricted model	1	0.0977
Productivity Prediction of Garment Employees		
Features as groups	14	0.6151
Random groups of size 3	4	0.5761
Unrestricted model	1	0.5533
Mobile Telecommunications		
Semantic groups	10	0.4722
Unrestricted model	1	0.4422

X. CONCLUSIONS

In this work, we investigated Asymmetric Shapley Values (ASV) for root-cause analysis. We formalized ASV feature attributions in terms of variance reduction and presented examples where ASV produces counter-intuitive results. We proposed using ASV in conjunction with generalized additive models (GAMs), and proved multiple results about the joint behavior, with Theorem B giving a sufficient condition for correct behavior.

We conclude that although ASV generally performs well, it may incorrectly assign contributions in various situations involving complex relationships in real-world datasets. By utilizing GAMs, most of these problematic cases can be mitigated; however, it is still possible for issues to arise under unusual circumstances. Therefore, we recommend considering the potential occurrence of the problematic cases outlined in this paper when applying ASV. Nevertheless, when such cases are absent, ASV remains a highly effective approach for assigning contributions with known causal relationships.

References

[1] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, vol. 30, 2017.

[2] C. Frye, C. Rowat, and I. Feige, "Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability," *NeurIPS*, vol. 33, pp. 1229–1239, 2020.

[3] T. Heskes, E. Sijben, I. G. Bucur, and T. Claassen, "Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models," *NeurIPS*, vol. 33, pp. 4778–4789, 2020.

[4] J. Wang, J. Wiens, and S. Lundberg, "Shapley flow: A graph-based approach to interpreting model predictions," in *AISTATS*. PMLR, 2021, pp. 721–729.

[5] Y. Jung, S. Kasiviswanathan, J. Tian, D. Janzing, P. Blöbaum, and E. Bareinboim, "On measuring causal contributions via do-interventions," in *ICML*. PMLR, 2022, pp. 10 476–10 501.

[6] J. Li, H. X. Tran, T. D. Le, L. Liu, K. Yu, and J. Liu, "Explanatory causal effects for model agnostic explanations," *arXiv:2206.11529*, 2022.

[7] J. Pearl, "The do-calculus revisited," in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, 2012, pp. 3–11.

[8] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton, "Neural additive models: Interpretable machine learning with neural nets," *NeurIPS*, vol. 34, pp. 4699–4711, 2021.

[9] T. J. Hastie, "Generalized additive models," in *Statistical models in S*. Routledge, 2017, pp. 249–307.

[10] S. Bordt and U. von Luxburg, "From shapley values to generalized additive models and back," in *AISTATS*. PMLR, 2023, pp. 709–745.

[11] I. Covert, S. M. Lundberg, and S.-I. Lee, "Understanding global feature contributions with additive importance measures," *NeurIPS*, vol. 33, pp. 17 212–17 223, 2020.

[12] C. Frye, D. de Mijolla, T. Begley, L. Cowton, M. Stanley, and I. Feige, "Shapley explainability on the data manifold," in *ICLR*, 2021.

[13] A. B. Owen, "Sobol' indices and shapley value," *SIAM/ASA Journal on Uncertainty Quantification*, vol. 2, no. 1, pp. 245–251, 2014.

[14] J. Z. Huang, "Functional anova models for generalized regression," *Journal of Multivariate Analysis*, vol. 67, no. 1, pp. 49–71, 1998.

[15] A. Herren and P. R. Hahn, "Statistical aspects of shap: Functional anova for model interpretation," *arXiv:2208.09970*, 2022.

[16] S. Da Veiga, "Kernel-based anova decomposition and shapley effects–application to global sensitivity analysis," *arXiv:2101.05487*, 2021.

[17] H. Chen, J. D. Janizek, S. M. Lundberg, and S. Lee, "True to the model or true to the data?" *CoRR*, vol. abs/2006.16234, 2020.

[18] X. Man and E. P. Chan, "The best way to select features? comparing mda, lime, and shap," *The Journal of Financial Data Science*, vol. 3, no. 1, pp. 127–139, 2021.

[19] C. M. Scavuzzo, J. M. Scavuzzo, M. N. Campero, M. Anegagrie, A. A. Aramendia, A. Benito, and V. Periago, "Feature importance: Opening a soil-transmitted helminth machine learning model via shap," *Infectious Disease Modelling*, vol. 7, no. 1, pp. 262–276, 2022.

[20] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent individualized feature attribution for tree ensembles," *arXiv:1802.03888*, 2018.

[21] C. Molnar, "A guide for making black box models explainable," 2018.

[22] A. Spanos, *Probability theory and statistical inference: Empirical modeling with observational data*. Cambridge University Press, 2019.

[23] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proc. 18th ACM SIGKDD*, 2012, pp. 150–158.

[24] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*. Springer science & business media, 2009.

[25] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," in *NeurIPS*, vol. 30, 2017.

[26] S. Lipovetsky and M. Conklin, "Analysis of regression in game theory approach," *Applied Stochastic Models in Business and Industry*, vol. 17, no. 4, pp. 319–330, 2001.

[27] D. Dua and C. Graff, "UCI machine learning repository," 2017.

[28] X. Liang, S. Li, S. Zhang, H. Huang, and S. X. Chen, "Pm2.5 data reliability, consistency, and air quality assessment in five chinese cities," *Journal of Geophysical Research: Atmospheres*, vol. 121, no. 17, pp. 10,220–10,236, 2016.

[29] K. Hamidieh, "A data-driven statistical model for predicting the critical temperature of a superconductor," *Computational Materials Science*, vol. 154, pp. 346–354, 2018.

[30] M. S. Rahim, A. A. Imran, and T. Ahmed, "Mining the productivity data of garment industry," *International Journal of Business Intelligence and Data Mining*, vol. 1, no. 1, p. 1, 2021.

[31] A. A. Imran, M. N. Amin, M. R. I. Rifat, and S. Mehreen, "Deep neural network approach for predicting the productivity of garment employees," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*. IEEE, apr 2019.

XI. Appendix

A. Existence of E^r

As described in [24], if

$$M = \{g(Y) \,|\, g \text{ is a measurable function and } \mathrm{E}(g(Y)^2) < \infty\},$$

then M is a closed subspace of $L^2(\Omega)$. Let us define

$$K = \left\{ f(X) \,\middle|\, f(X) = \sum_{i=1}^{n} f_i(X_i); f,\ f_i \text{ are measurable and } E(f_i(X_i)^2) < \infty,\ i = 1, 2, \ldots, n \right\}. \tag{26}$$

Then $K \subseteq \overline{K} \subseteq M$, where $\overline{K}$ is the closure of K. The function f such that $E^r(T \,|\, X) = f(X)$ always exists with $E(f(X)^2) < \infty$, if we define $E^r(T \,|\, X_1, \ldots, X_n)$ as

$$E^r(T \,|\, X_1, \ldots, X_n) = \underset{f(X) \in \overline{K}}{\operatorname{argmin}} \sigma^2(T - f(X)). \tag{27}$$

Notice that this does not guarantee $\exists f_i : E^r(T \,|\, X) = f(X) = \sum f_i(X_i)$. In Theorem B, an additional assumption is made to account for this.

B. Lemma about L_T

Lemma A.

$$L_T(X) = \sigma^2(T) - \sigma^2(T - E(T \,|\, X)) = \sigma^2(E(T \,|\, X)) \tag{28}$$

Proof.

$$L_T(X) = \sigma^2(T) - \sigma^2(T - E(T \,|\, X)) \tag{29}$$

$$= 2cov(T, E(T \,|\, X)) - \sigma^2(E(T \,|\, X)) \tag{30}$$

$$= 2cov(T, E(T \,|\, X)) - cov(E(T \,|\, X), E(T \,|\, X)) \tag{31}$$

$$= cov(T, E(T \,|\, X)) + cov(T - E(T \,|\, X), E(T \,|\, X)) \tag{32}$$

$$= cov(T, E(T \,|\, X)) = \sigma^2(E(T \,|\, X)) \tag{33}$$

The second term of (32) is zero because we know that $cov(T - g(T), f(T))$ is zero for all f if g minimizes $\min_g \sigma^2(T - g(T))$ which follows from the Hilbert projection theorem. □

C. Contributions in Example B using E

We are going to calculate ϕ_1' and ϕ_2', the variance reductions assigned to variables X_1 and X_2 using the ASV distal weighting scheme, using the permutation (X_1, X_2). First,

$$E(T \,|\, X_1) = E(4X_1^2 \,|\, X_1) + E(8X_1X_2 \,|\, X_2) + \tag{34}$$

$$+ E(4X_2^2 \,|\, X_1) = 4X_1^2 + 0 + E(4X_2^2) = 4X_1^2 + c. \tag{35}$$

It is known that if $Z \sim N(0, s^2)$, then $\sigma^2(Z^2) = 2s^4$, thus

$$\phi_1' = -L_T(X_1) = \sigma^2(T - E(T \,|\, X_1)) - \sigma^2(T) \tag{36}$$

$$= -\sigma^2(4X_1^2) = -16\,\sigma^2(X_1^2) = -32 \tag{37}$$

because of Lemma A and $E(T \mid X_1, X_2) = T$. Since $T = (2X_1 + 2X_2)^2$ where $(2X_1 + 2X_2) \sim N(0, 8)$,

$$\phi'_0 = \sigma^2(T) = 128 \text{ and} \tag{38}$$

$$\phi'_0 + \phi'_1 + \phi'_2 = 128 - 32 + \phi'_2 = \sigma^2(T \mid X_1, X_2) = 0 \tag{39}$$

$$\text{therefore } \phi'_2 = -96. \tag{40}$$

D. Contributions in Example B using E^r

With using E^r, the behavior is a bit different. We still have $\phi'_0 = \sigma^2(T) = 128$ and $E^r(T \mid X_1) = 4X_1^2$, thus $\phi'_1 = -32$, however $E^r(T \mid X_1, X_2) \neq E(T \mid X_1, X_2)$. The best restricted prediction we can wish for is the sum of the individual expectations, i.e., $E^r(T \mid X_1, X_2) = E(T \mid X_1) + E(T \mid X_2)$, which is the case if $cov(E(T \mid X_1), E(T \mid X_2)) = 0$. Since $E^r(T \mid X_1) = 4X_1^2 + c_1$ and $E^r(T \mid X_2) = 4X_2^2 + c_2$, they are indeed uncorrelated, so $E^r(T \mid X_1, X_2) = 4X_1^2 + 4X_2^2$. From this, $E((T - E^r(T \mid X_1, X_2))^2) = \sigma^2(8X_1X_2) = 64$, and we can determine ϕ_2 as

$$\phi'_0 + \phi'_1 + \phi'_2 = 128 - 32 + \phi'_2 = \sigma^2(T \mid X_1, X_2) = 64,$$

therefore $\phi'_2 = -32$. This means that with E^r, both variables get equal contribution. In this case, $\phi_{\mathcal{I}}$, defined in Section VIII, gets the most contribution: $\phi_{\mathcal{I}} = -64$.

E. Proof of Theorem A

Proof. Because of the independence of X and Y, for any functions $f(X)$ and $g(Y)$

$$\big[\sigma^2(T) - \sigma^2(T - f(X))\big] + \big[\sigma^2(T) - \sigma^2(T - g(X))\big] \tag{41}$$

$$= 2cov(T, f(X)) - \sigma^2(f(X)) + 2cov(T, g(Y)) - \sigma^2(g(Y)) \tag{42}$$

$$= 2cov(T, f(X) + g(Y)) - \sigma^2(f(X) + g(Y)) \tag{43}$$

$$= \big[\sigma^2(T) - \sigma^2(T - (f(X) + g(Y)))\big]. \tag{44}$$

This trivially implies $L_T^r(X) + L_T^r(Y) \leq L_T^r(X, Y)$, as the function $L_T^r(X, Y) = \sigma^2(T) - \sigma^2(T - h(X, Y))$ is maximal for functions of the form $h(X, Y) = h_X(X) + h_Y(Y)$.

We can also use it to prove $L_T^r(X) + L_T^r(Y) \geq L_T^r(X, Y)$, however some care needs to be taken due to the considerations of Appendix XI-A. To do this, let us take a series of measurable square-integrable $h_i, h_{i,X}(X), h_{i,Y}(Y)$ functions such that

$$L_T^r(X, Y) = \lim_{i \to \infty} \big[\sigma^2(T) - \sigma^2(T - h_i(X, Y))\big] \tag{45}$$

$$= \lim_{i \to \infty} \big[\sigma^2(T) - \sigma^2(T - (h_{i,X}(X) + h_{i,Y}(Y)))\big]. \tag{46}$$

Since $L_T^r(X), L_T^r(Y)$ are also maximal for for functions of the given form, this means that for each term of the series h_i

$$\sigma^2(T) - \sigma^2(T - (h_{i,X}(X) + h_{i,Y}(Y)) \leq L_T^r(X) + L_T^r(X), \tag{47}$$

which implies that the same is also true for the limit, i.e.

$$L_T^r(X, Y) \leq L_T^r(X) + L_T^r(X).$$

□

F. Proof of Theorem B

Proof. Let us introduce the notations

$$E^r(T \mid X_1, \dots, X_n, Y_1, \dots, Y_n) = F_X(X) + F_Y(Y) \tag{48}$$

$$= \sum f_{X_i}(X_i) + \sum f_{Y_i}(Y_i) = F \tag{49}$$

$$E^r(T \mid X_1, \dots, X_n) = G, \text{ and } E^r(T \mid Y_1, \dots, Y_m) = H. \tag{50}$$

Since $\sigma^2(T) - \sigma^2(T - A) = 2cov(T, A) - \sigma^2(A)$, thus

$$L_T^r(X_1, \dots, X_n) = \sigma^2(T) - \sigma^2(T - G) = 2cov(T, G) - \sigma^2(G)$$
$$L_T^r(X_1, \dots, X_n) = \sigma^2(T) - \sigma^2(T - H) = 2cov(T, H) - \sigma^2(H)$$

and

$$L_T^r(X_1, \dots, X_n, Y_1, \dots, Y_m) = \sigma^2(T) - \sigma^2(T - F) \tag{51}$$

$$= 2cov(T, F) - \sigma^2(F) \tag{52}$$

$$= 2cov\Big(T, \sum f_{X_i}(X_i)\Big) - \sigma^2\Big(\sum f_{X_i}(X_i)\Big) + \tag{53}$$

$$+ 2cov\Big(T, \sum f_{Y_j}(Y_j)\Big) - \sigma^2\Big(\sum f_{Y_j}(Y_j)\Big) - \tag{54}$$

$$- 2cov\Big(\sum f_{X_i}(X_i), \sum f_{Y_j}(Y_j)\Big). \tag{55}$$

Thus

$$W_T^r(X_1, \dots, X_n; Y_1, \dots, Y_m) \tag{56}$$

$$= \Big(2cov\Big(T, \sum f_{X_i}(X_i)\Big) - \sigma^2\Big(\sum f_{X_i}(X_i)\Big)\Big) - \tag{57}$$

$$- (2cov(T, G) - \sigma^2(G)) + \tag{58}$$

$$+ \Big(2cov\Big(T, \sum f_{Y_i}(Y_i)\Big) - \sigma^2\Big(\sum f_{Y_i}(Y_i)\Big)\Big) - \tag{59}$$

$$- (2cov(T, H) - \sigma^2(H)) + \tag{60}$$

$$+ \Big(-2cov\Big(\sum f_{X_i}(X_i), \sum f_{Y_i}(Y_i)\Big)\Big). \tag{61}$$

Since (58) and (60) are maximal for such expressions, i.e.,

$$2cov\Big(T, \sum f_{X_i}(X_i)\Big) - \sigma^2\Big(\sum f_{X_i}(X_i)\Big) \tag{62}$$

$$\leq 2cov(T, G) - \sigma^2(G), \text{thus} \tag{63}$$

$$W_T^r(X_1, \dots, X_n; Y_1, \dots, Y_m) + c \tag{64}$$

$$= -2cov\Big(\sum f_{X_i}(X_i), \sum f_{Y_i}(Y_i)\Big) \tag{65}$$

for some $c \geq 0$, which proves the proposition. □

G. Datasets and Variables

In the Communities and Crime Unnormalized dataset, we define the following feature groups: Race, Age, Income, Race/Income, Education, Family, Immigration, Housing, Homelessness, Native, Police, Race/Police, Land/Population.

In the Telecommunications dataset, the feature groups indicated by $\phi_1, \dots, \phi_{10}$ are, respectively: Spectrum (1); Antennas, MIMO and modulations (2); Dimensioning (3); Cell load (4); Neighbor cell load (5); UE capability distribution (6); TA distribution (7); Interference (8); Path loss (9); Channel quality (10) and the label is Downlink Throughput. Causal relations are displayed in Figure 5.

For full description of the groups in both the Telco and CaCu datasets, please refer to the source repository.