Journal of Machine Learning Research 24 (2023) 1-42 Submitted 2/22; Revised 9/22; Published 3/23

Faith-Shap: The Faithful Shapley Interaction Index

Che-Ping Tsai CHEPINGT@CS.CMU.EDU
Department of Machine Learning
Carnegie Mellon University
PA 15213, USA

Chih-Kuan Yeh CJYEH@CS.CMU.EDU
Department of Machine Learning
Carnegie Mellon University
PA 15213, USA

Pradeep Ravikumar PRADEEPR@CS.CMU.EDU
Department of Machine Learning
Carnegie Mellon University
PA 15213, USA

Editor: Tommi Jaakkola

Abstract

Shapley values, which were originally designed to assign attributions to individual players in coalition games, have become a commonly used approach in explainable machine learning to provide attributions to input features for black-box machine learning models. A key attraction of Shapley values is that they uniquely satisfy a very natural set of axiomatic properties. However, extending the Shapley value to assigning attributions to interactions rather than individual players, an *interaction index*, is non-trivial: as the natural set of axioms for the original Shapley values, extended to the context of interactions, no longer specify a unique interaction index. Many proposals thus introduce additional less "natural" axioms, while sacrificing the key axiom of efficiency, in order to obtain unique interaction indices. In this work, rather than introduce additional conflicting axioms, we adopt the viewpoint of Shapley values as coefficients of the most faithful linear approximation to the pseudo-Boolean coalition game value function. By extending linear to ℓ-order polynomial approximations, we can then define the general family of *faithful interaction indices*. We show that by additionally requiring the faithful interaction indices to satisfy interaction-extensions of the standard individual Shapley axioms (dummy, symmetry, linearity, and efficiency), we obtain a *unique* Faithful Shapley Interaction index, which we denote Faith-Shap, as a natural generalization of the Shapley value to interactions. We then provide some illustrative contrasts of Faith-Shap with previously proposed interaction indices, and further investigate some of its interesting algebraic properties. We further show the computational efficiency of computing Faith-Shap, together with some additional qualitative insights, via some illustrative experiments.

Keywords: Feature Attribution, Feature Interaction, Interaction Index, Shapley Value, Explainable AI

1. Introduction

Explaining the prediction of a black-box machine learning model via attributions to its features is an increasingly important task. Most approaches have focused on attributions to *individual features*, which does not always suffice to provide insight into the model when there are heavy feature interactions. For instance, when explaining models with text input, we might also ask for attributions to phrases and sequences of words rather than just individual words. Similarly, in Question Answering (QA) (Ye et al., 2021), it is of interest to measure attributions to query answer tuples, rather than just individual entities associated with answers. Such feature interactions are also salient with images as input, where instead of attributions to individual pixels, we might prefer attributions to groups of pixels.

A large class of recent approaches for individual feature attributions reduces the task to a cooperative game theory problem. Given a machine learning model, a test point, and the underlying data distribution, one can devise a "set value function" that takes as input a set of features and outputs the value of that set of features. There are many choices for such a reduction to a set function (Lundberg and Lee, 2017; Sundararajan and Najmi, 2019; Frye et al., 2020; Chen et al., 2020). We can then relate this to a cooperative game theory problem where the features are players, the set function above is the value function of the coalition game that specifies the value of various player coalitions, and we wish to derive feature attributions given such a value function. This meta-approach has led to a slew of explanation approaches when the goal is to obtain individual feature attributions. The key question we focus on in this paper is to obtain attributions to *feature interactions* instead. In this setting, any feature interactions (up to a given order), along with each individual feature, should get some attribution score. This question has attracted some attention in the cooperative game theory and the explainable AI literature, with the broad strategy of extending popular approaches for individual feature attributions, such as Shapley and Banzhaf values (Shapley, 1953; Harsanyi, 1963), to the interaction context. But these existing proposals come with many caveats.

Part of the attraction of the cooperative game theory based explanations above is that for the case of individual feature attributions, if we stipulate some natural axioms such as linearity, symmetry, dummy, and efficiency (detailed in a later section), there exist unique attributions such as Shapley and Banzhaf (depending on the notion of efficiency). Thus we have both a strong axiomatic foundation to the explanations, as well as a very compelling uniqueness result that there can exist no other explanations that satisfy these axioms. These have thus led to an explosion of Shapley value based explanations in the XAI literature that assign attributions to features, data, and even concepts (Lundberg and Lee, 2017; Grömping, 2007; Lindeman, 1980; Owen, 2014; Owen and Prieur, 2017; Datta et al., 2016; Ghorbani and Zou, 2019; Jia et al., 2019; Yeh et al., 2020). However, when we move to the context of feature interactions, while the axioms above have natural extensions from the individual feature to the feature interaction context, they no longer result in a *unique feature attribution value*.

Approaches to address this have thus focused on adding additional less natural axioms to ensure uniqueness. One set of unique feature attributions — Shapley interaction and Banzhaf interaction indices (Grabisch and Roubens, 1999) — derive unique attributions via a *recursive axiom*, which specifies how higher-order feature attributions be derived from

lower order feature interaction attributions (all the way to individual feature attributions). Thus, given the uniqueness at the level of individual feature attributions, we in turn get uniqueness at all levels of interaction attributions. One major caveat of these Shapley interaction and Banzhaf interaction indices is that they do not satisfy the efficiency axiom for interaction feature attributions, and hence can no longer be viewed as distributing the total contribution of the model prediction among all feature interactions. The other caveat is that the recursive axiom, while convenient to extend uniqueness from individual to interaction feature attributions, is much less "natural" when compared with the original Shapley axioms, which specifically defined the forms of first-order indices for certain value functions. To address these caveats, Sundararajan et al. (2020) proposed the *interaction distribution axiom* that entails distributing higher-order interactions to the topmost interaction indices at the expense of impoverished lower-order interactions. This makes the interaction attributions unique for unanimity games (Shapley, 1953), and since these act as a basis for set value functions, by linearity this ensures uniqueness of interaction attributions for general games. The caveat however is that the specified attribution distribution inordinately favors the topmost interactions, which in turn affects the usefulness of both the lower and highest-order interactions as we show in our examples. And arguably, the interaction distribution axioms too are much less natural when compared to the original Shapley axioms. Thus, there remains an open problem to specify a "natural" restriction or axiom that allows for unique interaction attributions.

An additional desideratum is that the feature interaction attributions be cognizant of the *maximum interaction order* of the interaction attributions we require. For instance, with individual feature attributions, the maximum interaction order is one, while with pairwise feature attributions, the maximum interaction order is two. This would allow the explanations to be tailored to the set of possible interactions and satisfy the relevant axioms with respect to just these interactions, instead of all possible subsets of feature interactions.

In this work, rather than devising potentially less natural axioms to ensure uniqueness, we work from yet another viewpoint of Shapley values, that they are faithful to the set value function: for all subsets, the sum of individual feature attributions over a subset should approximate the set value function evaluated on that subset. When formalized as a weighted regression problem, this yields Shapley and Banzhaf values depending on the weights in the weighted regression (Banzhaf III, 1964; Ruiz et al., 1996). We then extend the above weighted regression to feature interactions up to a given maximum interaction order, which then yields what we call Faith-Interaction indices. We show that when restricting to the class of Faith-Interaction indices, together with the (interaction extensions of the individual) Shapley axioms, we obtain a unique interaction index, which we term the Faith-Shap (for Faithful Shapley Interaction) index, which reduces to the individual feature Shapley values when the top interaction order is one. We thus posit Faith-Shap as the natural extension of Shapley values from individual features to interaction indices. Similarly, when the efficiency axiom is replaced by the generalized 2-efficiency axiom, we obtain a unique interaction index, which we term Faith-Banzhaf (for Faithful Banzhaf Interaction) index. The latter has also appeared in other guises in prior work (Hammer and Holzman, 1992; Grabisch et al., 2000). Unlike the other restrictive axioms discussed earlier, here we only require that the explanations be faithful to the model, which has always been a big attraction of Shapley values in the explainable AI (XAI) context. We corroborate the usefulness of these

Faith-Interaction indices by contrasting them with prior indices in two illustrative coalition games, as well as real-world XAI applications. We then discuss the algebraic properties of Faithful Shapley Interaction index by relating them to cardinal indices, i.e. indices that can be expressed as a linear combination of marginal contributions, as well as in terms of approximations to multilinear extensions of the coalition set value function. An additional benefit of the Faith Interaction indices is that the estimation becomes much more efficient via leveraging the weighted linear regression formulation, which we validate in our experiments.

2. Preliminaries

2.1 Notations

Suppose we are given a black-box model $f : \mathcal{X} \mapsto \mathbb{R}$, with input domain $\mathcal{X} \subseteq \mathbb{R}^d$; and suppose we wish to explain its prediction at a given test point $x \in \mathcal{X}$. Suppose also given the tuple f, x (and possibly with additional information about the underlying data distribution on which f is trained on, and from which x is drawn), there is a well-defined set function $v_x : 2^d \to \mathbb{R}$. We can interpret such a set function as specifying the value of a subset of the set of d features. Many popular explanations employ such a reduction of the model and its prediction context to set value functions; see Ribeiro et al. (2016); Lundberg and Lee (2017); Sundararajan et al. (2020) for many examples. When clear from the context, and for notational simplicity, we will often omit x and simply use v to denote the set function. Such a reduction allows us to leverage results from cooperative game theory, by relating the set of features to a set of players, and the set function above as specifying the values of coalitions of players.

We are then interested in quantifying the importance of interactions between different features up to some order $\ell \in [d]$. Note that in this context, when we mean interactions between features, we mean non-self interactions between distinct features, since self-self interactions could simply be identified with the individual features. In other words, we require an importance function $\mathcal{E}$ which for each coalition $S \subseteq [d]$ where $0 \leq |S| \leq \ell$, outputs a scalar $\mathcal{E}_S(v, \ell)$. Let $\mathcal{S}_\ell$ denote the set of all subsets of $[d]$ with size less than or equal to ℓ; the size of this set can be seen to be $d_\ell \stackrel{\text{def}}{=} \sum_{j=0}^{\ell} \binom{d}{j}$. We then use the shorthand $\mathcal{E}(v, \ell) = (\mathcal{E}_S(v, \ell))_{S \in \mathcal{S}_\ell} \in \mathbb{R}^{d_\ell}$. To simplify notation, we omit braces for small sets and write $T \cup i$ to represent $T \cup \{i\}$.

2.2 Definitions

We begin by recalling the concept of discrete derivatives.

Definition 1 *(Discrete Derivative) Given a set function $v : 2^d \mapsto \mathbb{R}$ and two finite disjoint coalitions $S, T \subseteq [d]$ with $S \cap T = \emptyset$, the S-derivative of v at T, $\Delta_S(v(T))$, is defined recursively as follows:*

$$\Delta_i v(T) = v(T \cup i) - v(T), \quad \forall i \in [d], \text{ and} \tag{1}$$

$$\Delta_S(v(T)) = \Delta_i[\Delta_{S \setminus i}(v(T))] = \sum_{L \subseteq S} (-1)^{|S|-|L|} v(T \cup L), \forall i \in S. \tag{2}$$

The second equality in Eqn. (2) can be shown via induction on S (Fujimoto et al., 2006). As an illustration of discrete derivatives, for a subset S of size 2, the discrete derivative can be written as

$$\Delta_{\{i,j\}}v(T) = v(T \cup \{i,j\}) - v(T \cup j) - v(T \cup i) + v(T).$$

$\Delta_{\{i,j\}}v(T)$ captures the joint effect of features i and j co-occurring compared to the individual effects of i and j. If $\Delta_{\{i,j\}}v(T) > 0$ (resp. < 0), we say i and j have positive (resp. negative) interaction effect in the presence of T since the presence of i increases (resp. decreases) the marginal contribution of j to coalition T. Following the intuition from the two features example, the discrete derivative $\Delta_S(v(T))$ can be viewed as a measurement of the *marginal interaction of S in the presence of T*. When a set of features have a positive (negative) interaction effect, the discrete derivative is positive (negative). Discrete derivatives play a fundamental role in measurement of interaction effects. As we will see in the following section, the Shapley and Banzhaf interaction indices can be viewed as a weighted average of S-derivatives over all subsets $T \subseteq [d] \backslash S$.

Next, let us recall the concept of the Möbius transform.

Definition 2 *(Möbius transform) Given set function $v : 2^d \mapsto \mathbb{R}$, the Möbius transform of $v(\cdot)$ is*

$$a(v,S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} v(T) \quad \textit{for all} \quad S \subseteq [d]. \tag{3}$$

An important property (Shapley, 1953) of the Möbius transform is that any set function $v(\cdot)$ can be expressed as:

$$v = \sum_{R \subseteq [d]} a(v,R)\, v_R, \tag{4}$$

where v_R for any $R \subseteq [d]$ has the form $v_R(S) = 1$ if $S \supseteq R$ and 0 otherwise; and is also known as a *unanimity game* value function in game theory. Eqn. (4) states that any set function can be expressed as a linear combination of these unanimity game value functions (so that $\{v_R\}_{R \subseteq [d]}$ form a basis for real-valued set value functions), with the Möbius transforms $a(v,R)$ as their coefficients. Note that if an interaction index satisfies the **interaction linearity axiom** (to be discussed in the sequel), the interaction index for general set value functions can be expressed as a linear combination of the interaction indices for unanimity games.

3. Background: Axioms for Interaction Indices

In this section, we present natural extensions of Shapley axioms for individual features to the feature interactions (Grabisch and Roubens, 1999; Sundararajan et al., 2020). We then discuss the key interaction indices proposed so far in the literature — the Shapley interaction index, Banzhaf interaction index and Shapley-Taylor interaction index — with respect to these axioms. In all these axioms, we allow for dependence on the maximum interaction order $\ell \in [d]$. A summarization of axioms that these interaction indices satisfy is in Table 1.

Indices	Interaction linearity	Interaction symmetry	Interaction dummy	Interaction efficiency	Interaction recursive	Generalized Interaction 2-efficiency	Interaction distribution	Is Faith-Interaction Index
Shapley Interaction	✓	✓	✓		✓			
Banzhaf Interaction	✓	✓	✓		✓	✓		
Shapley Taylor	✓	✓	✓	✓			✓	
Faithful Shapley	✓	✓	✓	✓				✓
Faithful Banzhaf	✓	✓	✓			✓		✓

Table 1: A table of axioms that different interaction indices satisfy.

Axiom 3 *(Interaction Linearity): For any maximum interaction order $\ell \in [d]$, and for any two set functions v_1 and v_2, and any two scalars $\alpha_1, \alpha_2 \in \mathbb{R}$, the interaction index satisfies:* $\mathcal{E}(\alpha_1 v_1 + \alpha_2 v_2, \ell) = \alpha_1 \mathcal{E}(v_1, \ell) + \alpha_2 \mathcal{E}(v_2, \ell)$.

The interaction linearity axiom states that the feature interaction index is a linear functional of the set function $v(\cdot)$. It ensures that the corresponding indices scale with the value function $v(\cdot)$.

Axiom 4 *(Interaction Symmetry): For any maximum interaction order $\ell \in [d]$, and for any set function $v : 2^d \mapsto \mathbb{R}$ that is symmetric with respect to elements $i, j \in [d]$, so that $v(S \cup i) = v(S \cup j)$ for any $S \subseteq [d] \backslash \{i, j\}$, the interaction index satisfies: $\mathcal{E}_{T \cup i}(v, \ell) = \mathcal{E}_{T \cup j}(v, \ell)$ for any $T \subseteq [d] \backslash \{i, j\}$ with $|T| < \ell$.*

The interaction symmetry axiom entails that if the value function treats two features the same, their corresponding feature interaction index values should be the same as well.

Axiom 5 *(Interaction Dummy): For any maximum interaction order $\ell \in [d]$, and for any set function $v : 2^d \mapsto \mathbb{R}$ such that $v(S \cup i) = v(S)$ for some $i \in [d]$ and for all $S \subseteq [d] \backslash \{i\}$, the interaction index satisfies: $\mathcal{E}_T(v, \ell) = 0$ for all $T \in \mathcal{S}_\ell$ with $i \in T$.*

The interaction dummy axiom entails that a dummy feature $i \in [d]$ that has no influence on the function v should have no interaction effect with the other features.

Axiom 6 *(Interaction Efficiency): For any maximum interaction order $\ell \in [d]$, and for any set function $v : 2^d \to \mathbb{R}$, the interaction index satisfies: $\sum_{S \in \mathcal{S}_\ell \backslash \emptyset} \mathcal{E}_S(v, \ell) = v([d]) - v(\emptyset)$ and $\mathcal{E}_\emptyset(v, \ell) = v(\emptyset)$.*

The interaction efficiency ensures that the interaction index distributes the total value $v([d])$ among the different subsets in $\mathcal{S}_\ell$. This axiom lends itself a natural explanation of $\mathcal{E}_S(v, \ell)$: it represents the marginal contribution that the group S makes to the total value, which has also been considered by Sundararajan et al. (2020). As we will detail in the sequel, some of the recently proposed interaction indices do not satisfy such an efficiency axiom. For instance, the chaining interaction and Shapley interaction indices only require the total sum of *individual feature importances* to sum to $v([d]) - v(\emptyset)$, without consideration of the higher-order interaction importances.

Challenge: Lack of Uniqueness: These axioms are natural extensions to the interaction setting of classical axioms for individual feature attributions; see Fujimoto et al. (2006); Grabisch and Roubens (1999) for a counterpart of these interaction axioms without consideration of the maximum interaction order $\ell \in [d]$. As Sundararajan et al. (2020) note, though the linearity, symmetry, dummy, and efficiency axioms uniquely specify a feature attribution when the maximum interaction order $\ell = 1$ (i.e. for individual feature attributions), they no longer do when $\ell > 1$. In other words, there could exist many interaction indices that all satisfy the axioms specified above. A big attraction of the individual Shapley value was its uniqueness given the corresponding individual attribution axioms. Accordingly, a line of work has focused on specifying additional axioms that together specify a unique interaction index.

Axiom 7 *(Recursive Interaction): For any maximum interaction order $2 \leq \ell \leq d$, and for any set function $v : 2^d \to \mathbb{R}$, and for any $j \in [d]$, let the reduced set functions $v^{[d]\setminus j}, v^{[d]\setminus j}_{\cup j} : 2^{d-1} \to \mathbb{R}$ be defined as:*

$$\text{for all } \; T \subseteq [d]\setminus j, \quad v^{[d]\setminus j}(T) = v(T), \quad \text{and} \quad v^{[d]\setminus j}_{\cup j}(T) = v(T \cup j) - v(j).$$

Then the interaction index satisfies: $\mathcal{E}_S(v,\ell) = \mathcal{E}_{S\setminus j}(v^{[d]\setminus j}_{\cup j}, \ell) - \mathcal{E}_{S\setminus j}(v^{[d]\setminus j}, \ell), \;\; \forall S \in \mathcal{S}_\ell$ with $|S| \geq 2$.

The recursive axiom above is an extension of the recursive axiom of Grabisch and Roubens (1999) to account for arbitrary maximum interaction orders. The axiom can be informally interpreted as "how does the presence or absence of feature j influence the share of feature set S". But more importantly (and the reason it is termed the recursive axiom) is that it specifies how higher-order interaction scores are *uniquely determined* given lower-order interaction indices. By recursion, the higher-order interaction indices are thus uniquely specified given just the singleton feature attributions. The reason this helps with uniqueness is that so long as the axioms entail unique singleton attributions, together with this recursive axiom, they would entail unique interaction attributions. Thus, we argue that the recursive axiom is less "natural" compared to previously introduced axioms since the recursive axiom only ensures the uniqueness property, at the potential expense of other axiomatic properties.

Shapley Interaction Index: Grabisch and Roubens (1999) thus show that there is a unique interaction index that satisfies the interaction linearity, symmetry, dummy, and the recursive axioms (but not the interaction efficiency axiom), and whose restrictions to singleton sets correspond to Shapley values. They term this interaction index Shapley interaction index. This Shapley interaction index has the following closed form:

$$\mathcal{E}^{\text{Shap}}_S(v,\ell) = \sum_{T \subseteq [d]/S} \frac{|T|!(d-|S|-|T|)!}{(d-|S|+1)!} \Delta_S(v(T)), \quad \forall S \in \mathcal{S}_\ell. \tag{5}$$

A critical caveat of the resulting Shapley interaction value is that it no longer satisfies the interaction efficiency axiom when the maximum interaction order $\ell > 1$. Indeed, simply summing the contributions to singleton sets (i.e. the classical individual attribution Shapley values) is already equal to $v([d]) - v(\emptyset)$, so the only way for the interaction efficiency axiom to be satisfied if all the other interaction attributions sum to zero, which they do not.

Banzhaf Interaction Index: Grabisch and Roubens (1999) further show that there is a unique interaction index that satisfies the interaction linearity, symmetry, dummy, and recursive axioms (but not the interaction efficiency axiom), and whose restrictions to singleton sets correspond to the Banzhaf values. They term this interaction index Banzhaf interaction index, which has the following closed form:

$$\mathcal{E}_S^{\text{Bzf}}(v,\ell) = \sum_{T \subseteq [d]/S} \frac{1}{2^{d-|S|}} \Delta_S(v(T)), \quad \forall S \in \mathcal{S}_\ell. \tag{6}$$

It can be again shown that the Banzhaf interaction index does not satisfy the interaction efficiency axiom even when $\ell = 1$; though they do satisfy the generalized 2-efficiency axiom, which can be stated as follows.

Axiom 8 *(Generalized Interaction 2-Efficiency): Define the reduced function $v_{[ij]} : 2^{d-1} \to \mathbb{R}$ given any $i, j \in [d]$ as $v_{[ij]}(S) = v(S)$ for all sets S containing both i and j, and $v_{[ij]}(S \cup [ij]) = v(S \cup \{i,j\})$ for all S containing neither i nor j. That is, the reduced function considers features i and j together as a group $[ij]$. Then the interaction index satisfies: $\mathcal{E}_{S\cup[ij]}(v_{[ij]},\ell) = \mathcal{E}_{S\cup i}(v,\ell) + \mathcal{E}_{S\cup j}(v,\ell)$ for all $S \subseteq [d]\backslash\{i,j\}$, and $\ell = |S| + 1$.*

The generalized interaction 2-efficiency axiom above is an extension of the generalized 2-efficiency axiom of Grabisch and Roubens (1999) to account for arbitrary maximum interaction orders. It states that when features i, j form a group in the set function $v_{[ij]}$ with $d-1$ features, the importance of $S \cup [ij]$ equals the sum of importances of $S \cup i$ and $S \cup j$ with respect to the original set value function. When $S = \emptyset$ and $\ell = 1$, it reduces to the classical 2-efficiency axiom (Harsanyi, 1963) that indicates that the importance of $[ij]$ as a group should be equal to the sum of importances of individual features i and j.

Shapley Taylor Interaction Index: Sundararajan et al. (2020) stipulate an additional *interaction distribution (ID) axiom*, which can be stated as follows.

Axiom 9 *(Interaction distribution (Sundararajan et al., 2020)): Define v_T parameterized by a set $T \subseteq [d]$ as $v_T(S) = 0$ if $T \not\subseteq S$ and $v_T(S) = 1$ otherwise. Then for all $\ell \in [d]$, and for all S with $S \not\subseteq T$ and $|S| < \ell$, the interaction index satisfies: $\mathcal{E}_S(v_T,\ell) = 0$.*

The key idea behind the ID axiom is to uniquely specify an interaction index for unanimity games $\{v_T\}_{T\subseteq[d]}$, given the interaction linearity, symmetry, dummy, and efficiency axioms. Since unanimity games form a basis for the set of all games, in the presence of interaction linearity axiom, we then get unique interaction indices. They thus show that there exists a unique interaction index that satisfies interaction linearity, symmetry, dummy, efficiency, and interaction distribution axioms and which they term Shapley Taylor index (for reasons which will become clearer in a later section when we discuss algebraic properties of various interaction indices). The Shapley Taylor interaction index has the following closed form:

$$\mathcal{E}_S^{\text{Taylor}}(v,\ell) = \begin{cases} \Delta_S(v(\emptyset)) & \text{, if } |S| < \ell. \\ \sum_{T\subseteq[d]/S} \frac{|T|!(d-|T|-1)!|S|}{d!} \Delta_S(v(T)) & \text{, if } |S| = \ell. \end{cases} \tag{7}$$

A key advantage of this interaction index is that it depends on the maximum interaction order ℓ, in contrast to previously proposed interaction indices such as the Shapley interaction and Banzhaf interaction indices. Indeed, in order for an interaction index to satisfy the

interaction efficiency axiom for maximum interaction order ℓ, it has to distribute the contributions among subsets in $\mathcal{S}_\ell$, and hence has to be cognizant of the maximum interaction order ℓ. However, a key caveat of the interaction distribution axiom is that the specified attribution distribution inordinately favors the topmost interaction. As can be seen from Eqn.(7), the importance of a set S with $|S| < \ell$ is only specified by the marginal contribution of S in the presence of the empty set, and not the presence of other subsets $T \subseteq [d] \backslash S$. This impoverishes lower-order interactions, which in turn hurts the meaningfulness of both lower and highest-order interactions as we will show in Section 5.

Thus a key open question that this section has made salient is: how do we more naturally constrain interaction indices beyond interaction linearity, symmetry, dummy, and efficiency axioms, so as to obtain a unique interaction index?

4. Faith-Interaction Indices

In this section, in contrast to additional axioms, we draw from another viewpoint of singleton Shapley feature attributions: that they are faithful to the underlying value function.

Faithfulness of Singleton Shapley Values: Given singleton feature attributions $\{\mathcal{E}_i\}_{i\in[d]}$, we can require that:

$$v(S) \approx \sum_{i\in S} \mathcal{E}_i,\ \forall S \subseteq [d].$$

Note that we can only ask for approximate rather than exact equality for all sets S, since exact equality would entail we solve 2^d linear equalities (corresponding to the subsets of $[d]$) with d variables (corresponding to the d singleton feature attributions $\{\mathcal{E}_i\}_{i\in[d]}$), which may not always have a feasible solution. One approach to formalize such approximate equality is via weighted regression:

$$\min_{\mathcal{E}\in\mathbb{R}^{d+1}} \sum_{S\subseteq[d]} \mu(S) \left(v(S) - \mathcal{E}_\emptyset - \sum_{i\in S} \mathcal{E}_i \right)^2, \tag{8}$$

where $\mu : 2^{[d]} \mapsto \mathbb{R}^+ \cup \{\infty\}$ is some weighting over the subsets $S \subseteq [d]$ which can be interpreted as the importance of different coalitions. Note that the range of μ is the extended positive reals. When $\mu(S) = \infty$ for some sets S, we can interpret the above as solving the constrained problem:

$$\min_{\mathcal{E}\in\mathbb{R}^{d+1}} \sum_{S\subseteq[d]\,:\,\mu(S)<\infty} \mu(S) \left(v(S) - \sum_{i\in S} \mathcal{E}_i \right)^2 \text{ s.t. } v(S) = \sum_{i\in S} \mathcal{E}_i,\ \forall S :\ \mu(S) = \infty.$$

It has been shown that we can recover the singleton Shapley values as the solution of the weighted regression problem above by setting $\mu(S) \propto \frac{d-1}{\binom{d}{|S|}\,|S|\,(d-|S|)}$ and $\mu(\emptyset) = \mu([d]) = \infty$ (Charnes et al., 1988). And we can recover singleton Banzhaf values by using the uniform distribution $\mu(S) = 1/2^d$ (Hammer and Holzman, 1992).

From Singleton Attributions to Interaction Indices: In this section, we consider the generalization of the above to *interaction indices*, so that we now require:

$$v(S) \approx \sum_{T \subseteq S, |T| \leq \ell} \mathcal{E}_T(v, \ell), \ \forall S \subseteq [d].$$

Again here we ask for approximate rather than exact equality since when the order of interactions is less than the number of features, so that $\ell < d$, the latter would entail we solve 2^d linear equalities with d_ℓ variables, which may not always have a feasible solution. Accordingly, we consider the following weighted regression problem as a formalization of the above:

$$\mathcal{E}(v, \ell) \ = \underset{\mathcal{E} \subseteq \mathbb{R}^{d_\ell}}{\arg\min} \sum_{S \subseteq [d]} \mu(S) \left(v(S) - \sum_{T \subseteq S, |T| \leq \ell} \mathcal{E}_T(v, \ell) \right)^2, \tag{9}$$

where $\mu : 2^d \to \mathbb{R}^+ \cup \{\infty\}$ is a coalition weighting function. And as before of $\mu(S) = \infty$ for some sets S, we can interpret above as solving the constrained problem:

$$\begin{aligned} \mathcal{E}(v, \ell) \ &= \underset{\mathcal{E} \subseteq \mathbb{R}^{d_\ell}}{\arg\min} \sum_{S \subseteq [d] \,:\, \mu(S) < \infty} \mu(S) \left(v(S) - \sum_{T \subseteq S, |T| \leq \ell} \mathcal{E}_T(v, \ell) \right)^2 \\ &\text{s.t. } v(S) = \sum_{T \subseteq S, |T| \leq \ell} \mathcal{E}_T(v, \ell), \ \ \forall S : \mu(S) = \infty. \end{aligned} \tag{10}$$

We note that the range of the weighting function μ is not allowed to include zero since it is a necessary condition to ensure that there exists a unique minimizer (See Proposition 26 in the Appendix). This is not an issue in practice since we can always choose an arbitrary small positive value instead of zero to approximate the intended constraint that $\mu(S) = 0$ for some $S \subseteq [d]$.

We can also see from Eqn. (9) that when the weighting function is infinite for many subsets, this entails corresponding equality constraints on the interaction index, which may not have a feasible solution. We thus consider the following set of what we term *proper* weighting functions.

Definition 10 *(Proper weighting function) We say that a weighting function $\mu : 2^d \mapsto \mathbb{R}^+ \cup \{\infty\}$ is proper if $\mu(S)$ is finite for all $S \subseteq [d]$ with $1 \leq S \leq d-1$.*

This then leads to our definition of Faith-interaction indices.

Definition 11 *(Faith-Interaction Indices): We say that $\mathcal{E}$ is a Faith-Interaction index, given any set value function $v : 2^d \to \mathbb{R}$ and any maximum interaction order $\ell \in [d]$, if there exists a proper weighting function $\mu : 2^d \to \mathbb{R}^+ \cup \{\infty\}$ such that $\mathcal{E}(v, \ell)$ minimizes the corresponding weighted regression objective in Eqn.*(10)*.*

When the coalition weighting function μ is fully finite so that $\mu(S)$ are finite for all sets $S \subseteq [d]$, Faith-interaction indices have a simple closed-form expression as detailed in the following proposition.

Proposition 12 *Any Faith-Interaction index $\mathcal{E}(v, \ell)$ with respect to a finite weighting function $\mu(\cdot)$ has the form:*

$$\mathcal{E}(v,\ell) = \left(\sum_{S\subseteq[d]} \mu(S)p(S)p(S)^T\right)^{-1} \sum_{S\subseteq[d]} \mu(S)v(S)p(S), \tag{11}$$

where $p : 2^{[d]} \to \{0,1\}^{d_\ell}$ is specified as: $p(S)[T] = \mathbb{1}[(T \subseteq S)]$ for any $T \in \mathcal{S}_\ell$.

When the coalition weighting function $\mu(\cdot)$ is not fully finite, we have a linearly constrained least squares problem that does not have a closed form, but whose solution can be characterized via its Lagrangian (see more details in Proposition 29 in the Appendix).

4.1 Axiomatic Characterization of Faith-Interaction Indices

In this section, we investigate the axiomatic properties of our class of Faith-Interaction indices. We first show that all faith-interaction indices satisfy the interaction linearity axiom.

Proposition 13 *Faith-Interaction indices $\mathcal{E}$ satisfy the interaction linearity axiom.*

For Faith-Interaction indices corresponding to finite coalition, weighting functions $\mu(\cdot)$, this result easily follows from Proposition 12 that these are linear functionals of the set value function $v(\cdot)$. For Faith-Interaction indices where the weighting function is no longer finite for some sets $S \in \{\emptyset, [d]\}$, they solve a linearly constrained least squares problem which does not have a closed-form solution. But by a more nuanced analysis of its Lagrangian, we can again show that the interaction indices are linear functionals of the set value function $v(\cdot)$.

We next show that Faith-Interaction indices also satisfy the interaction symmetry axiom provided that the weighting functions are permutation invariant ("symmetric"), and hence the weighting functions only depend on the size of the set.

Proposition 14 *Faith-Interaction indices $\mathcal{E}$ satisfy the interaction symmetry axiom if and only if the weighting functions are permutation invariant, and hence only depend on the size of the set so that $\mu(S)$ is only a function of $|S|$.*

We next consider the dummy axiom.

Proposition 15 *Faith-Interaction indices $\mathcal{E}$ satisfy the interaction dummy axiom if the features behave independently of each other when forming coalitions in the weighting function so that the coalition weighting functions can be expressed as $\mu(S) \propto \prod_{i\in S} p_i \prod_{j\notin S}(1-p_j)$ for all $S \subseteq [d]$, where $0 < p_i < 1$ is the probability of the feature i to be present.*

Proposition 15 implies that a dummy feature has no impact on other features when the weighting function treats features independently.

So far, we have analyzed when Faith-Interaction indices satisfy the interaction linearity, symmetry, and dummy axioms. When they satisfy all three simultaneously, and the coalition weighting function is finite, then we can show that the latter has a specific algebraic form.

Theorem 16 *Faith-Interaction indices $\mathcal{E}$ with a finite weighting function satisfy the interaction linearity, symmetry, and dummy axioms if and only if the weighting function $\mu(\cdot)$ has the following form:*

$$\mu(S) \propto \sum_{i=|S|}^{d} \binom{d-|S|}{i-|S|}(-1)^{i-|S|} g(a,b,i), \quad \text{where } g(a,b,i) = \begin{cases} 1 & \text{if } i = 0 \\ \prod_{j=0}^{j=i-1} \frac{a(a-b)+j(b-a^2)}{a-b+j(b-a^2)} & \text{if } 1 \le i \le d, \end{cases} \tag{12}$$

for some $a, b \in \mathbb{R}^+$ with $a > b$ such that $\mu(S) > 0$ for all $S \subseteq [d]$.

Theorem 16 shows the surprising fact that Faith-Interaction indices satisfying the interaction linearity, symmetry, and dummy axioms with finite weighting functions have only two degrees of freedom: $a, b \in \mathbb{R}$. Given these, we can fully specify the weighting function, and hence the corresponding Faith-Interaction indices. In Appendix D, we additionally show that the condition $1 > a > b \ge a^2 > 0$ ensures that $\mu(\cdot)$ is positive everywhere and also provides generalized guidance on setting the values a, b.

Faith-Banzhaf Interaction Index: As a first application of this theorem, suppose in addition to the three axioms above, we additionally require the Faith-Interaction indices to satisfy generalized 2-efficiency. The following theorem shows that there is a unique Faith-Interaction index satisfying these four axioms, which we term the Faith-Banzhaf index.

Theorem 17 *(Faith-Banzhaf) For any $d \ge 3$, there is a unique Faith-Interaction index that satisfies the interaction linearity, symmetry, dummy, and generalized 2-efficiency axioms, with its coalition weighting function given as $\mu(S) \propto \frac{1}{2^d}$ for all $S \subseteq [d]$. We term this unique interaction index as **Faithful Banzhaf Interaction index** (Faith-Banzhaf), which has the form:*

$$\mathcal{E}_S^{\text{F-Bzf}}(v,\ell) = a(v,S) + (-1)^{\ell-|S|} \sum_{T \supseteq S, |T|>\ell} \left(\frac{1}{2}\right)^{|T|-|S|} \binom{|T|-|S|-1}{\ell-|S|} a(v,T), \forall S \in \mathcal{S}_\ell, \tag{13}$$

where $a(v,\cdot)$ is the Möbius transform of $v(\cdot)$. Moreover, its highest-order interaction terms coincide with corresponding interaction terms from the Banzhaf interaction index introduced earlier:

$$\mathcal{E}_S^{\text{F-Bzf}}(v,\ell) = \sum_{T \subseteq [d] \setminus S} \frac{1}{2^{d-|S|}} \Delta_S(v(T)) \quad \text{for all } S \in \mathcal{S}_\ell \text{ with } |S| = \ell. \tag{14}$$

Our derivation of Faith-Banzhaf indices follows the pseudo-Boolean function approximation results from Grabisch et al. (2000).

Faith-Shapley Interaction Index: When moving from generalized 2-efficiency to the more natural interaction efficiency axiom, we have the following proposition.

Proposition 18 *Faith-Interaction indices satisfy the interaction efficiency axiom if and only if the weighting functions satisfy $\mu(\emptyset) = \mu([d]) = \infty$.*

That the condition in the proposition is sufficient is a straight-forward consequence of the fact that $\mu(\emptyset) = \mu([d]) = \infty$ entails that the corresponding linear constraint be exactly satisfied, so that: $\sum_{S\in\mathcal{S}_\ell} \mathcal{E}_S(v,\ell) = v([d])$ and $\mathcal{E}_\emptyset(v,\ell) = v(\emptyset)$, which is precisely the interaction efficiency axiom. We now have the machinery to present our main result on the unique Faith-Interaction index that satisfies the four (interaction counterparts of the) standard axioms that the singleton Shapley value satisfies.

Theorem 19 *(Faith-Shap) There is a unique Faith-Interaction index that satisfies the interaction linearity, symmetry, dummy, and efficiency axioms, with its coalition weighting function given as:*

$$\mu(S) \propto \frac{d-1}{\binom{d}{|S|}\,|S|\,(d-|S|)} \text{ for all } S \subseteq [d] \text{ with } 1 \le |S| \le d-1, \text{ and } \mu(\emptyset) = \mu([d]) = \infty. \tag{15}$$

We term this unique interaction index as the ***Faithful Shapley Interaction index*** *(Faith-Shap), which has the form:*

$$\mathcal{E}_S^{\text{F-Shap}}(v,\ell) = a(v,S) + (-1)^{\ell-|S|}\frac{|S|}{\ell+|S|}\binom{\ell}{|S|}\sum_{T\supset S,|T|>\ell}\frac{\binom{|T|-1}{\ell}}{\binom{|T|+\ell-1}{\ell+|S|}}a(v,T), \quad \forall S\in\mathcal{S}_\ell, \tag{16}$$

where $a(v,\cdot)$ is the Möbius transform of $v(\cdot)$. Moreover, its highest-order interaction terms can be expressed as a weighted average of discrete derivatives:

$$\mathcal{E}_S^{\text{F-Shap}}(v,\ell) = \frac{(2\ell-1)!}{((\ell-1)!)^2}\sum_{T\subseteq[d]\setminus S}\frac{(\ell+|T|-1)!(d-|T|-1)!}{(d+\ell-1)!}\Delta_S(v(T)) \quad \text{for all } S\in\mathcal{S}_\ell \text{ with } |S|=\ell. \tag{17}$$

When the maximum interaction order $\ell = 1$, so that we only require singleton feature contributions, the explanation coincides with the classical singleton Shapley values. Thus for larger orders with $\ell > 1$, Faith-Shap can be seen to be a "natural" generalization of the first-order Shapley value. Note that the set of axioms it satisfies are (interaction extensions of) the classical linearity, symmetry, dummy, and efficiency axioms. As noted before in an interaction context these axioms alone do not uniquely specify an interaction index. In contrast to the less intuitive axioms such as recursive and interaction distribution axioms, we merely require an interaction extension of the faithfulness property of singleton Shapley values: that the interaction Shapley values approximate the given set value function for all possible subsets.

5. Contrasting Faith-Interaction with other Interaction Indices

In this section, we compare our Faith-Interaction indices, specifically Faith-Shap, with the other interaction indices introduced earlier.

Comparison with Shapley Interaction and Banzhaf Interaction Indices: As noted earlier, the Shapley interaction and Banzhaf interaction indices do not satisfy the interaction efficiency axiom, which states that the sum of interaction weights should equal the difference

between the value function evaluated over the complete and empty sets. A critical advantage of the interaction efficiency axiom is that it forces the interaction index to distribute a fixed contribution (difference between the value function evaluated over the complete and empty sets) among the different interactions; without such a distributive requirement, the resulting weights can become quite non-intuitive. For instances of such non-intuitive behaviors, we refer to Sundararajan et al. (2020), who provided many simple examples where the sum of Shapley interaction values over all subsets diverges as the number of features increases, even when the value function is bounded and $v([d]) = 1$. Another caveat with these two interaction indices is that they are not cognizant of the maximum interaction order, and hence we cannot compute Shapley values that differ with varying maximum interaction orders.

Comparison with Shapley Taylor index: The Shapley Taylor index does satisfy the four axioms of interaction linearity, symmetry, dummy, and efficiency. However, as noted earlier these four axioms do not uniquely determine interaction indices. The fifth axiom Shapley Taylor index then imposes for uniqueness is the interaction distribution axiom, which has caveats of imbalanced distributions of values to coalitions of different orders, namely, inordinately favoring the maximum interaction order. In particular, the interaction distribution axiom states that higher than max-order interaction values (order $> \ell$) be distributed to the max-order interactions (order $= \ell$), but these max-order terms end up unable to solely explain all higher-order interactions. On the other hand, it entails lower than max order interactions (order $< \ell$) that do not take into account sub-coalitions other than the empty set, which can be contrasted for instance with singleton Shapley value that explicitly takes into account even higher order coalitions that contain the single feature. Thus the interaction distribution has the consequence of making both lower and max-order interactions less faithful to the model.

In contrast, in our Faith-Interaction indices, even lower-order interaction weights take into account all possible coalitions, and where the weights are balanced so that the overall set of interaction indices optimally approximates the behavior of the underlying value function.

5.1 Examples

Example 1: We illustrate the difference between these interaction indices using a function with diminishing marginal utility. Consider the following value function with 11 features:

$$v(S) = \begin{cases} 0 & \text{, if } |S| \leq 1. \\ |S| - p \times \binom{|S|}{2} & \text{, otherwise.} \end{cases} \tag{18}$$

This function represents the payoff when any subset of 11 people work on a task. Each person contributes 1 unit to the overall payoff, and the task requires at least 2 people. However, the marginal utility is diminishing in nature, since any two people also have a probability of p of being non-cooperative. Given this payoff function, it is worth reflecting on what the attributions to individuals should be. While it might seem that zero is a good value since at least two people are needed for the task, this attribution would only correspond to the *marginal contribution* of an individual player i.e. how much a player would contribute when they are by themselves. Whereas we would like our attributions to also take into account

larger coalitions, and marginal contributions to such larger coalitions: this is one of the motivations for considering coalitional game-theoretic indices. Once we do so, then it can be seen that an individual effect of one is much more reasonable. Similarly, we would expect that the pairwise interaction effects be close to $-p$.

In Table 2, we list the values for different interaction indices for $p = 0.1, 0.2$. When the maximum interaction order $\ell = 1$, all indices are similar since their restrictions to singleton are the Shapley/Banzhaf values. When the maximum interaction order $\ell = 2$, our Faith-Shap accurately captures individual contribution and pairwise interaction effects by assigning $0.95/0.95$ and $-0.091/-0.191$ for order 1 and 2 and for $p = 0.1/0.2$ respectively, which are very close to the intuitions we outlined earlier. However, the Shapley Taylor index assigns the individual effect of i by using the marginal $v(\{i\}) - v(\emptyset)$, which can be highly inaccurate since such a marginal contribution does not take into account marginal contributions to larger coalitions.

For $p = 0.1$, Shapley Taylor along with Interaction Shapley assigns a positive/zero value to the interaction effect, which suggests that forming groups has complimentary/no effects. On the contrary, Banzhaf interaction and Faith-Banzhaf give negative values for interaction between players, which correctly reflects the decrease in the marginal utility of this game.

For $p = 0.2$, the Shapley Taylor index is uniformly zero for any order. This highlights the other drawback of the Shapley Taylor index: the impoverished lower-order interaction indices make the max-order indices less faithful to the model. Specifically, for $p = 0.2$, and with $d = 11$ players, we can see that $v([d]) = v(\emptyset) = 0$. We have already seen that $\mathcal{E}^{\text{Taylor}}_{\{i\}}(v, \ell) = 0$, for $i \in [d]$. For $\ell = 2$, we then have that the summation of the max-order (i.e. order two) indices equals $v([d]) - v(\emptyset) - \sum_{i=1}^{d} \mathcal{E}^{\text{Taylor}}_{\{i\}}(v, \ell) = 0$ by the efficiency axiom. Since all max-order indices have the same value by the symmetry axiom, the max-order indices are uniformly zero. In this case, the Shapley Taylor indices do not take into account the function values $v(S)$ with $\ell \leq |S| < d$, and can be arbitrarily unfaithful to these orders. Here, the Banzhaf interaction and Faith-Banzhaf again correctly reflect the negative interaction between players. However, the Banzhaf interaction value gives a value close to 0 for the first-order indices. Taken together with its negative interaction effects, it might seem that coalitions can only be hurtful to the payoff, which is misleading since the total utility is positive when 2 to 10 players are present. On the other hand, our Faith-Banzhaf gives a positive value close to 1 for individual effects of order 1. Taken together with its negative interaction effects, the value given by the Faith-Banzhaf seems more intuitive: every single player contributes to the utility, while each pair of players hurts the utility.

Another instructive viewpoint for interaction values is by inspecting their utility for approximating the overall payoff function. In Figure 1 and 2, we approximate the function $v(S)$ using $\sum_{T \subseteq S, |T| \leq 2} \mathcal{E}_T(v, \ell)$ for different interaction indices. We can see that our Faith-Shap/Faith-Banzhaf are (almost) faithful to all orders except for $|S| = 1$. However, the Shapley Taylor index is only fully faithful to the model when the order is $0, 1, 11$, and curves for other interaction indices are unfaithful.

Example 2: We provide another example, this time with increasing marginal utility. Consider a family who is in the wind energy business, with $d = 11$ family members. Currently, the family owns 1 wind turbine, and they can get 3 units of revenue per wind turbine they own. Now, each family member is considering whether to manage a wind turbine. To build

Indices	$p = 0.1$			$p = 0.2$		
	$\ell = 1$	$\ell = 2$		$\ell = 1$	$\ell = 2$	
	Order 1	Order 1	Order 2	Order 1	Order 1	Order 2
Faith-Shap	0.5	0.95	-0.091	0	0.95	-0.191
Shapley Taylor	0.5	0	0.1	0	0	0
Interaction Shapley	0.5	0.5	0	0	0	-0.1
Banzhaf Interaction	0.51	0.51	-0.113	0.009	0.009	-0.213
Faith-Banzhaf	0.51	1.08	-0.113	0.009	1.08	-0.213

Table 2: Values for different interaction indices of different orders for $p = 0.1, 0.2$ with different maximum interaction orders. Note that the value function is symmetric with respect to players, so we use order 1 and 2 to denote importance scores of any single player and interaction of any two players. Note that $\mathcal{E}_{\emptyset}^{\text{F-Shap}}(v, \ell) = 0$ and $\mathcal{E}_{\emptyset}^{\text{F-Bzf}}(v, \ell) = -0.24$ for both $p = 0.1$ and $p = 0.2$.

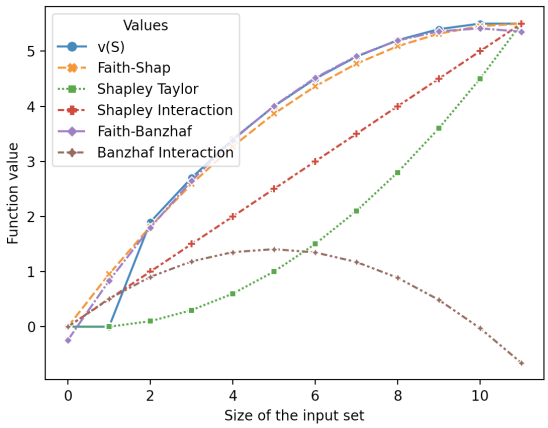


Figure 1: Function approximation of Eqn.(18) using different interaction indices for $p = 0.1$ with the maximum interaction order $\ell = 2$.

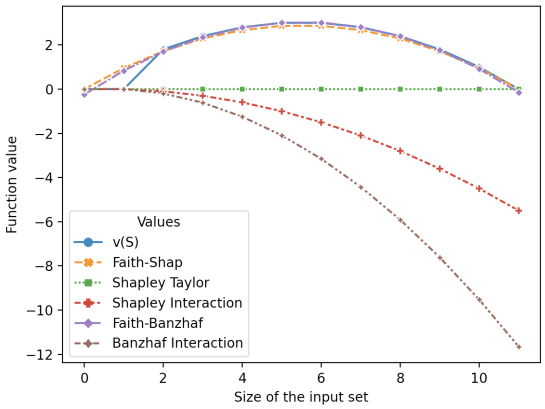


Figure 2: Function approximation of Eqn.(18) using different interaction indices for $p = 0.2$ with the maximum interaction order $\ell = 2$.

x wind turbines, the cost is described by the function $\text{cost}(x) = x + 2\log(x + 1)$, as they may get a discount from the constructor to build more wind turbines at the same time. If exactly one member chooses to manage a wind turbine, the building cost will be 0 since the family already owns one wind turbine. The total revenue for the family when S is the set of members that participate in building new wind turbines can be described by the following function:

$$v(S) = \begin{cases} 0 & \text{, if } |S| = 0. \\ 3 & \text{, if } |S| \leq 1. \\ 3|S| - (|S| - 2\log(|S| + 1)) & \text{, if } 2 \leq |S| \leq 11. \end{cases} \tag{19}$$

This function has an increasing marginal utility since the marginal cost is decreasing. Therefore, we would expect the interaction effect to be positive. However, from Table 3, only Faith-Shap, Faith-Banzhaf and Banzhaf interaction indices capture this effect.

Moreover, the Faith-Shap and Faith-Banzhaf indices have the following intuitive interpretation: Having one more member joining the family business increases the total revenue by 1.20/1.19 unit, with 0.07/0.09 additional unit of revenue when two members join together since they are cooperative. In contrast, we can not interpret the Banzhaf interaction index for orders 1 and 2 jointly since it is not cognizant of the maximum interaction order ℓ.

Indices	$\ell = 1$	$\ell = 2$	
	Order 1	Order 1	Order 2
Faith-Shap	1.55	1.20	0.07
Shapley Taylor	1.55	3	-0.29
Shapley Interaction	1.55	1.55	-0.12
Faith-Banzhaf	1.65	1.19	0.09
Banzhaf Interaction	1.65	1.65	0.09

Table 3: Values for different interaction indices of different orders with the maximum interaction order $\ell = 2$. Note that the value function is symmetric with respect to players, so we use order 1 and 2 to denote importance scores of any single player and interaction of any two players. Note that $\mathcal{E}_{\emptyset}^{\text{F-Shap}}(v, \ell) = 0$ and $\mathcal{E}_{\emptyset}^{\text{F-Bzf}}(v, \ell) = 0.48$ for the indices corresponding to empty sets.

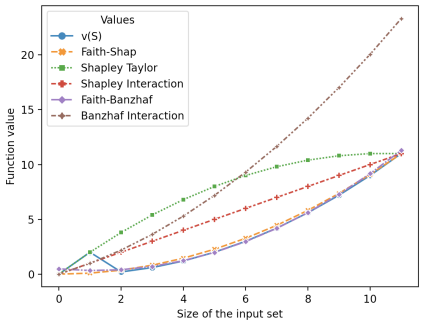


Figure 3: Function approximation of Eqn.(19) using different interaction indices with the maximum interaction order $\ell = 2$.

6. Computation of Faithful Shapley Interaction Index

The exact computation of Faith-Shap indices via Eqn.(16) is intractable in general since it involves computations of $\mathcal{O}(2^d)$ Möbius transforms of different subsets. However, when the value function is lower-order, the Faith-Shap interaction indices can be computed in polynomial time.

Definition 20 *(Order of value functions) A value function v has order $\ell_v \in \mathbb{N}$ if ℓ_v is the smallest integer such that $a(v, R) = 0$ for all $R \subseteq [d]$ with $|R| > \ell_v$.*

From Eqn.(4), this entails the value function with order ℓ_v can be written as $v = \sum_{R \subseteq [d], |R| \leq \ell_v} a(v, R) v_R$. When a game only involves the cooperation of a small number of players, its value function can usually be written as a summation of lower-order basis functions. For instance, a basis function v_R has order 1, example value functions in Section 5.1 have order 2, the value function (Garg et al., 2012) defined as summations of pairwise distances within a coalition used in clustering also have order 2. When $\ell_v = \mathcal{O}(\ell) > \ell$, by using Eqn.(16), the exact computation of Faith-Shap indices of a subset S only requires time complexity $\mathcal{O}(d^{O(\ell)})$ since we only need to consider $\binom{d}{\ell_v - |S|} = \mathcal{O}(d^{O(\ell)})$ Möbius transforms of subsets.

For the computation of the Faith-Shap for general functions, calculating the exact Faith-Shap values requires 2^d model evaluations. We sample each coalition $S \subseteq [d]$ with probability $\propto \frac{d-1}{\binom{d}{|S|}|S|(d-|S|)}$, and solve

$$\underset{\mathcal{E} \subseteq \mathbb{R}^{d_\ell}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \left(v(S) - \sum_{T \subseteq S, |T| \leq \ell} \mathcal{E}_T(v, \ell) \right)^2, \text{s.t.} \sum_{T \subseteq [d], |T| \leq \ell} \mathcal{E}_T(v, \ell) = v([d]), \text{ and } \mathcal{E}_\emptyset(v, \ell) = v(\emptyset). \tag{20}$$

We empirically show the sampling approach provides more accurate estimates with fewer model evaluations in Section 8.1. We defer deriving approximation results for sampling as well as other approximation methods for computing Faith-Shap to future work. There is a rich body of work on developing such approximation results for the first-order Shapley values (see Mitchell et al. (2022); Covert and Lee (2020), which could be extended to the Faith-Shap setting.

7. Algebraic Properties of Faith-Interaction Indices

In the following two sub-sections, we discuss how Faith-Shap can be represented as a cardinal index, as well as through the lens of a multilinear approximation.

7.1 Cardinal Indices

Grabisch and Roubens (1999) show that any interaction index (they only consider the classical case with maximum interaction order $\ell = d$) that satisfies the linearity, dummy, and symmetry axioms necessarily has the following form:

$$\mathcal{E}_S(v, d) = \sum_{T \subseteq [d] \setminus S} p_{|T|}^{|S|} \Delta_S v(T), \quad \forall S \subseteq [d], \tag{21}$$

and for some family of constants $\{p_t^s\}_{s \in [0:d], t \in [0:d-s]}$. They term this class of interaction indices as *cardinal* interaction indices. Of course this is a large class, and it is not apriori clear how to further constrain the indices so as to get specific values for the constants $\{p_t^s\}$. We remark in passing that Shapley and Banzhaf interaction indices impose additional structure on the constants $\{p_t^s\}$.

We can also consider the class of probabilistic interaction indices:

$$\mathcal{E}_S(v, d) = \sum_{T \subseteq [d] \setminus S} p_T^S \Delta_S v(T),$$

where for any $S \subseteq [d]$, the constants $\{p_T^S\}_{T \subseteq [d] \setminus S}$ form a probability distribution on $[d] \setminus S$. We can then define cardinal-probabilistic indices as those indices that are both cardinal and probabilistic interaction indices, so that $p_T^S = p_{|T|}^{|S|}$, for some family of constants $\{p_t^s\}_{s \in [1:d], t \in [0:d-s]}$ that satisfy:

$$\sum_{t=0}^{d-s} \binom{d-s}{t} p_t^s = 1.$$

Fujimoto et al. (2006) shows that indices that satisfy certain additivity, monotonicity, symmetry, and dummy partnership axioms are necessarily cardinal probabilistic indices. As Fujimoto et al. (2006) shows, Shapley and Banzhaf interaction indices do fall into this class.

One could of course extend these notions of cardinal, probabilistic, and cardinal-probabilistic indices to be cognizant of the maximum interaction order $\ell \in [d]$. It is an interesting open question to investigate extensions of results of Fujimoto et al. (2006) to such a sub-class of cardinal-probabilistic indices cognizant of the max-interaction order. In this section, we provide a modest initial result along these lines, focusing on the top interaction level of the interaction index.

Proposition 21 *For any maximum interaction order $1 \leq \ell \leq d$, and for any set value function $v : 2^d \mapsto \mathbb{R}$, the top level of the Faithful Shapley Interaction index can be expressed as a cardinal-probabilistic index:*

$$\mathcal{E}_S^{F\text{-}Shap}(v, \ell) = \sum_{T \subseteq [d] \setminus S} p_{|T|}^{\ell} \Delta_S(v(T)), \quad \forall S \subseteq [d] \text{ with } |S| = \ell, \tag{22}$$

where $p_t^{\ell} = \frac{(2\ell-1)!(\ell+t-1)!(d-t-1)!}{((\ell-1)!)^2(d+\ell-1)!}$. Moreover, it satisfies $\sum_{t=0}^{d-\ell} \binom{d-\ell}{t} p_t^{\ell} = 1$.

Therefore, the top level of the Faithful Shapley Interaction index captures the interactions of features in S *in the presence of all subsets* $T \subseteq [d] \setminus S$.

7.2 Multilinear Formulation

Any set value function $v : 2^{[d]} \mapsto \mathbb{R}$ has a unique multi-linear extension $g : [0,1]^d \mapsto \mathbb{R}$, also referred to the *Owen multilinear extension* (Owen, 1972), given as:

$$g(x) := \sum_{T \subseteq [d]} v(T) \prod_{i \in T} x_i \prod_{i \notin T} (1 - x_i), \quad \forall x \in [0,1]^d.$$

For any set $S \subseteq [d]$, with $S = \{i_1, \ldots, i_s\}$, denote its S-derivative as $\Delta_S g(x) := \frac{\partial^s g(x)}{\partial x_{i_1} \ldots \partial x_{i_s}}$.

7.2.1 PATH INTEGRALS

Grabisch et al. (2000) show that Shapley interaction index can be written as:

$$\mathcal{E}_S^{\text{Shap}}(v, d) = \int_{x=0}^{1} \Delta_S g(x, \ldots, x) dx, \ \forall S \subseteq [d].$$

That is, we can obtain the Shapley interaction index by integrating the S-derivative along the diagonal of the unit hypercube.

On the other hand, the Banzhaf interaction index can be written as:

$$\mathcal{E}_S^{\text{Bzf}}(v, d) = \int_{x \in [0,1]^d} \Delta_S g(x) dx, \ \forall S \in \mathcal{S}_d.$$

That is, we can obtain the Banzhaf interaction index by integrating the S-derivative over the entire unit hypercube. In this case, it also has the closed form: $\Delta_S g(1/2, \ldots, 1/2)$.

Fujimoto et al. (2006) show that any cardinal probabilistic index $\mathcal{E}$ has the form:

$$\mathcal{E}_S(v,d) = \int_{x=0}^{1} \Delta_S g(x,\ldots,x) dF_{|S|}(x),\ \forall S \in \mathcal{S}_d,$$

for some family of CDFs $\{F_s\}_{s\in[d]}$. That is, we can obtain any cardinal probabilistic index by integrating the S-derivative along the diagonal of the unit hypercube with respect to some distribution over $[0,1]$.

It is an interesting open question whether we could extend these results from Grabisch et al. (2000) and Fujimoto et al. (2006) to interaction indices that are cognizant of the maximum interaction order $\ell \in [d]$. In this section, we provide a modest initial result along these lines, focusing on the top interaction level of the interaction index.

Proposition 22 *For any maximum interaction order $1 \leq \ell \leq d$, and for any set function $v : 2^d \mapsto \mathbb{R}$, the top level of the Faithful Shapley Interaction index value can be expressed as:*

$$\mathcal{E}_S^{\textit{F-Shap}}(v,\ell) = \int_{x=0}^{1} \Delta_S g(x,\cdots,x) dI_x(\ell,\ell), \quad \forall S \in \mathcal{S}_\ell \text{ with } |S| = \ell, \tag{23}$$

where $I_x(\ell,\ell)$ is cumulative distribution function of the beta distribution $B(x;\ell,\ell)$.

7.2.2 Taylor Expansion

In contrast to path integrals, Sundararajan et al. (2020) use the Taylor expansion of $g(\mathbf{1}) = v([d])$ around $g(\mathbf{0}) = v(\emptyset)$ Taylor derivations to derive their interaction index. Specifically, they show that Shapley Taylor index $\mathcal{E}_S^{\text{Taylor}}(v,\ell)$ is equal to the $|S|^{\text{th}}$ term of the $(\ell-1)^{\text{th}}$ order Taylor expansion of $g(\cdot)$ with Lagrange remainder:

$$\begin{aligned}
g(\mathbf{1}) &= \sum_{j=0}^{\ell-1} \frac{g^{(j)}(\mathbf{0})}{j!} g(\mathbf{0}) + \int_{x=0}^{1} \frac{(1-x)^{\ell-1}}{(\ell-1)!} g^{(\ell)}(x,\cdots,x) dx \\
&= \sum_{j=0}^{\ell-1} \sum_{|S|=j} \Delta_S g(\mathbf{0}) + \sum_{|S|=\ell} \int_{x=0}^{1} \ell(1-x)^{\ell-1} \Delta_S g(x,\cdots,x) dx \\
&\quad \text{(Sundararajan et al., 2020, Theorem 3)} \\
&= \sum_{j=0}^{\ell-1} \sum_{|S|=j} \mathcal{E}_S^{\text{Taylor}}(v,\ell) + \sum_{|S|=\ell} \mathcal{E}_S^{\text{Taylor}}(v,\ell),
\end{aligned}$$

where $g^{(j)}(x)$ is the j^{th} derivative of the function $g(x,\cdots,x)$, $\mathcal{E}_S^{\text{Taylor}}(v,\ell) = \Delta_S g(\mathbf{0})$ for $|S| < \ell$ and $\mathcal{E}_S^{\text{Taylor}}(v,\ell) = \int_{x=0}^{1} \ell(1-x)^{\ell-1} \Delta_S g(x,\cdots,x) dx$ with $|S| = \ell$. This can be seen to result in impoverished lower-order subset interactions, which now no longer take into account higher-order coalitions that include that subset.

7.2.3 Pseudo-Boolean Function Approximation

While we have so far discussed the continuous multi-linear extension of a set value function $v : 2^{[d]} \mapsto \mathbb{R}$, we can also simply consider its equivalent pseudo-Boolean counterpart $g \in \mathcal{F}$

with $\mathcal{F} = \{g : \{0,1\}^d \mapsto \mathbb{R}\}$:

$$g(x) := \sum_{T \subseteq [d]} v(T) \prod_{i \in T} x_i \prod_{i \notin T} (1 - x_i), \quad \forall x \in \{0,1\}^d.$$

One can also derive the pseudo-Boolean function $g_{\mathcal{E}}$ corresponding to interaction indices $\mathcal{E}$, and ask for interaction indices with pseudo-Boolean counterparts $g_{\mathcal{E}}$ that best approximate the pseudo-Boolean counterpart g of the set value function. Specifically, given a maximum interaction order $\ell \in [d]$ and an interaction index $\mathcal{E} \in \mathbb{R}^{d_\ell}$, its pseudo-Boolean counterpart $g_{\mathcal{E}} \in \mathcal{F}$ is defined as:

$$g_{\mathcal{E}}(x) := \sum_{T \subseteq [d], |T| \leq \ell} \mathcal{E}_T(v, \ell) \prod_{i \in T} x_i, \quad \forall x \in \{0,1\}^d.$$

Hammer and Holzman (1992) and Grabisch et al. (2000) consider solving for the best ℓ_2-norm approximation by the function $g_{\mathcal{E}}(\cdot)$ with degree up to ℓ. That is, $\|g - g_{\mathcal{E}}\|_2 = \sqrt{\sum_{x \in \{0,1\}^d} (g(x) - g_{\mathcal{E}}(x))^2}$. Using this perspective, we can see that Faith-Banzhaf interaction indices can in turn be related to such a function approximation:

$$\mathcal{E}^{\text{F-Bzf}}(v, \ell) = \min_{\mathcal{E} \in \mathbb{R}^{d_\ell}} \|g(x) - g_{\mathcal{E}}(x)\|_2 = \min_{\mathcal{E} \in \mathbb{R}^{d_\ell}} \sum_{S \subseteq [d]} \left(v(S) - \sum_{T \subseteq S, |T| \leq \ell} \mathcal{E}_T \right)^2,$$

and where the solution has the closed-form expression we detail in Theorem 16.

For the singleton attribution case, with max order $\ell = 1$, Ding et al. (2008) and Ruiz et al. (1998) consider μ-norm function approximations $\|g(x) - g_{\mathcal{E}}(x)\|_\mu = \sqrt{\sum_{x \in \{0,1\}^d} \mu(x)(g(x) - g_{\mathcal{E}}(x))^2}$, but where μ only depends on $\|x\|_1$, and where $\mu(\mathbf{0})$ and $\mu(\mathbf{1})$ can both be infinity. Ding et al. (2008) provide a closed-form expression for $g_{\mathcal{E}}(x)$, while Ruiz et al. (1998) analyze its axiomatic properties.

For the specific case where the probability of coalition S can be expressed as $\mu(x) = \prod_{i:x_i=1} p_i \prod_{j:x_j=0} (1 - p_j)$ for some $0 < p_i < 1$ indicating the probability of the feature i being present, Ding et al. (2010) and Marichal and Mathonet (2011) considers solving the best ℓ^{th} order polynomial approximation under $\|\cdot\|_\mu$ norm.

In contrast to the above work, our developments could be cast as pseudo-Boolean approximations for the general weighted norm case $\|\cdot\|_\mu$, for general weighting functions $\mu(\cdot)$ without stringent structural assumptions, and while allowing for arbitrary maximum interaction orders $\ell \in [d]$.

8. Experiments

We first provide some experiments validating the relative computational efficiency of computing our Faith-Interaction indices, followed by quantitative and qualitative demonstrations of their use as explanations of ML models over a language dataset.

The language dataset we use throughout the experiment is the simplified IMDB (Maas et al., 2011) dataset, where the model only uses the first two sentences of movie reviews as input, and predicts the probability of the reviews being positive. The model being explained is a BERT language model (Devlin et al., 2018) with 0.82 accuracy on the test set.

8.1 Computational Efficiency

Exact computation of interaction indices that aggregate over all possible feature subsets exactly typically requires 2^d model evaluations (with d features) which is impractical in most machine learning applications. A key advantage of our Faith-Interaction indices, as compared to other recently proposed interaction indices such as the Shapley Taylor index and Shapley Interaction index, is that they can be computed by solving a weighted least squares problem. As we empirically show in this section, this enables us to provide more accurate estimates with fewer model evaluations, compared to the other recent approaches that employ permutation-based sampling methods.

Setup: To demonstrate the computational efficiency of Faith-Interaction indices, we compare our proposed Faith-Shap with Shapley interaction and Shapley Taylor interaction indices using different estimation methods. For the Faith-Shap interaction index, We use Eqn.(20) and solve the corresponding linear regression problem with ℓ_1 regularization, and regularization parameter $\alpha = 10^{-3}$ and $\alpha = 10^{-6}$ for the simplified IMDB dataset and the bank dataset. For the Shapley Taylor interaction and Shapley Interaction indices, we use the permutation-based sampling methods (see exact algorithms in the Appendix B).

Methods	Simplified IMDB	Bank marketing
Shapley Taylor	2781.2	7368.3
Shapley Interaction	3960.4	10421.7
Faith-Shap	**887.4**	**893.7**

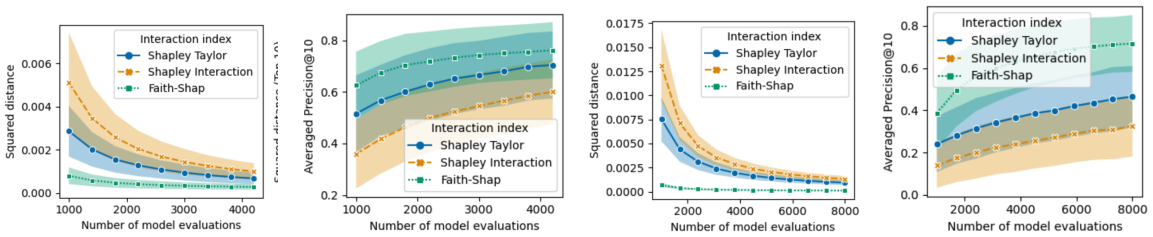
Table 4: Run-time comparison of different Shapley values. Each value represents the number of evaluations required to achieve averaged squared distance less than 10^{-3}.

We compare these indices in two datasets: (1) language dataset: we randomly choose 50 samples with $d = 15$ words from the test set of simplified IMDB and set $\ell = 2$. We treat each word in each text sentence as an input feature and set the baseline to empty text so that we simply remove a word if it does not appear in a coalition. We use BERT prediction scores as outputs of the value function. (2) Portuguese marketing dataset (Moro et al., 2014): this is a tabular dataset with $d = 17$ features. We train an xgboost model (Chen and Guestrin, 2016) with 90.5% accuracy on the test set. We also compare them on synthetic sparse functions in the Appendix.

To measure how close an interaction index is to its ground-truth value, we use two evaluation metrics: (1) averaged squared distance of all top indices, $\left\|\mathcal{E} - \mathcal{E}^{\text{est}}\right\|_2^2 / \binom{d}{\ell}$, and (2) precision at 10, which we measure the proportion of top-10 feature interactions (with respect to absolute value) in the top-10 ground-truth interactions as top interactions are more critical when these indices are used in XAI. (3) We also report run-time, as measured by the number of model evaluations required to achieve averaged squared distance to be smaller than 10^{-3}.

We also note that we drop the lower-order indices and only compare top-order indices (order=ℓ) since computing lower-order Shapley Taylor indices are trivial. We sampled all 2^d different coalitions to compute the ground truth of each index. Each evaluation metric is reported by averaging 50 different inputs with 20 different random seeds.

Results: From Figure 4 and Table 4, we see that Faith-Shap can be estimated more accurately and uses fewer model evaluations: in both language data and sparse settings, as well as in terms of all evaluation metrics.



(a) Averaged squared distance (language). (b) Precision@10 (language). (c) Average squared distance (bank). (d) Precision@10 (bank).

Figure 4: Comparison of Faith-Shap, Shapley Taylor and Shapley interaction indices in terms of computational efficiency in language data and synthetic sparse functions. The shaded area indicates the total area between one standard deviation above and below.

8.2 Explanations on a Language Dataset

In this section, we use our Faith-Shap interaction index to explain the BERT model on the simplified IMDB dataset. The experimental setting is the same as described in Section 8.1 except that we set the maximum interaction order $\ell = 2$, the regularization parameter $\alpha = 10^{-3}$ for Lasso, and sample 4000 coalitions for each text in the simplified IMDB. Table 5 shows some of the interesting interactions we found.

Index	Sentences (bold words are the interactions with the highest (absolute) importance values)	Model Prediction	Interaction score
1	I have **Never forgot** this movie. All these years and it has remained in my life.	Positive	0.818
2	TWINS EFFECT is a poor film in so many respects. The **only good** element is that it doesn't take itself seriously..	Negative	-0.375
3	I rented this movie to get an easy, entertained view of the history of Texas. I got a **headache instead**.	Negative	0.396
4	Truly **appalling waste** of space. Me and my friend tried to watch this film to its conclusion but had to switch it off about 30 minutes from the end.	Negative	0.357
5	I still remember watching Satya for the first time. I was completely **blown away**.	Positive	0.283

Table 5: Top interactions of different examples on IMDB. See more results in Appendix B.

In the first two examples, we see non-complementary interaction effects. In the first example, while the importance values of the individual words "Never" and "forgot" are negative (as shown in Tables 8, 10 in the Appendix), their joint effect as shown in the table here is

extremely positive. Similarly, for the second, the words "only" and "good" are individually positive, while their joint effect is strongly negative. The fourth and fifth examples show more subtle non-complementarity effects. In the fourth example, while the individual words "headache" and "instead" have negative importance scores, their joint effect is positive, since the total effect of the phrase is less than the sum of the individual importance of these two words. The last example shows the effect of complementarity: words in a phrase are only meaningful when all words are present, and hence have a positive interaction effect.

In Appendix B, we further show the top-15 important interactions and compare them to those from Shapley Taylor index, Shapley interaction index, Integrated Hessian (Janizek et al., 2021) and Archipelago (Michael et al., 2020).

We find that although Faith-Shap, Shapley Taylor index, and Shapley interaction index capture similar feature interactions, the later two methods are not able to find meaningful singleton features. The reasons are (1) the first-order terms of the Shapley Taylor indices are trivial, which is the difference between predicted probabilities of a sentence containing only one word and an empty sentence (a baseline) (2) importance scores for the first order Shapley value and Shapley interaction index are not comparable since the Shapley interaction index does not satisfy the efficiency axiom. For integrated Hessian, we empirically find that the BERT model assigns higher values for self-interactions and punctuation marks.

9. Related work

Related work in cooperative theory: in cooperative game theory, a set function $v(\cdot)$ with $v(\emptyset) = 0$ corresponds to a transferable utility game (TU-game), and a set function with order $\leq \ell$ is called an ℓ-additive TU-game (Grabisch et al., 2016). Therefore, our approach can be viewed as a least squares approximation of a TU-game by an ℓ-additive TU-game; see for instance Eqn. (10). Variants and special cases of this least squares approximation problem have been studied in the cooperative game theory field. For $\ell = 1$, Charnes et al. (1988) first give general solutions when the weighting function is symmetric and positive, and show that the Shapley value results from a particular choice of the weighting function. Ruiz et al. (1996, 1998) consider the same setting, and study the axiomatic properties of the solutions of the least squares problems. Ding et al. (2008) further generalizes the previous results by considering the cases where some weights are allowed to be zero. For the case where maximum interaction order $\ell > 1$, Hammer and Holzman (1992) and Grabisch et al. (2000) solve the least squares problem when the weighting function is a constant, and show that the top-level coefficients coincide with those of the Banzhaf interaction indices of order ℓ. Ding et al. (2010) and Marichal and Mathonet (2011) consider a certain weighted version of the problem, and propose weighted Banzhaf interaction indices. Grabisch and Rusinowska (2020) consider the approximation problem under the constraints that both TU-games yield the same Shapley value. Marichal and Roubens (1999) extend the Shapley value and propose the chaining interaction index whose definition is based on maximal chains of ordered sets. For more details on this line of work, see the recent book (Grabisch et al., 2016). From the lens of TU-game approximation, our work could be viewed as allowing for general weighting functions $\mu(\cdot)$ without stringent structural assumptions, as well as arbitrary maximum interaction orders $\ell \in [d]$.

While the Shapley value focuses on a fair allocation among players, there exist other solution concepts in cooperative game theory that have different purposes. For example, core (Gillies, 1953) allocates the total payoff in a stable manner, nucleolus (Schmeidler, 1969) is a solution lying in the core with unique axiomatic properties, and the Nash bargaining solution (Nash Jr, 1950) focuses on two-player bargaining problems. Extending these concepts to interaction contexts may lead to different solutions with different properties, and are interesting topics for future work.

Feature attribution in XAI: When Faith-Shap is used in XAI, it can be seen as a local and post-hoc approach that extracts singleton features and feature interaction importances for a given prediction. It can be viewed as an order ℓ polynomial model, with desired axiomatic properties, that explains how a black-box model behaves locally. Explaining complex models with an interpretable *local surrogate model* has been substantially studied in XAI. LIME (Ribeiro et al., 2016) use a local linear model to describe a prediction made by the model being explained. Model Agnostic Supervised Local Explanations (MAPLE) (Plumb et al., 2018) utilize local linear modeling and dual interpretation of random forests. AnchorLIME (Ribeiro et al., 2018) uses IF-THEN rules to generate explanations. Model Understanding through Subspace Explanations (MUSE) (Lakkaraju et al., 2019) explains how the model behaves in subspaces characterized by certain features of interest. Kernel SHAP (Lundberg and Lee, 2017) can be viewed as first-order Faith-Shap. These approaches assign credits to each individual feature based on how much it influences the models' prediction and do not aim to explain how feature interactions affect the model.

Feature interactions in XAI: Feature interactions have also been investigated in the machine learning community. Tsang et al. (2017) detect feature interactions by examining weight matrices of DNNs. Tsang et al. (2018) disentangle complex feature interactions within DNNs by forcing the weights matrices to be block-diagonal. Singh et al. (2018) build hierarchical explanations within a feed-forward neural network using hierarchical clustering of features. Cui et al. (2019) and Janizek et al. (2021) explain pairwise interactions in neural networks, and Bayesian neural networks respectively via second-order derivatives. Lundberg et al. (2020) quantify feature interactions in tree-based models using the Shapley interaction index. Tsang et al. (2020) proposes Archipelago, which quantifies the interaction within a feature group S via the marginal importance $v(S) - v(\emptyset)$.

While these approaches have taken significant steps towards understanding feature interactions, they are limited to a certain kind of model architecture. Tsang et al. (2017) and Tsang et al. (2018) can only be applied to feed-forward neural network architectures, but not LSTMs and CNNs. While Singh et al. (2018) can be applied to LSTMs and CNNs, it is unclear how to apply it to recent innovations such as transformers. The approach of Cui et al. (2019) can only be applied to Bayesian neural networks, and Janizek et al. (2021) can only be applied to models where second-order derivatives exist everywhere. Lundberg et al. (2020) only study tree-based models. While Archipelago (Tsang et al., 2020) is a post-hoc explanation approach that can be applied to any model, Archipelago measures the importance of a feature group as a whole, while Faith-Shap measures the marginal effects of interaction among feature groups. Also, the Archipelago does not obey the dummy axiom and satisfies the efficiency axiom only for certain kinds of functions.

10. Conclusion

Deriving unique interaction indices that satisfy the interaction extensions of the individual Shapley axioms has been a long-standing open problem. Existing approaches introduce additional less natural axioms, with some even sacrificing natural ones such as efficiency, in order to specify unique interaction indices. In this work, we take the alternate route of considering the family of what we term faithful interaction indices, which similar to individual Shapley values, aim to approximate the given set value function for all feature subsets. We show that when restricting to the class of faithful interaction indices, we obtain a unique interaction index that satisfies the interaction extensions of the individual Shapley axioms, which we term the Faithful Shapley Interaction Index (Faith-Shap). We show the benefits of the faithful Shapley interaction index via specific games of interest where there is diminishing return and increasing return and connect the Faith-Shap to cardinal probabilistic indices and multilinear approximations. Finally, we show that Faith-Shap is efficient to estimate thanks to its connection to weighted linear regression in sparse settings, and provide some qualitative results for their use as explanations of machine learning models on a real language dataset.

Acknowledgments

The authors would like to thank Michel Grabisch for his generous feedback and thank Hung-Hsun Yu for providing assistance in deriving the closed-form solution of Faith-Shap. The authors also acknowledge the support of NSF via IIS-1909816 and IIS-1955532.

References

John F Banzhaf III. Weighted voting doesn't work: A mathematical analysis. *Rutgers L. Rev.*, 19:317, 1964.

A Charnes, B Golany, M Keane, and J Rousseau. Extremal principle solutions of games in characteristic function form: core, chebychev and shapley value generalizations. In *Econometrics of planning and efficiency*, pages 123–133. Springer, 1988.

Hugh Chen, Joseph D Janizek, Scott Lundberg, and Su-In Lee. True to the model or true to the data? *arXiv preprint arXiv:2006.16234*, 2020.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

Ian Covert and Su-In Lee. Improving kernelshap: Practical shapley value estimation via linear regression. *arXiv preprint arXiv:2012.01536*, 2020.

Tianyu Cui, Pekka Marttinen, and Samuel Kaski. Learning global pairwise interactions with bayesian neural networks. *arXiv preprint arXiv:1901.08361*, 2019.

Anupam Datta, Shayak Sen, and Yair Zick . Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 598–617. IEEE, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Guoli Ding, Robert F Lax, Jianhua Chen, and Peter P Chen. Formulas for approximating pseudo-boolean random variables. *Discrete Applied Mathematics*, 156(10):1581–1597, 2008.

Guoli Ding, Robert F Lax, Jianhua Chen, Peter P Chen, and Brian D Marx. Transforms of pseudo-boolean random variables. *Discrete Applied Mathematics*, 158(1):13–24, 2010.

Christopher Frye, Damien de Mijolla, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley-based explainability on the data manifold. *arXiv preprint arXiv:2006.01272*, 2020.

Katsushige Fujimoto, Ivan Kojadinovic, and Jean-Luc Marichal. Axiomatic characterizations of probabilistic and cardinal-probabilistic interaction indices. *Games and Economic Behavior*, 55(1):72–99, 2006.

Vikas K Garg, Y Narahari, and M Narasimha Murty. Novel biobjective clustering (bigc) based on cooperative game theory. *IEEE Transactions on Knowledge and Data Engineering*, 25(5):1070–1082, 2012.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pages 2242–2251. PMLR, 2019.

Donald B Gillies. Some theorems on n-person games. princeton university. *Unpublished doctoral dissertation.)[aAMC]*, 1953.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999.

Michel Grabisch and Agnieszka Rusinowska. k-additive upper approximation of tu-games. *Operations Research Letters*, 48(4):487–492, 2020.

Michel Grabisch, Jean-Luc Marichal, and Marc Roubens. Equivalent representations of set functions. *Mathematics of Operations Research*, 25(2):157–178, 2000.

Michel Grabisch et al. *Set functions, games and capacities in decision making*, volume 46. Springer, 2016.

Ulrike Grömping. Estimators of relative importance in linear regression based on variance decomposition. *The American Statistician*, 61(2):139–147, 2007.

Peter L Hammer and Ron Holzman. Approximations of pseudo-boolean functions; applications to game theory. *Zeitschrift für Operations Research*, 36(1):3–21, 1992.

John C Harsanyi. A simplified bargaining model for the n-person cooperative game. *International Economic Review*, 4(2):194–220, 1963.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104): 1–54, 2021.

Ruoxi Jia, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gürel, Bo Li, Ce Zhang, Dawn Song, and Costas J Spanos. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1167–1176. PMLR, 2019.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 131–138, 2019.

Richard Harold Lindeman. *Introduction to bivariate and multivariate analysis*. Scott Foresman & Co, 1980.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1): 56–67, 2020.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics, 2011.

Jean-Luc Marichal and Pierre Mathonet. Weighted banzhaf power and interaction indexes through weighted approximations of games. *European journal of operational research*, 211 (2):352–358, 2011.

Jean-Luc Marichal and Marc Roubens. The chaining interaction index among players in cooperative games. In *Advances in Decision Analysis*, pages 69–85. Springer, 1999.

Tsang Michael, Cheng Dehua, Liu Hanpeng, Feng Xue, Zhou Eric, and Yan Liu. Extracting and leveraging feature interaction interpretations. In *International Conference on Learning Representations*, 2020.

Rory Mitchell, Joshua Cooper, Eibe Frank, and Geoffrey Holmes. Sampling permutations for shapley value estimation. *Journal of Machine Learning Research*, 23(43):1–46, 2022.

Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

John F Nash Jr. The bargaining problem. *Econometrica: Journal of the econometric society*, pages 155–162, 1950.

Art B Owen. Sobol'indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):245–251, 2014.

Art B Owen and Clémentine Prieur. On shapley value for measuring importance of dependent inputs. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):986–1002, 2017.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5-part-2):64–79, 1972.

Gregory Plumb, Denali Molitor, and Ameet S Talwalkar. Model agnostic supervised local explanations. In *Advances in Neural Information Processing Systems*, pages 2515–2524, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Luis M Ruiz, Federico Valenciano, and Jose M Zarzuelo. The least square prenucleolus and the least square nucleolus. two values for tu games based on the excess vector. *International Journal of Game Theory*, 25(1):113–134, 1996.

Luis M Ruiz, Federico Valenciano, and Jose M Zarzuelo. The family of least square values for transferable utility games. *Games and Economic Behavior*, 24(1-2):109–130, 1998.

David Schmeidler. The nucleolus of a characteristic function game. *SIAM Journal on applied mathematics*, 17(6):1163–1170, 1969.

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

Chandan Singh, W James Murdoch, and Bin Yu. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*, 2018.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. *arXiv preprint arXiv:1908.08474*, 2019.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. The shapley taylor interaction index. In *International Conference on Machine Learning*, pages 9259–9268. PMLR, 2020.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. *arXiv preprint arXiv:1705.04977*, 2017.

Michael Tsang, Hanpeng Liu, Sanjay Purushotham, Pavankumar Murali, and Yan Liu. Neural interaction transparency (nit): Disentangling learned interactions for improved interpretability. *Advances in Neural Information Processing Systems*, 31:5804–5813, 2018.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. How does this interaction affect me? interpretable attribution for feature interactions. *arXiv preprint arXiv:2006.10965*, 2020.

Xi Ye, Rohan Nair, and Greg Durrett. Connecting attributions and qa model behavior on realistic counterfactuals. *arXiv preprint arXiv:2104.04515*, 2021.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Appendix A. Organization

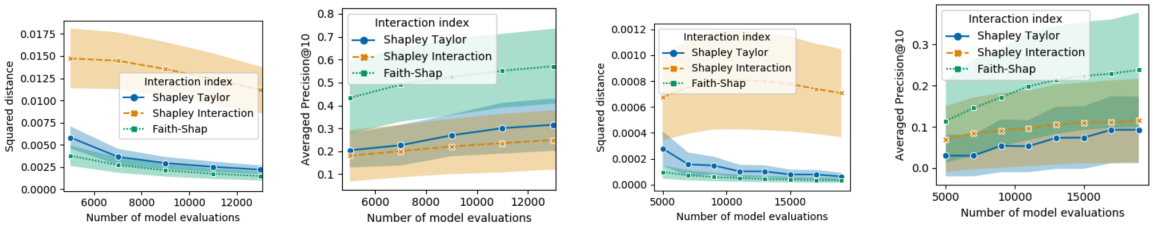
The Appendices contain additional technical content and are organized as follows: In Appendix B, we provide details for sampling algorithms for different indices and supplementary results for different setups for the computational efficiency experiment in Section 8.1. In Appendix C, we give experimental details and show the detailed results of Faithful Shapley Interaction value and Shapley Taylor indices. In Appendix D, we provide additional guidance on Theorem, where we clarify how to choose the parameters a, b to design Faith-Interaction indices. In Appendix E, we provide auxiliary theoretical results of the Faith-Interaction indices, which will be subsequently used in our proof of main theorems. We **leave proof of propositions and theorems to arxiv** (https://arxiv.org/abs/2203.00870).

Appendix B. Experimental Details and Supplementary Results of Computational Efficiency

In this section, we provide implementation details of the sampling algorithms for different indices as well as supplementary experimental results for computational efficiency experiments.

The sampling algorithms for the Shapley Taylor and Shapley interaction indices are shown in Algorithm 1 and 2. These algorithms are based on the fact that these two indices are the expected value of discrete derivatives over different ordering processes (Sundararajan et al., 2020, Section 2.2). These algorithms are more efficient since they may use $v(S)$ of the same coalition S to compute indices of different subsets. Also, to measure the run-time of each index, we measure its average squared distance every 200/300 model evaluation.

We also measure computation efficiency on the synthetic sparse functions, which is constructed as follows: we parameterize the synthetic sparse function $v : \{0,1\}^d \to \mathbb{R}$ with $\sum_{i=1}^{N} a_i \prod_{j \in S_i} x_j$, where $x = \{x_1, \cdots, x_d\} \in \{0,1\}^d$ are the input of the value function, $S_1, S_2, \cdots, S_N$ are subsets of $[d]$ and $a_1, \cdots, a_N$ are coefficients. We set $d = 70$, $N = 30$, $\ell = 2$ and $d = 90$, $N = 10$, $\ell = 2$, sample each a_i uniformly over $[-\frac{i}{10}, \frac{i}{10}]$. Each S_i is uniformly sampled over subsets of $|S|$ with sizes ≤ 5 and ≤ 10, respectively. We use Eqn.(16) to compute the ground truth of interaction indices for sparse synthetic functions. The results are shown in Figure 5. We can see that Faith-Shap is more efficient in terms of all evaluation metrics.



(a) Averaged squared distance for $d = 70$, $\ell = 2, N = 30, |S| \leq 5$.

(b) Precision@10 for $d = 70$, $\ell = 2, N = 30, |S| \leq 5$.

(c) Averaged squared distance for $d = 90$, $\ell = 2, N = 20, |S| \leq 10$.

(d) Precision@10 for $d = 90$, $\ell = 2, N = 20, |S| \leq 10$.

Figure 5: Comparison of Faith-Shap, Shapley Taylor and Shapley interaction indices in terms of computational efficiency on synthetic sparse functions for $d = 70$, $\ell = 2, N = 30, |S| \leq 5$, and $d = 90$, $\ell = 2, N = 20, |S| \leq 10$.

Algorithm 1: Permutation-based sampling algorithm for the top-order Shapley Taylor index

input : a value function $v : 2^d \mapsto \mathbb{R}$, maximum order ℓ.
begin
  sum$[S] \leftarrow 0$ for all sets $S \subseteq [d]$ with size ℓ.
  count$[S] \leftarrow 0$ for all sets $S \subseteq [d]$ with size ℓ.
  for $t = 1, 2, \ldots$ **do**
    $\pi \leftarrow \{i_1, \cdots, i_d\}$ be a random ordering of $\{1, 2, \cdots, d\}$.
    for *all set* $S \subseteq [d]$ *with size* ℓ **do**
      $i_k \leftarrow$ the leftmost element of S in the ordering π.
      $T \leftarrow \{i_1, \cdots, i_{k-1}\}$ the set of predecessors of i_k in π.
      sum$[S] \leftarrow$ sum$[S] + \Delta_S(v(T))$.
      count$[S] =$ count$[S] + 1$.
    end
  end
  indices$[S] \leftarrow$ sum$[S]$/count$[S]$ for all sets $S \subseteq [d]$ with size ℓ.
  return indices
end

Appendix C. Experimental details for Language Dataset

For the dataset, the Internet Movie Review Dataset (IMDb) (Maas et al., 2011) consists of 50,000 binary labeled movie reviews. Each review is annotated as a positive or negative review. We used 25,000 reviews for training and 25,000 reviews for evaluation.

Here, the set function $v(x)$ represents the predicted probability of input texts being positive sentiment, which is between 0 and 1. We remove a word in a text sequence if the corresponding entry of the word in a binary perturbation variable x is 0. we use 4000 samples to estimate both Faithful Shapley Interaction indices and Shapley Taylor indices. We use Lasso with regularization parameter $\alpha = 0.001$ to estimate Faithful Shapley Interac-

Algorithm 2: Permutation-based sampling algorithm for the Shapley Interaction index.

input : a value function $v : 2^d \mapsto \mathbb{R}$, maximum order ℓ.
begin
  sum$[S] \leftarrow 0$ for all sets $S \subseteq [d]$ with size ℓ.
  count$[S] \leftarrow 0$ for all sets $S \subseteq [d]$ with size ℓ.
  for $t = 1, 2, ...$ **do**
    $\pi \leftarrow \{i_1, \cdots, i_d\}$ be a random ordering of $\{1, 2, \cdots, d\}$.
    for $k = 1, \cdots, d - \ell + 1$ **do**
      $S \leftarrow \{i_k, \cdots, i_{k+\ell-1}\}$.
      $T \leftarrow \{i_1, \cdots, i_{k-1}\}$ the set of predecessors of i_k in π.
      sum$[S] \leftarrow$ sum$[S] + \Delta_S(v(T))$.
      count$[S] =$ count$[S] + 1$.
    end
  end
  indices$[S] \leftarrow$ sum$[S]$/count$[S]$ for all sets $S \subseteq [d]$ with size ℓ.
  return indices
end

tion indices and permutation-based sampling method to estimate the highest order Shapley Taylor indices ($\ell = 2$).

For comparison with other feature interactions methods in XAI, we provide top-15 important features (interactions) for Faith-Shap, Shapley Taylor interaction indices, Shapley Interaction indices, Integrated Hessian (Janizek et al., 2021), and Archipelago (Michael et al., 2020) in Table 6 to 10. We note that Archipelago is run with the number of interactions $k = 3$, and its usage is slightly different than our methods: it constructs a feature hierarchy as an explanation rather than measuring the importance scores of feature interactions.

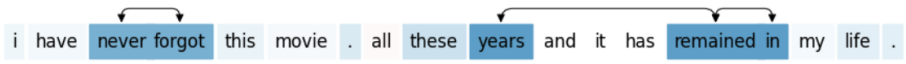
Index	Sentences			Predicted Prob.
1	I have Never forgot this movie. All these years and it has remained in my life.			0.992
	Faithful Shapley indices		*Shapley Taylor indices*	
	Feature (interactions)	Scores	Feature (interactions)	Scores
	Never, forgot	0.818	Never, forgot	1.077
	life	0.383	Never, life	-0.211
	forgot	-0.254	remained, movie	-0.177
	and	0.168	Never, this	-0.160
	it	0.168	forgot, life	-0.149
	Never	-0.163	and, forgot	-0.149
	years	0.156	in, life	-0.143
	All	0.132	Never, it	-0.122
	my	0.126	Never, movie	-0.114
	has	0.120	have, Never	-0.110
	have	0.112	I, have	0.106
	Never, life	-0.106	forgot, in	-0.105
	forgot, it	-0.096	Never, All	-0.104
	my, life	-0.086	years, life	-0.101
	this	0.081	it, forgot	-0.101
	Shapley interaction indices		*Integrated Hessian*	
	Never, forgot	1.166	this, this	1.796
	these, life	-0.194	in, in	-1.406
	have, Never	-0.164	my, my	-1.357
	Never, it	-0.154	., .	1.123
	it, forgot	-0.149	have, this	1.050
	forgot, life	-0.148	movie, movie	1.038
	forgot, remained	0.146	it, it	-0.972
	have, forgot	-0.139	never, this	0.820
	and, Never	-0.136	in, my	0.792
	it, life	-0.135	in, .	0.719
	and, it	0.131	., .	-0.602
	forgot, in	-0.129	this, in	0.554
	I, it	-0.125	remained, remained	-0.532
	years, Never	-0.121	this, life	0.526
	forgot, my	-0.119	remained, my	0.526
	Archipalego			
				

Table 6: Top-15 important feature (interactions) for different methods for language dataset. The predicted probability is the out probability of the sentence having positive sentiment. The archipalego algorithm is run with $k = 3$ interactions.

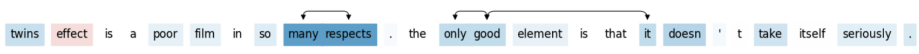
Index	Sentences		Predicted Prob.	
2	TWINS EFFECT is a poor film in so many respects. The only good element is that it doesn't take itself seriously.		0.012	
	Faithful Shapley indices		*Shapley Taylor indices*	
	Feature (interactions)	Scores	Feature (interactions)	Scores
	poor	-0.341	only, good	-0.450
	respects	0.297	EFFECT, good	-0.182
	only, good	-0.243	good, is	-0.171
	poor, only	0.206	poor, film	-0.169
	good	0.176	only, element	-0.168
	poor, respects	-0.173	doesn't, poor	0.151
	doesn't	-0.169	only, poor	0.150
	poor, good	0.122	respects, poor	-0.149
	only, doesn't	0.115	itself, poor	-0.142
	poor, doesn't	0.111	respects, good	-0.137
	many	0.095	it, doesn't	-0.108
	it	0.084	it, only	-0.098
	itself	0.083	take, seriously	0.095
	element	0.076	doesn't, good	-0.094
	poor, many	-0.070	doesn't, only	0.093
	Shapley interaction indices		*Integrated Hessian*	
	only, good	-0.280	., .	19.441
	a, only	-0.259	is, .	3.506
	only, poor	0.223	the, .	2.802
	poor, good	0.171	only, .	2.374
	doesn't, poor	0.159	take, .	2.004
	respects, poor	-0.154	it, .	1.463
	good, is	-0.150	good, .	1.462
	The, good	-0.146	seriously, .	1.372
	poor, element	-0.146	doesn, .	1.282
	doesn't, only	0.142	is, .	1.226
	doesn't, take	-0.130	that, .	1.169
	respects, only	-0.120	', '	1.143
	good, element	-0.119	., .	-1.001
	it, take	-0.117	is, is	0.996
	so, that	-0.112	itself, .	0.914
	Archipalego			
				

Table 7: Top-15 important feature (interactions) for different methods for language dataset. The predicted probability is the out probability of the sentence having positive sentiment. The archipalego algorithm is run with $k = 3$ interactions.

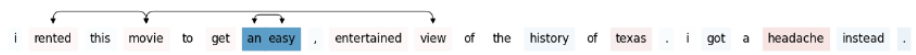
Index	Sentences			Predicted Prob.
3	I rented this movie to get an easy, entertained view of the history of Texas. I got a headache instead.			0.026
	Faithful Shapley indices		*Shapley Taylor indices*	
	Feature (interactions)	Scores	Feature (interactions)	Scores
	instead	-0.321	headache, instead	0.268
	headache, instead	0.252	view, instead	-0.178
	headache	-0.205	headache, Texas	-0.139
	easy	0.158	rented, instead	0.137
	view	0.130	instead, easy	-0.125
	history	0.123	got, headache	-0.118
	rented	-0.122	entertained, instead	-0.115
	Texas	0.101	rented, headache	0.109
	entertained	0.095	got, easy	-0.108
	rented, instead	0.085	got, history	-0.105
	Texas, headache	-0.069	a, I	-0.100
	history, instead	-0.064	view, history	0.100
	the	0.059	got, rented	-0.100
	entertained, instead	-0.057	got, a	-0.099
	this	0.052	history, an	0.094
	Shapley interaction indices		*Integrated Hessian*	
	headache, instead	0.333	., .	-13.363
	easy, I	-0.248	i, .	-2.441
	movie, instead	-0.226	., a	-2.117
	history, to	-0.162	., .	-1.420
	Texas, an	-0.135	texas, .	1.171
	rented, easy	-0.135	this, .	-1.087
	entertained, easy	0.130	., i	-1.056
	to, easy	-0.115	to, .	-1.035
	view, instead	-0.114	of, .	0.843
	entertained, instead	-0.112	entertained, entertained	-0.761
	of, instead	-0.102	headache, headache	0.753
	an, easy	0.095	history, history	-0.688
	instead, easy	-0.093	., got	-0.673
	of, easy	-0.087	view, .	0.657
	get, easy	-0.085	to, to	-0.599
	Archipalego			
				

Table 8: Top-15 important feature (interactions) for different methods for language dataset. The predicted probability is the out probability of the sentence having positive sentiment. The archipalego algorithm is run with $k = 3$ interactions.

Index	Sentences			Predicted Prob.
4	Truly appalling waste of space. Me and my friend tried to watch this film to its conclusion but had to switch it off about 30 minutes from the end.			0.002
	Faithful Shapley indices		*Shapley Taylor indices*	
	Feature (interactions)	Scores	Feature (interactions)	Scores
	waste	-0.345	appalling, waste	0.298
	appalling, waste	0.257	Truly, waste	-0.296
	appalling	-0.251	switch, it	-0.248
	Truly	0.169	tried, waste	0.230
	waste, tried	0.167	but, watch	-0.210
	friend	0.162	friend, waste	-0.184
	space	0.149	friend, tried	-0.172
	tried	-0.134	friend, but	-0.169
	Truly, waste	-0.118	Truly, but	-0.145
	watch	0.087	but, waste	0.145
	off	-0.086	waste, watch	-0.140
	and	0.078	waste, off	0.138
	waste, friend	-0.074	had, space	-0.128
	waste, space	-0.058	Truly, film	0.126
	of	0.055	30, waste	0.124
	Shapley interaction indices		*Integrated Hessian*	
	tried, watch	-0.365	the, the	-31.568
	appalling, waste	0.293	the, end	-13.784
	tried, waste	0.259	the, .	9.472
	from, end	0.230	end, end	-5.719
	conclusion, its	0.228	end, .	3.522
	Truly, waste	-0.210	from, the	3.390
	waste, space	-0.202	., .	2.616
	space, off	-0.191	had, the	1.540
	to, its	-0.180	from, end	1.441
	to, space	0.166	30, the	-0.959
	Me, its	0.162	its, the	0.941
	Truly, of	0.155	minutes, .	0.821
	the, Me	-0.154	off, the	0.796
	but, waste	0.148	., .	0.779
	had, space	-0.146	to, the	0.737
	Archipalego			
				

Table 9: Top-15 important feature (interactions) for different methods for language dataset. The predicted probability is the out probability of the sentence having positive sentiment. The archipalego algorithm is run with $k = 3$ interactions.

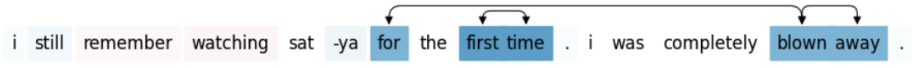
Index	Sentences			Predicted Prob.
5	I still remember watching Satya for the first time. I was completely blown away.			0.994
	Faithful Shapley indices		*Shapley Taylor indices*	
	Feature (interactions)	Scores	Feature (interactions)	Scores
	remember	0.337	blown, away	0.345
	blown, away	0.293	the, first	0.191
	time	0.281	time, first	0.182
	Satya	0.208	watching, for	-0.169
	remember, blown	-0.158	time, away	-0.167
	watching	0.153	time, Satya	-0.151
	blown	0.146	time, still	-0.145
	time, away	-0.127	still, watching	-0.144
	completely, away	-0.101	I, watching	-0.131
	Satya, time	-0.091	watching, first	-0.128
	remember, time	-0.073	remember, away	0.118
	I, watching	-0.071	Satya, away	-0.118
	completely, blown	0.063	was, watching	-0.115
	first, blown	-0.053	remember, blown	-0.110
	first	0.049	completely, away	-0.107
	Shapley interaction indices		*Integrated Hessian*	
	blown, away	0.318	., .	4.759
	was, remember	0.237	was, .	1.866
	remember, blown	-0.180	blown, blown	1.552
	time, Satya	-0.167	i, .	1.185
	the, first	0.144	was, was	1.105
	blown, first	-0.133	i, .	1.063
	completely, blown	0.126	blown, away	0.889
	time, away	-0.119	satya, .	0.857
	I, was	0.093	for, for	-0.763
	watching, blown	0.087	remember, remember	-0.745
	I, watching	-0.083	for, time	-0.745
	time, blown	0.080	i, i	-0.727
	watching, away	-0.078	., .	-0.616
	remember, watching	-0.076	watching, satya	0.592
	remember, was	-0.074	completely, .	0.579
	Archipalego			
				

Table 10: Top-15 important feature (interactions) for different methods for language dataset. The predicted probability is the out probability of the sentence having positive sentiment. The archipalego algorithm is run with $k = 3$ interactions.

Appendix D. Additional Guidance on Theorem 16

In this section, we clarify how to use Theorem 16 to design Faith-Interaction indices satisfying interaction linearity, symmetry, and dummy axioms by first explaining Theorem 16 and then providing some examples.

Theorem 16 states that the finite weighting function must be in the following form:

$$\mu(S) \propto \sum_{i=|S|}^{d} \binom{d-|S|}{i-|S|}(-1)^{i-|S|} g(a,b,i), \quad \text{where } g(a,b,i) = \begin{cases} 1 & , \text{if } \ i=0. \\ \prod_{j=0}^{j=i-1} \frac{a(a-b)+j(b-a^2)}{a-b+j(b-a^2)} & , \text{if } \ 1 \le i \le d. \end{cases}$$

for some $a, b \in \mathbb{R}^+$ with $a > b$ such that $\mu(S) > 0$ for all $S \subseteq [d]$.

To better understand this formula, some questions need to be answered: (1) What kind of a, b makes $\mu(S) > 0$ for all $S \subseteq [d]$? (2) What is the physical meaning of the parameters a and b?

To answer (1), we show that a simple condition $1 \ge a > b \ge a^2 > 0$ suffices to make $\mu(S) > 0$ for all $S \subseteq [d]$.

Proposition 23 *When $a, b \in \mathbb{R}^+$ such that $1 \ge a > b \ge a^2 > 0$, we have*

$$\sum_{i=|S|}^{d} \binom{d-|S|}{i-|S|}(-1)^{i-|S|} g(a,b,i) > 0 \quad \textit{for all } S \subseteq [d],$$

where $g(a,b,i)$ is defined in Eqn.(12).

We delayed the proof of this proposition to arxiv. We note that it is only a sufficient condition for selecting a and b: For some small $d \in \mathbb{N}$, we may have some a, b such that $1 > a^2 > b > 0$ but makes $\mu(S) > 0$ for all $S \subseteq [d]$. However, if $a = \bar{\mu}_1$ and $b = \bar{\mu}_2$ need to make the weighting function positive for all $d \in \mathbb{N}$, we must have the condition $1 \ge a > b \ge a^2 > 0$.

For question (2), we show that $\bar{\mu}_i = g(a,b,i) = \sum_{L \supseteq S} \mu(L)$ for subsets $S \subseteq [d]$ with $|S| = i$. Here, $\bar{\mu}_i$ is defined as the total weight of coalitions containing a group of features of size i (any group with size i will work due to the interaction symmetry axiom). By plugging in $i = 1, 2$, we get $a = \bar{\mu}_1$ and $b = \bar{\mu}_2$ are the total weights of coalitions containing a single feature and a pair of features.

In the following, we give some special cases with particularly chosen a and b to provide an intuition of Theorem 16,

Example 1 *When $a = 0.5$ and $b = 0.25$, the weighting function $\mu(\cdot)$ with respect to Theorem 16 is $\mu(S) = 1/2^d$ for all $S \subseteq [d]$. In this case, the explanations $\mathcal{E}_T(v, \ell)$ equals the Banzhaf Interaction value up to order ℓ for all $|T| = \ell$, which has the form $\mathcal{E}_T(v, \ell) = \sum_{S \subseteq [d] \setminus T} \Delta_T v(S)/2^{d-|S|}$.*

In this example, the Banzhaf interaction value satisfies interaction linearity, symmetry, and dummy axioms Fujimoto et al. (2006), which coincides with Theorem 16. We also provide another guideline to design the values of a, b based on the desired $\frac{\mu_d}{\mu_{d-1}}$ and $\frac{\mu_{d-1}}{\mu_{d-2}}$, where $\mu_i = \mu(S)$ when $|S| = i$. [1]

1. μ_i can be defined since the interaction symmetry axiom ensures that all coalitions with equal size have equal weights.

Proposition 24

$$\textit{Let } \frac{\mu_d}{\mu_{d-1}} = r_1 \quad \textit{and} \quad \frac{\mu_{d-1}}{\mu_{d-2}} = r_2 \quad \textit{with } r_1 > r_2 > \frac{(d-2)r_1}{d-1+r_1} > 0,$$

then a and b can be represented as functions of r_1 and r_2:

$$a = \frac{r_1(r_2+1) - (d-1)(r_1-r_2)}{(r_1+1)(r_2+1) - (d-1)(r_1-r_2)} \quad \textit{and} \quad b = \frac{r_1(r_2+1) - (d-2)(r_1-r_2)}{(r_1+1)(r_2+1) - (d-2)(r_1-r_2)}a. \tag{24}$$

In this case, a and b satisfy $1 > a > b \geq a^2 > 0$, which implies $\mu_i > 0$ for all $0 \leq i \leq d$.

This proposition provides a guideline to design a unique interaction value that satisfies interaction linearity, symmetry, dummy axioms based on given values of $\frac{\mu_d}{\mu_{d-1}}$ and $\frac{\mu_{d-1}}{\mu_{d-2}}$. For example, if the coalition μ_t has a higher probability to form when t is large, such as the case when the features of an image are explained. As an example, we may set $\frac{\mu_d}{\mu_{d-1}} = 10$. We then have $10 > r_2 > \frac{d-2}{d+9}10$, and we can set $r_2 = 9$ when $d < 101$. This narrows down a unique interaction value that satisfies these three axioms and the conditions of $\frac{\mu_d}{\mu_{d-1}} = 10$ and $\frac{\mu_{d-1}}{\mu_{d-2}} = 9$.

Appendix E. Auxiliary Theoretical Results

In this section, we provide auxiliary theoretical results of the Faith-Interaction indices. These properties are useful in the proof of our main theorems. The proof are delayed to Appendix arxiv.

First of all, we show that if the coalition weighting function $\mu(\cdot)$ is finite, Eqn.(9) is strictly convex.

Proposition 25 *If the coalition weighting function $\mu(\cdot)$ is finite such that $\mu(S) \in \mathbb{R}^+$ for all $S \subseteq [d]$, Eqn.(9) is strictly convex.*

Given that Eqn.(9) is strictly convex, we next show that the minimization problems have a unique minimizer.

Proposition 26 *The (constrained) regression problems defined in Eqn.(10) with a proper weighting function μ (Definition 10) have a unique minimizer.*

This proposition is a straightforward application of the following fact: For a minimization problem with linear constraints, if the objective is strictly convex, then it has a unique minimizer.

Also, we note that having a positive measure for all subsets of $[d]$ on the weighting function $\mu(\cdot)$ is necessary to ensure the uniqueness of the minimizer. Consider the case when the maximum interaction order equals the number of features, i.e. $\ell = d$, there are 2^d variables with 2^d equalities. That is, $v(S) - \sum_{T \subseteq S} \mathcal{E}_S(v, d) = 0$ for all $S \subseteq [d]$. In this case, we can not have any $S \subseteq [d]$ such that $\mu(S) = 0$ due to the lack of equations.

In this special case of $\ell = d$, we have the following closed-form expression. We note that these results are independent of the weighting function as long as we have $\mu(S) > 0$ for all $S \subseteq [d]$.

Proposition 27 *When the maximum interaction order $\ell = d$, the minimizer of Eqn.(10) the Möbius transform of v, i.e. $\mathcal{E}_S(v, d) = a(v, S) = \sum_{T \subseteq S} (-1)^{|S|-|T|} v(T)$ for all subsets $S \subseteq [d]$.*

Then we provide the expression of partial derivatives of the objective in Eqn.(9) with respect to each variable $\mathcal{E}_A(v, \ell)$ for all $A \subseteq [d]$ with $|A| \leq \ell$.

Proposition 28 *The partial derivative of Eqn.(9) with respect to $\mathcal{E}_A(v, \ell)$ is*

$$-2 \sum_{\substack{S: S \supseteq A, \\ \mu(S) < \infty}} \mu(S) v(S) + 2 \sum_{S \in \mathcal{S}_\ell} \mathcal{E}_S(v, \ell) \sum_{\substack{L: L \supseteq S \cup A, \\ \mu(L) < \infty}} \mu(L) \quad \textit{for all} \quad A \in \mathcal{S}_\ell. \tag{25}$$

This proposition is frequently used in our proof as we solve the minimization problem. Next, the following proposition illustrates how to solve the constrained regression problem via Lagrangian.

Proposition 29 *Any Faith-Interaction index $\mathcal{E}(v,\ell)$ with respect to a proper weighting function $\mu(\cdot)$ with $\mu(\emptyset)=\mu([d])=\infty$ has the form:*

$$\begin{bmatrix} \lambda_\emptyset \\ \lambda_{[d]} \\ \mathcal{E}_\emptyset(v,\ell) \\ \cdots \\ \mathcal{E}_S(v,\ell) \\ \mathcal{E}_T(v,\ell) \\ \cdots \end{bmatrix} = \underbrace{\begin{bmatrix} 0, & 0, & 1, & \cdots, & 0, & 0, & \cdots \\ 0, & 0, & 1, & \cdots, & 1, & 1, & \cdots \\ -\frac{1}{2}, & -\frac{1}{2}, & \bar{\mu}(\emptyset), & \cdots, & \bar{\mu}(S), & \bar{\mu}(T), & \cdots \\ \cdots, & \cdots, & \cdots, & \cdots, & \cdots, & \cdots, & \cdots, \\ 0, & -\frac{1}{2}, & \bar{\mu}(S), & \cdots, & \bar{\mu}(S), & \bar{\mu}(S\cup T), & \cdots \\ 0, & -\frac{1}{2}, & \bar{\mu}(T), & \cdots, & \bar{\mu}(S\cup T), & \bar{\mu}(T), & \cdots \\ \cdots, & \cdots, & \cdots, & \cdots, & \cdots, & \cdots, & \cdots \end{bmatrix}^{-1}}_{\mathbf{M}^{-1}} \underbrace{\begin{bmatrix} v(\emptyset) \\ v([d]) \\ \bar{v}(\emptyset) \\ \cdots \\ \bar{v}(S) \\ \bar{v}(T) \\ \cdots \end{bmatrix}}_{\mathbf{y}}, \tag{26}$$

where $\lambda_\emptyset$ and $\lambda_{[d]}$ are Lagrange multipliers with respect to the constraints on the empty set and the full set, $\bar{\mu}(S)=\sum_{L\supseteq S,\mu(L)<\infty}\mu(L)$, and $\bar{v}(S)=\sum_{L\supseteq S,\mu(L)<\infty}\mu(L)v(L)$.

Formally, the matrix $\mathbf{M}\in\mathbb{R}^{(d_\ell+2)\times(d_\ell+2)}$ and the vector $\mathbf{y}\in\mathbb{R}^{d_\ell+2}$ have the following definitions: we overuse the notations $\lambda_\emptyset,\lambda_{[d]}$ and let the rows and columns of $\mathbf{M}$ are indexed by $\{\lambda_\emptyset,\lambda_{[d]},\emptyset,\cdots,S,T,\cdots\}$, which are corresponding to variables $\lambda_\emptyset,\lambda_{[d]},\mathcal{E}_{\emptyset(v,\ell),\cdots,}\mathcal{E}_S(v,\ell),\mathcal{E}_T(v,\ell)$

$$\mathbf{M}_{S,T}=\begin{cases} 1 & \text{if } S=(\lambda_\emptyset)\wedge(T=\emptyset). \\ 0 & \text{if } (S=\lambda_\emptyset)\wedge(T\neq\emptyset). \\ 1 & \text{if } (S=\lambda_{[d]})\wedge(T\subseteq\mathcal{S}_\ell). \\ 0 & \text{if } (S=\lambda_{[d]})\wedge(T\in\{\lambda_\emptyset,\lambda_{[d]}\}). \\ -\frac{1}{2} & \text{if } (S=\emptyset)\wedge(T=\lambda_{[d]}). \\ 0 & \text{if } (S\in\mathcal{S}_\ell\backslash\emptyset)\wedge(T=\lambda_\emptyset). \\ -\frac{1}{2} & \text{if } (S\in\mathcal{S}_\ell)\wedge(T=\lambda_{[d]}). \\ \bar{\mu}(S\cup T) & \text{, otherwise.} \end{cases}, \quad \text{and} \quad \mathbf{y}_S=\begin{cases} v(\emptyset) & \text{if } S=\lambda_\emptyset. \\ v([d]) & \text{if } S=\lambda_{[d]} \\ \bar{v}(S) & \text{otherwise.} \end{cases}$$

where we use $\mathbf{M}_{S,T}$ to denote the entry of the intersection of S^{th} row and T^{th} column.