The Shapley Taylor Interaction Index

Kedar Dhamdhere [1] **Ashish Agarwal** [1] **Mukund Sundararajan** [1]

Abstract

The *attribution* problem, that is the problem of attributing a model's prediction to its base features, is well-studied. We extend the notion of attribution to also apply to feature interactions. The Shapley value is a commonly used method to attribute a model's prediction to its base features. We propose a generalization of the Shapley value called Shapley-Taylor index that attributes the model's prediction to interactions of subsets of features up to some size k. The method is analogous to how the truncated Taylor Series decomposes the function value at a certain point using its derivatives at a different point. In fact, we show that the Shapley Taylor index is equal to the Taylor Series of the *multilinear extension* of the set-theoretic behavior of the model. We axiomatize this method using the standard Shapley axioms—*linearity*, *dummy*, *symmetry* and *efficiency*—and an additional axiom that we call the *interaction distribution* axiom. This new axiom explicitly characterizes how interactions are distributed for a class of functions that model pure interaction. We contrast the Shapley-Taylor index against the previously proposed Shapley Interaction index (cf. (9)) from the cooperative game theory literature. We also apply the Shapley Taylor index to three models and identify interesting qualitative insights.

1. Introduction

1.1. Motivation

There is considerable literature on feature importance/attribution for deep networks(cf. (2; 23; 22; 3; 26; 14; 28; 14)). The basic idea of attribution is to distribute the prediction score of a model for a specific input to its base features; the attribution to a base feature can be interpreted as its contribution to the prediction. For instance, when attribution is applied to a network that predicts the sentiment associated with a paragraph of text, it quantifies the influence of every word in the text on the network's score. This can be used, for instance, to tell if the model bases its prediction on words that connote a protected category like a specific race/gender/religion. This would be indicative of the model possibly being biased along the protected dimension. Attribution/feature importance for Deep Networks has been applied to a variety of real world applications, for instance in health, drug-discovery, machine translation, natural language tasks, recommendation systems etc. Thus, attributions are quite useful despite their simple form; notice they don't reveal the logic of the network beyond base feature importance.

In this paper, we take a step towards making attributions a somewhat richer form of explanation by identifying the importance of *feature interactions*, either pairwise or of higher orders. We would like to identify to what extent a set of features exert influence in conjunction as opposed to independently. We expect the study of interactions to be fruitful. Deep networks are likely to have an abundance of strong feature interactions, because they excel at creating higher-order representations (e.g. filters) out of the base features (e.g. pixels). We also expect the study of interactions to be critical for the tasks that cannot be performed by features acting independently. We study two such tasks in Section 5. In the sentiment analysis task, *negation*, should manifest as an interaction between the negation word (e.g. not) and the sentiment bearing word (e.g. good or bad) that it modifies. If such an interaction is not detectable, then the network needs fixing. Another example is reading comprehension, i.e., question answering about paragraphs of text. A good model will match question words to certain words/phrases in the paragraph, and these matches should manifest as interactions between those words.

Before we describe our contributions, we mention some related work on feature interactions besides the attribution literature briefly described above.

[1]Google LLC. Correspondence to: Mukund Sundararajan <mukunds@google.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

1.2. More Related Work

1.2.1. Shapley Value, Shapley Interaction Value

Some of the deep network attribution literature (cf. (28; 12; 13; 6; 7; 27; 25)) is built on prior work in cooperative game theory, specifically, the Shapley value ((21)) and its continuous variant (1). These prescribe a way to distribute the value of a game among its players.[1] Shapley values have been used to study global feature/variable importance in statistics (cf. (17; 4)). The work most closely related to ours is the Shapley interaction value, first proposed by (18) to study pairwise interactions between players. (9) generalized it to study interactions of higher orders, and provided an axiomatic foundation. (13) applied it to studying feature interactions in trees. We provide comparison against the Shapley interaction value.

1.2.2. Interactions in Machine Learning

It is hard to describe all of the vast literature on interactions in machine learning and statistics. Most of this literature is focused on *global* feature importance, i.e., important interactions across the data set. In contrast, we study feature importance for individual inputs.

There is, for instance, the classic literature on ANOVA (cf. (8)), and the more recent literature on Lasso, (29), both of which can be used to quantify the importance of putative interactions.

We mention some recent deep learning literature: (30) constructs a generalized additive model that mimics the behavior of a deep network by investigating the structure of the inter-layer weight matrices. (31) forces the weight matrices to be block-diagonal, restricting the type of interactions, by designing the appropriate regularization. (5) studies *pairwise* interactions by building deep networks of a specific form and then interpreting the network via its gradients, and averaging appropriately over the data set. (24) combines agglomerative clustering with influence propagation across the layers of the deep-network, to produce a hierarchical decomposition of the input with influence scores for each group; this work, unlike the others in this section, is about feature importance of individual inputs.

1.3. Model

In this section, we formally model the attribution problem for interactions. We have a set of features N. The deep network is modeled as a function $F : 2^N \to \mathbb{R}$, i.e., we treat the features as discrete variables. Modeling the network as a function of Boolean features simplifies our theory of attribution, i.e., we can investigate the influence of variables via the change in score resulting from removing/ablating the variable. For many problems, treating features as Boolean is natural; for a text model, words are either present or absent. In our applications, we will model the absence of a feature by replacing it with a sentinel value (zero embedding vector/out of vocabulary word/average value of the feature across the training data).

To simplify the notation, we denote the cardinalities of a set S, T etc. using lowercase letters: s, t. We omit braces for small sets and write $T \cup i$ instead of $T \cup \{i\}$ and $T \cup ij$ instead of $T \cup \{i, j\}$.

Define

$$\delta_i F(T) = F(T \cup i) - F(T) \tag{1}$$

and

$$\delta_{ij} F(T) = F(T \cup ij) - F(T \cup i) - F(T \cup j) + F(T) \tag{2}$$

Here $i, j \notin T$.[2] These are discrete equivalents of first and second order derivatives. In general, for $T \subseteq N \setminus S$, we define the discrete derivative with respect to set S as:

$$\delta_S F(T) = \sum_{W \subseteq S} (-1)^{w-s} F(W \cup T) \tag{3}$$

We use a fixed number k as the **order of explanation** to mean that we'll compute the Shapley-Taylor indices for subsets of size up to k. For instance, $k = 2$ corresponds to computing main effects as well as pairwise interaction effects. For a set S such that $|S| \leq k$, $\mathcal{I}^k_S(v)$ denotes the interaction effect for the set S.

1.4. Our Results

Our goal is to axiomatically identify the *best kth order explanation*. For instance, when $k = 2$, we would like to identify the main effects of each feature and the interaction effects for every pair of features.

- We propose the *Shapley-Taylor index* to solve this problem (Section 2).
- We axiomatize the Shapley-Taylor indices in Section 3. We introduce a new axiom called the *interaction distribution* axiom. This axiom explicitly defines how interactions are distributed for a class of functions called *interaction functions*.
- We compare the Shapley-Taylor indices against a previously proposed method called Shapley interaction indices ((9)). This method is the unique method that satisfies variants of the standard Shapley axioms, namely

[1] Games are analogous to models, players to features, and the shares to the feature importance.

[2] If features i and j don't interact, this quantity will be zero, because adding j to the set T or to the set $T \cup i$ will have the same influence.

dummy, linearity and symmetry (see Section 3 for axiom definitions) and an additional axiom called *the recursive axiom*[3]. Critically, the Shapley interaction indices do not satisfy efficiency, i.e., the attributions do not sum to the score of the function at the input minus its score at the empty set. Axiomatically, the main contribution of this paper is to replace the recursive axiom with the efficiency axiom; the consequence is that the method of choice goes from being Shapley interaction indices to Shapley-Taylor indices. Section 4 contrasts the results of the two methods. We find that the lack of efficiency causes Shapley interaction indices to amplify interaction effects (Section 4.1), or causes the interaction effects to have seemingly incorrect signs (Section 4.2).

- In Section 3.3, we connect the Shapley-Taylor interaction index to the Taylor series (with a Lagrangian remainder); we show that the Taylor series applied to the so called multilinear extension of the function is equivalent to the Shapley-Taylor index applied to the function.
- Though our key contributions and evaluations are mainly theoretical, we demonstrate the applicability of our work in Section 5, which studies models for three tasks (sentiment analysis, random forest regression, and question answering). We identify certain interesting interactions.

2. Shapley-Taylor indices

2.1. Shapley Value

Before we discuss feature interactions, let us revisit the Shapley value. The central concept in the Shapley value is that of the *marginal*, i.e., the change in the function value F by the addition of a feature, i.e. $\delta_i F(S) = F(S \cup i) - F(S)$. In general, for a nonlinear function F, the value of this expression depends on the set $S \subseteq N \setminus \{i\}$ at which we compute the marginal and there are several choices for this set S. The Shapley value defines a random process that implicitly prescribes a certain weighting over these sets. Given an ordering of the features, add the features one by one in this order. Each feature i gets ascribed the marginal value of adding it to the features that precede it. The Shapley value of the feature i is the expected value of this marginal over an ordering of features chosen uniformly at random.

The Shapley Value is known to be the unique method that satisfies four axioms (see Section 3 for formal definitions of the axioms): *efficiency* (the attributions of all the features sum to the difference between the prediction scores at the input minus that at the empty set.), *symmetry* (symmetric features receive equal attributions), *linearity* (the sum of the attributions of a feature for two functions is identical to the attribution of the feature in the game formed by the sum of the two functions) and the *dummy* (if the marginal value of a feature is zero for every set S, it has a zero attribution) axioms.

[3]The recursive axiom requires that the interaction index for a pair of features i, j is equal to the difference in the Shapley value of feature i in a game with feature j omnipresent and the Shapley value of feature i in a game with feature j absent.

2.2. Definition of Shapley-Taylor indices

In this section we define Shapley-Taylor indices. Just like the Shapley value, Shapley-Taylor indices can also be computed as an expectation over orderings chosen uniformly at random. The output of the Shapley-Taylor indices is more extensive compared to the Shapley value; it includes attributions to interaction terms.

For an order of explanation $k = 2$, i.e., the index specifies values for individual features and pairs of features. The indices for an individual feature i is just the marginal $\delta_i F(\emptyset)$. To compute the indices for pairs, we pick an ordering of the features, and ascribe to each pair of features the expression 2 computed at the set S of features that precede both i and j in the ordering. The Shapley Taylor interaction index for a pair is the expectation of this quantity over an ordering chosen uniformly at random.

Similarly for the general case, Shapley-Taylor indices are defined by random process over orderings of features. Let $\pi = (i_1, i_2, \ldots, i_n)$ be an ordering. Let $\pi^{i_k} = \{i_1, \ldots, i_{k-1}\}$ be the set of predecessors of i_k in π. For a fixed ordering π and a set, we define Shapley-Taylor indices $\mathcal{I}^k_{S,\pi}(F)$ as follows:

$$\mathcal{I}^k_{S,\pi}(F) = \begin{cases} \delta_S F(\emptyset) & \text{if } |S| < k \\ \delta_S F(\pi^S) & \text{if } |S| = k, \end{cases} \tag{4}$$

Here, π^S is defined as $\cap_{i \in S} \pi^i$, the set of elements that precede all of the features in S. Let us briefly discuss the three cases. When the size of the interaction term is strictly less than the order of explanation, its interaction value (for the fixed permutation) is simply equal to the discrete derivative (Equation 3) at the empty set. Notice that this quantity does not depend on the permutation itself. When the order of approximation is $k = 2$, this is just the marginal value of adding the feature to the empty set. When the size of the interaction term is *equal* to the order of explanation, its interaction value (for the fixed permutation) is equal to the discrete derivative at the largest set of elements that precede all of the elements of S. when the order of approximation is $k = 2$, the discrete derivatives match Equation 2.

The Shapley-Taylor indices are defined as the expectation of $\mathcal{I}^k_{S,\pi}(F)$ over an ordering of the N features chosen uni-

formly at random:

$$\mathcal{I}^k_S(F) = \mathbb{E}_\pi(\mathcal{I}^k_{S,\pi}(F)) \tag{5}$$

A noteworthy special case is $k = 1$. Here the first case of Equation 4 does not occur and the discrete derivatives correspond to marginals. The resulting definition is precisely the well-known Shapley value. Shapley-Taylor indices thus generalize the Shapley value.

We have defined the Shapley-Taylor indices in terms of permutations. The theorem below derives a closed form expression for Shapley-Taylor indices. This provides a method to compute the Shapley-Taylor indices. We have given the proof in the Appendix (Section 7.1).

Theorem 1. *Let k be the order of explanation. For a set $S \subseteq N$, such that $|S| = k$, Shapley-Taylor indices satisfy the following expression:*

$$\mathcal{I}^k_S(F) = \frac{k}{n} \sum_{T \subseteq N \setminus S} \delta_S F(T) \frac{1}{\binom{n-1}{t}} \tag{6}$$

Remark 1. *The formula in Theorem 1 gives us a way to compute the Shapley-Taylor indices. It involves computation over all subsets of the feature set and hence it takes exponential time. In practice, we can trade off accuracy for speed. One way to obtain a fast approximation is to apply Equation 4 over a sample of permutations. This is similar to the Shapley value sampling methods(15). In Section 5, we use a different approximation. We first identify a subset of features with high attribution (using Shapley values or Integrated Gradients method). We use Theorem 1 formula on this subset.*

We now define the previously proposed **Shapley interaction index** ((9)) [4]. The Shapley interaction index for a subset $S \subseteq N$ for a function $F \subseteq \mathcal{G}^N$ is defined as:

$$I_{Sh}(F, S) := \sum_{T \subseteq N \setminus S} \frac{(n-t-s)!t!}{(n-s+1)!} \delta_S F(T) \tag{7}$$

3. Axiomatization of the Shapley-Taylor Interaction Index

In this section, we axiomatize Shapley-Taylor indices. Let $\mathcal{G}^N$ denote the set of functions on N features.

The first three axioms are all variants of the standard Shapley axioms generalized to interactions by (9); they were used in the axiomatization of Shapley interaction indices.

[4]The Shapley interaction index can also defined by a random order process. To compute the interaction index for a set S, the set S is fused into an artificial player. The players are ordered randomly as before and the discrete derivative δ_S is evaluated at the set of players which precede the artificial player in the ordering. The combinatorial weights that occur in Equation 7 arise from this random process.

1. **Linearity axiom:** $\mathcal{I}^k(\cdot)$ is a linear function; i.e. for two functions $F_1, F_2 \in \mathcal{G}^N$, $\mathcal{I}^k_S(F_1 + F_2) = \mathcal{I}^k_S(F_1) + \mathcal{I}^k_S(F_2)$ and $\mathcal{I}^k_S(c \cdot F_1) = c \cdot \mathcal{I}^k_S(F_1)$.

2. **Dummy axiom:** If i is a dummy feature for F, i.e. $F(S) = F(S \setminus i) + F(i)$ for any $S \subseteq N$ with $i \in S$, then
   (i) $\mathcal{I}^k_i(F) = F(i)$
   (ii) for every $S \subseteq N$ with $i \in S$, we have $\mathcal{I}^k_S(F) = 0$

3. **Symmetry axiom:** for all functions $F \in \mathcal{G}^N$, for all permutations π on N, :
   $$\mathcal{I}^k_S(F) = \mathcal{I}^k_{\pi S}(\pi F)$$
   where $\pi S := \{\pi(i) | i \in S\}$ and the function πv is defined by $(\pi F)(\pi S) = F(S)$, i.e. it arises from relabeling of features $1, \ldots, n$ with the labels $\pi(1), \ldots, \pi(n)$.

In addition to these three axioms from (9), we use the efficiency axiom. Again, this is a generalization of the standard efficiency axiom for Shapley values.

4. **Efficiency axiom:** for all functions F, $\sum_{S \subseteq N, |S| \leq k} \mathcal{I}^k_S(F) = F(N) - F(\emptyset)$

Finally, we introduce the *Interaction Distribution* axiom. This axiom is defined for functions that we call *interaction functions*. An interaction function parameterized by a set T, has the form $F_T(S) = 0$ if $T \not\subseteq S$ and has a constant value $F_T(S) = c$ when $T \subseteq S$. These functions model pure interaction among the members of T (when $|T| > 1$—the combined presence of the members of T is necessary and sufficient for the function to have non-zero value. We call $|T|$ the order of the interaction of the interaction function. In machine learning terminology, the function $F_T(S)$ is a model with a single feature cross of all the (categorical) features in T, with a coefficient of c for this cross.

5. **Interaction Distribution Axiom:** For an interaction function F_T parameterized by the set T, for all S with $S \subsetneq T$ and $|S| < k$, where k denotes the order of explanation, we have $\mathcal{I}^k_S(F_T) = 0$.

3.1. The Roles of the Axioms

Here we discuss informally the roles that the various axioms play, and specifically, the need for the Interaction Distribution axiom. For simplicity, we consider explanations of order $k = 2$. Consider the space of Interaction Distribution methods that satisfy linearity. It is well-known that the set of unanimity functions (defined as $u_T(S) = 1$ if $T \subseteq S$ and 0 otherwise) constitute a basis for the space of all functions. That is, every function v, can be expressed as a unique linear

combination of these functions. By linearity, the interaction indices for the function v should be the same linear combination of the interaction indices for the unanimity functions. Therefore, defining the interaction indices for unanimity functions specifies the interaction indices for all functions.

Specifically, consider the unanimity function defined for a set $T \subseteq N$, such that $|T| > 1$. Now let us consider the effect of the Dummy and Symmetry axioms. Dummy constrains us from attributing to interaction terms that contain elements not in T. Symmetry dictates that all the singleton elements in T get the same credit (say a_T), and all the pairwise terms (i, j) such that $i \in$ and $j \in T$, get the same credit (say b_T).

Recall that Shapley interaction indices satisfy Linearity, Dummy and Symmetry, and as we show later, so do Shapley-Taylor indices. Pairwise Shapley interaction indices assign a value of $b_T = 1/(|T| - 1)$ to every one of the $\binom{|T|}{2}$ pairs of elements. Pairwise Shapley interaction indices does not define assignments to the singleton terms, but for simplicity, we can consider $a_T = 0$. Pairwise Shapley interaction indices thus violate efficiency.

Shapley-Taylor indices assign a value of $b_T = 1/\binom{|T|}{2}$ to every pair of elements from the set T and $a_T = 0$; this satisfies symmetry, dummy, linearity, but also efficiency. However, this is not the unique way to satisfy efficiency; there are other assignments to a_T and b_T that still satisfy efficiency. [5] This ambiguity is tied to non-orthogonality—-the singleton sets are not orthogonal to the pairs. The consequence is that we need an additional axiom to pin down the interaction index. In our case, the Interaction Distribution axiom causes the explicit choice of setting the singleton assignments a_T to zero. This prevents the singleton elements from accumulating higher order interactions (recall that we are discussing the case $|T| > 1$).

In contrast, notice that the assignment to the pairwise shares is reductive—the pairwise shares capture not only the pairwise interactions ($|T| = 2$) , but also larger interactions ($|T| > 2$). This reductiveness is necessitated by the constraint on the order of explanation ($k = 2$) and the efficiency axiom. Notice that the very same reductiveness is present in the Shapley value (order of explanation $k = 1$); recall that the Shapley value is the Shapley-Taylor indices when $k = 1$, see Section 3.3. In the Shapley value, all the interactions are redistributed among the singleton shares. As we increase the order of explanation, the amount of reductiveness decreases—progressively larger interactions (of size $k - 1$ or less) are captured unambiguously. However, this is at the cost of more computation and a longer explanation.

[5] A different method is to apply the Shapley value to the Shapley value; i.e. treat the Shapley value of an element itself as a set function and apply Shapley value on that set function. It turns out that this is well-defined. Doing so sets $a_T = b_T$. While it is an open question to identify an axiom that causes this assignment, this does show that the approach of explicitly defining interactions for interaction functions (recall that every basis function is an interaction functions) is interesting beyond Shapley-Taylor indices.

3.2. Main result

We are now ready to state our main result:

Theorem 2. *Shapley-Taylor indices are the only interaction indices that satisfy axioms 1-5.*

Proof. We prove this in two steps. First we show that Shapley-Taylor indices satisfy the axioms (Propositions 1–3). Next we show that any method that satisfies the axioms assigns specific interaction values to unanimity functions (Proposition 4).

Proposition 1. *Shapley-Taylor indices satisfy the Linearity, Dummy and Symmetry axioms*

Proof. In Equation 4, Shapley-Taylor indices are defined as expected values of certain discrete derivatives δ_S. The discrete derivative satisfies linearity conditions. Hence Shapley-Taylor indices satisfy the linearity axiom. The symmetry axiom follows from the fact that Shapley-Taylor indices are defined as expectations over all permutations. To show the dummy axiom, we note that the discrete derivative $\delta_S F(T)$ can be rewritten as $\delta_{S\setminus i} F(T \cup i) - \delta_{S\setminus i} F(T)$ for any $i \in S$. If i is a dummy feature, it follows that $\delta_S F(T) = 0$. Consequently $\mathcal{I}^k_S(v) = 0$. Furthermore, $\mathcal{I}^k_{\{i\}}(F) = \delta_i F(\emptyset) = F(i)$. Thus Shapley-Taylor indices also satisfy the dummy axiom. □

Proposition 2. *Shapley-Taylor indices satisfy the Interaction Distribution axiom.*

Proof. Consider an interaction function F_T and S such that $|S| < |T|$. Notice that $\mathcal{I}^k_S(F_T) = \delta_S(F_T) = \sum_{W\subseteq S}(-1)^{w-s} F_T(W)$. Since $|W| < |T|$, we know that $F_T(W) = 0$. Hence $\mathcal{I}^k_S(F_T) = 0$. This shows that Shapley-Taylor indices satisfies the Interaction Distribution axiom. □

Proposition 3. *Shapley-Taylor indices satisfy the efficiency axiom. Formally,*

$$\sum_{S,|S|\leq k} \mathcal{I}^k_S(F) = F(N) - F(\emptyset)$$

Proof. We prove the proposition for unanimity functions. By using additivity axiom, this extends to all functions. Consider the unanimity function u_T defined as $u_T(S) = 1$ if $S \supseteq T$, and 0 otherwise.

Lemma 1. *For all sets W, $\delta_S u_T(W) = 0$ if either $S \not\subseteq T$ or $T \not\subseteq S \cup W$.*

Fix a permutation π as the ordering of the features. Let k be the order of explanation. The two lemmas above have following implications:
- $\mathcal{I}^k_{S,\pi}(u_T) = 0$ if $|S| > |T|$.
- For $|T| < k$, $\mathcal{I}^k_{S,\pi}(u_T) = 1$ iff $S = T$, otherwise 0.
- For $|T| < k$, this proves efficiency.

The remaining case is $|T| \geq k$. For $|S| < k$, we have $\mathcal{I}^k_{S,\pi}(u_T) = \delta_S u_T(\emptyset)$. This is 0 from second lemma.

Finally, consider $|S| = k$. Here, we claim that $\mathcal{I}^k_{S,\pi}(u_T) = 1$ iff S is same as the last k elements of T in the permutation order π; otherwise it is 0.

Let $T_k := \{\text{last } k \text{ elements of } T\}\}$. Notice that if $S = T_k$, then $T \subseteq \pi^S \cup S$. Therefore, $u_T(S \cup \pi^S) = 1$, but $u_T(S' \cup \pi^S) = 0$ for any subset $S' \subset S$. Hence $\delta_S u_T(\pi^S) = 1$.

To prove the other side, there are two cases:

(a) S has an element that is not in T: first lemma gives $\mathcal{I}^k_{S,\pi}(u_T) = 0$.
(b) S does not contain an element of T_k: then $T \not\subseteq S \cup \pi^S$. Second lemma implies $\mathcal{I}^k_{S,\pi}(u_T) = 0$.

This proves the efficiency axiom for $|T| \geq k$. □

To prove the uniqueness, we investigate interactions for unanimity functions. Recall that, a unanimity function parameterized by a set $T \subseteq N$ is defined as $F_T(S) = 1$ iff $S \supseteq T$, and 0 otherwise. Thus unanimity functions are a subset of Interaction functions used for the Interaction Distribution Axiom.

Since the family of unanimity functions $\{F_T\}_{T \subseteq N, T \neq \emptyset}$ forms a linear basis of $\mathcal{G}^N$, it is sufficient to show that any interaction index that satisfies the axioms 1-5 assigns specific values on *unanimity functions*. This is shown in the next proposition.

Proposition 4. *Consider the unanimity function F_T defined as $F_T(S) = 1$ if $T \subseteq S$, otherwise 0. Let k be the order of explanation. Let ϕ be an interaction index that satisfies the axioms 1-5, then*

$$\phi_S(F_T) = \begin{cases} 1 & \text{if } S = T \text{ and } |S| < k \\ 0 & \text{if } S \neq T \text{ and } |S| < k \\ \frac{1}{\binom{t}{k}} & \text{if } S \subseteq T \text{ and } |S| = k \\ 0 & |S| = k, \text{ but } S \not\subseteq T \end{cases} \tag{8}$$

Proof. Consider a unanimity function F_T and an interaction index ϕ_S that satisfies axioms 1-5. We want to derive $\phi_S(F_T)$ using the axioms.

Start with the dummy axiom. If $i \notin T$, then i is a dummy feature for F_T. This implies that if $S \setminus T \neq \emptyset$, then $\phi_S(F_T) = 0$.

Consider $|S| < k$. The interaction distribution axiom states that $\phi_S(F_T) = 0$ if $S \subsetneq T$. Next we use the dummy axiom. Note that, for all $j \notin T$, $\delta_j F_T(S) = 0$. Thus j is dummy feature. The dummy axiom implies that $\phi_S(F_T) = 0$ if $j \in S$. Hence, $\phi_S(F_T) = 0$ for all S such that $|S \setminus T| > 0$.

Using the efficiency axiom, we have $\sum_S \phi_S(F_T) = F_T(N) = 1$. As we saw before, $\phi_S(F_T) = 0$ if $S \neq T$. Hence the sum reduces to $\phi_S(F_S) = 1$, when $S = T$.

Finally, consider the case where size of S is same as the order of explanation, i.e. $|S| = k$. As we saw earlier, $\phi_S(F_T) = 0$ if $S \not\subseteq T$. Using the efficiency axiom: $\sum_{S \subseteq T} \phi_S(F_T) = 1$. Furthermore, since k is the order of explanation, the interaction index is defined to be 0 for sets larger than k. Hence, $\sum_{S \subseteq T, |S|=k} \phi_S(F_T) = 1$.

The symmetry axiom implies that each of these terms must be equal. Hence $\phi_S(F_T) = \frac{1}{\binom{t}{k}}$. □

Proposition 4 shows that the interaction values for a unanimity function F_T depend only on $|T|$ and the order of explanation k. Since unanimity functions form a basis of $\mathcal{G}^N$, the interaction values extend in a unique way to all the functions. Thus there is a unique method (Shapley-Taylor indices) that satisfies the axioms 1-5. □

3.3. Connection to Taylor series

In this section, we connect Shapley-Taylor indices to the Taylor series of the multilinear extension.

The multilinear extension of F is defined as follows:

$$f(x) = \sum_{S \subseteq N} F(S) \prod_{i \in S} x_i \prod_{i \notin S} (1 - x_i) \tag{9}$$

where $x_i \in [0, 1], i \in N$. The multilinear extension has a probabilistic interpretation: if feature i is set with probability x_i, then $f(x)$ denotes the expected value of the function. Let $g(t) = f(t, t, \ldots, t)$ for $t \in [0, 1]$. Then $g(0) = F(\emptyset)$ and $g(1) = F(N)$. For a set $S \subseteq N$, with $|S| = j$, define

$$\Delta_S f(x) := \frac{\partial^j f}{\partial x_{i_1} \ldots \partial x_{i_j}} \quad \text{where } S = \{i_1, \ldots, i_j\}$$

Consider the $(k-1)^{\text{th}}$ order Taylor expansion of $g(\cdot)$ at 0 with Lagrange remainder and the corresponding multivariate expansion of each term in terms of $\Delta_S f(\cdot)$:

$$
\begin{aligned}
&F(N) - F(\emptyset) = g(1) - g(0) \\
&= \frac{g'(0)}{1!} + \ldots + \frac{g^{(k-1)}(0)}{(k-1)!} + \int_{t=0}^{1} \frac{(1-t)^{k-1}}{(k-1)!} g^{(k)}(t)dt \qquad (10) \\
&= \sum_{j \leq k-1} \sum_{|S|=j} \Delta_S f(0, \ldots, 0) + \\
&\qquad \sum_{|S|=k} \int_{t=0}^{1} k(1-t)^{k-1} \Delta_S f(t, \ldots, t)dt
\end{aligned}
$$

Theorem 3. *Let k be the order of explanation and $j < k$. Then j^{th} order Shapley-Taylor indices can be obtained from the j^{th} order terms in the Taylor series.*

$$\Delta_S f(0, \ldots, 0) = \mathcal{I}_S^k(F) \quad \textit{where} \ \ |S| < k$$

The k^{th} order Shapley-Taylor indices can be obtained from the Lagrange remainder term:

$$\int_{t=0}^{1} k(1-t)^{k-1} \Delta_S f(t, \ldots, t)dt = \mathcal{I}_S^k(F) \ \ \textit{where} \ |S| = k$$

We give the proof of the Theorem and Equation 10 in the appendix. For $k = 1$, we see that the definition of Shapley-Taylor indices in Equation 4 is exactly the Shapley value. Furthermore, Theorem 3 reduces to the result by (18) (Theorem 5) showing that the Shapley value can be obtained by integrating the gradient of f along the diagonal line.

4. Comparison of Shapley interaction indices and Shapley-Taylor indices

4.1. A Linear Model with Crosses

We illustrate the difference between Shapley-Taylor indices and Shapley interaction indices using a function that models linear regression with crosses. Consider the following function with 3 binary features: $F(x_1, x_2, x_3) := [x_1] + [x_2] + [x_3] + c[x_1] * [x_2] * [x_3]$; $[x_i]$ is 1 if feature x_i is present in the input, and zero otherwise. The last term of the function models a cross between the three features. This is a very simple function; we would expect the main effects to be 1 each, and the interaction effect of c to be divided equally among the three pairs.

Indeed, this is what we see with the Shapley-Taylor indices. The pairwise interaction terms for each of the three pairs is $c/3$, and the total interaction effect matches the coefficient of the cross term, c. The main effect terms for Shapley-Taylor indices are 1 each, as expected.

In contrast, those for Shapley interaction indices are $c/2$ each, and the total interaction effect is $3c/2$. Whether c is negative or positive, Shapley interaction indices *amplifies* the magnitude of interaction effect [6]. While Shapley interaction indices does not directly define main effect terms, a natural way to define them is to subtract the interaction terms from the Shapley values: $\Phi_{i,i} = \Phi_i - \frac{1}{2}\sum_{i \neq j} \Phi_{i,j}$, where Φ_i is the ith Shapley value and $\Phi_{i,j}$ is the Shapley interaction indices for $\{i, j\}$. With this definition, the main effect terms and interaction terms satisfy the efficiency axiom.

Using this definition, the main effect terms for Shapley interaction indices are $1 - \frac{c}{3}$ each. When c is larger than 3, the main effect term is negative, which does not match our expectation about the function.

In general, we observe that Shapley interaction indices returns inflated interaction values, possibly because it was not designed to satisfy efficiency, and a consequence of this is that the main effect terms, computed by subtracting the inflated interaction terms from the Shapley value, can have incorrect signs.

In general, there is no simple way to account for the amount of inflation in Shapley interaction indices. The following example shows that the size of the inflation varies with the size of the cross-terms present in the model: For example, for $F(x_1, \ldots, x_n) := [x_1] * \ldots * [x_n]$, the total interaction effect in Shapley-Taylor indices is 1, i.e., $1/\binom{n}{2}$ for each pair), while that for Shapley interaction indices is $n/2$ (i.e. $\frac{1}{n-1}$ for each pair). The inflation factor $n/2$ depends on the size of the interaction.

4.2. The Majority Function

Next, we consider the majority function F_{maj} over the set of features N is defined by $F_{maj}(S) = 1$ if $|S| \geq n/2$ and is 0 otherwise; here $n = |N|$.

It is easy to check that the singleton terms for pairwise Shapley-Taylor indices for a majority functions are uniformly zero (for $n > 2$). The pairwise terms are each equal to reciprocal of the number of pairs, $2/(n \cdot (n-1))$, a simple consequence of symmetry and efficiency. The analytical conclusion from this is that majority functions are all about interaction, and this is intuitively reasonable.

In contrast, the pairwise Shapley interaction indices are uniformly zero for every pair of features! This is unexpected; one would imagine that in a three feature majority function there would be non-zero interaction values for sets of size 2, the size at which majority arises. The results from studying pairwise Shapley interaction indices seems to suggest that there is *no* pairwise interaction! However, Shapley interaction indices of larger interaction sizes do have non-zero values. For instance, for a majority function with 3 features,

[6] If the function does not have crosses of size more than 2, the two methods coincide in the interaction terms.

the interaction value corresponding to all three features is -2, a non-intuitive number. The pattern becomes even more confusing for majority functions of more features.

Figure 2 (page 8) shows the sum of the Shapley interaction indices (in log scale) for all subsets of features for majority functions as number of features increases. This displays two non-intuitive characteristics. First, the total interaction diverges (recall the plot is in log scale) despite the fact that the function value being shared among the features is constant (1). Second, the *sign* of the total interaction alternates. In fact, for a function of a fixed size, *every* non-singleton interaction value has the same sign. (So the sign alternation is not due to cancellation across the interactions.) There is no intuitive reason for this sign alternation.

This discussion highlights the value of the Shapley-Taylor indices interaction distribution approach over that of Shapley interaction indices. To study interactions using Shapley interaction indices, one has to compare 2^n quantities, and even then the results appear unstable. In contrast, studying $O(n^2)$ quantities using the pairwise Shapley-Taylor indices gives us some insight into the functions behavior, no matter what the size of the set of interacting elements.

5. Applications

We study three tasks. The first scores sentences of text for sentiment. The model is a convolutional neural network from (11) trained over movie reviews[7]. The second is a random forest regression model built to predict house prices. We use the Boston house price dataset ((10)). The dataset has 12 numerical features and one binary feature.[8] The third model, called QANet (32), solves reading comprehension, i.e., identifying a span from a context paragraph as an answer to a question; it uses the SQuAD dataset ((20)). [9]

5.1. Insights

For the sentiment task, we ran Shapley-Taylor indices across a hundred test examples and identified word pairs with large interactions. Table 1 shows some of the interesting interactions we found.

The first example captures negation. The second example shows intensification; the effect of 'inspiring' is amplified by the 'both' that precedes it. The third example demonstrates a kind of intensification that would not typically appear in grammar texts; 'crisp' is intensified because it appears at the beginning of the review, which is captured by the interaction with 'A', whose main effect is nearly zero. The fourth examples shows complementarity; the interaction effect is of opposite sign to the main effects. The final example shows that sentiment expressed in third person is suppressed. This is natural because reviews are first-person artifacts.

Example sentence (interaction between bolded words)	interaction effect
Aficionados of the whodunit wo**n't** be **disappointed**	0.778
Watching these eccentrics is **both inspiring** and pure joy	0.1
A crisp psychological drama (and) a fascinating little thriller	0.48
With three **excellent** principal singers, a **youthful** and good-looking diva . . .	-0.224
Australian actor/director **John** Polson and . . . make a **terrific** effort . . .	-0.545

Table 1: Examples of different types of interactions. The interaction effect is the fraction of total change.

For the random forest regression task, we found that most of the influential pairwise feature interactions are substitutes; e.g. when the predicted price is lower than the average price, the main effects are negative, but the pair-wise interaction values are positive. We think this is because the features are correlated. We show a plot of main effects vs interaction effects of pairs of features in Figure 1.

For the reading comprehension task, previous analyses (cf. (16)) focused on whether the model ignored important question words. These analyses did not identify which paragraph words did the important question words match with. Our analysis identifies this. See the examples in Figure 3.

6. Discussion

In this section, we discuss issues of computational efficiency of Shapley-Taylor indices and future directions.

6.1. Computational Efficiency

The exact computation of Shapley-Taylor indices can be done using the formulas from Section 2 (Equations 5 and 6). Both the equations involve evaluating an exponential number of terms. This is similar to the exact computation of Shapley Values. To address this drawback, sampling-based approximations ((15)) have been used. The same method works for Shapley-Taylor indices. The basic idea is to evaluate Equation 5 over a sample of permutations.

[7] To ablate a word, we zero out the embedding for that word.

[8] There are 506 data points. We split the data into 385 training and 121 test examples. We used `scikit-learn` to train a random forest model. When we ablate a feature, we replace it by its training-data mean.

[9] We use Equation 6 to compute the Shapley-Taylor indices. However, the computational efficiency is $O(2^n)$, where n is the number of features. For the sentiment analysis and question answering tasks, n is large enough for this to be prohibitive. So we first use an attribution technique (Integrated Gradients (28)) to identify a small subset of influential features and then study interactions among that subset.

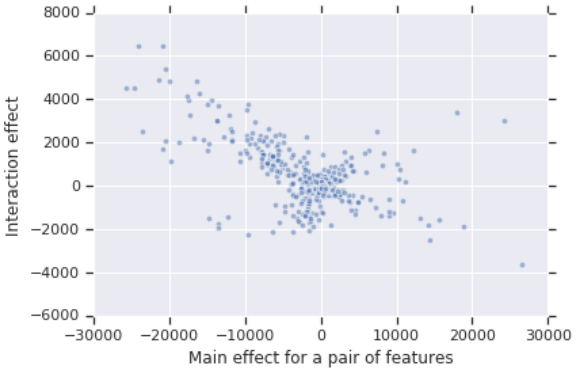

Figure 1: A plot of main effects of pairs of features vs interaction effect for the pair. Negative slope indicates substitutes.

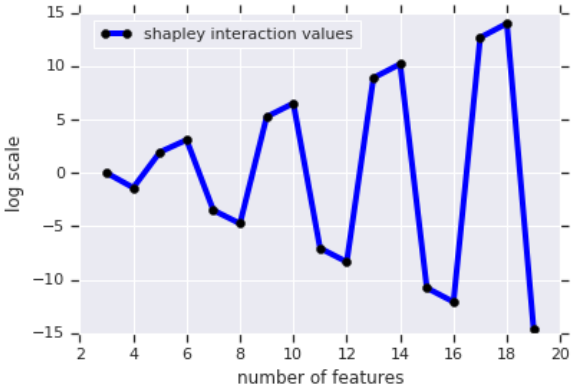

Figure 2: The sum of Shapley interaction indices (in log scale) for all subsets of features for majority functions as a function of the number of features.

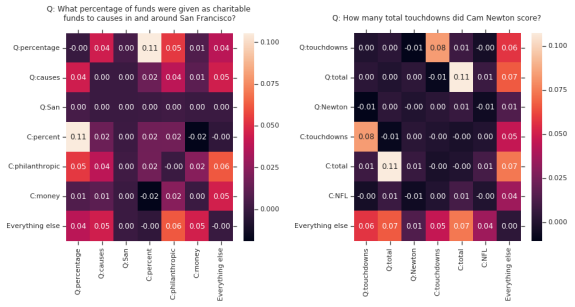

Figure 3: First example shows a match between *percentage* in question to *percent* in the context. Second example shows *total touchdowns* matching between the question and the context.

In our applications (Section 5), we used a different approximate method — first use a feature importance method to compute a subset of important features and compute feature importance over this subset.

Finally, it is possible that an approach like Kernel SHAP (cf. (14)) based on Kernelizing the Equation 6 can work. We leave this as a future direction to explore.

6.2. Continuous Extension

In this paper, we introduced Shapley-Taylor indices as a method to define interactions for functions over binary features. We also showed that the Shapley-Taylor indices are equivalent to the Taylor series terms of the multilinear extension of the original function (Theorem 3). This raises the question whether we can use the Taylor series to define interactions for continuous functions. For instance, pairwise interactions could be defined in terms of integral of the Hessian. This amounts to a generalization of the Integrated Gradients method cf. (28).

The main open question whether we can establish an equivalent of the Interaction distribution axiom to prove uniqueness of such a method. Another open question is about practical application – how well the method would work if the Hessian is not well-defined (e.g. due to presence of ReLUs).

ACKNOWLEDGEMENTS

We would like to thank Qiqi Yan and Amir Najmi for their feedback. We would also like to thank the anonymous reviewers for their suggestions to improve presentation.

References

[1] AUMANN, R. J., AND SHAPLEY, L. S. *Values of Non-Atomic Games*. Princeton University Press, Princeton, NJ, 1974.

[2] BAEHRENS, D., SCHROETER, T., HARMELING, S., KAWANABE, M., HANSEN, K., AND MÜLLER, K.-R. How to explain individual classification decisions. *Journal of Machine Learning Research 11* (2009), 1803–1831.

[3] BINDER, A., MONTAVON, G., LAPUSCHKIN, S., MÜLLER, K.-R., AND SAMEK, W. Layer-wise relevance propagation for neural networks with local renormalization layers. In *ICANN* (2016).

[4] COHEN, S., RUPPIN, E., AND DROR, G. Feature selection based on the shapley value. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (San Francisco, CA, USA, 2005), IJCAI'05, Morgan Kaufmann Publishers Inc., p. 665–670.

[5] CUI, T., MARTTINEN, P., AND KASKI, S. Recovering pairwise interactions using neural networks. *CoRR abs/1901.08361* (2019).

[6] DATTA, A., DATTA, A., PROCACCIA, A. D., AND ZICK, Y. Influence in classification via cooperative game theory. In *Proceedings of the 24th International Conference on Artificial Intelligence* (2015), IJCAI'15, AAAI Press, p. 511–517.

[7] DATTA, A., SEN, S., AND ZICK, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. *2016 IEEE Symposium on Security and Privacy (SP)* (2016), 598–617.

[8] GELMAN, A., ET AL. Analysis of variance—why it is more important than ever. *The annals of statistics 33*, 1 (2005), 1–53.

[9] GRABISCH, M., AND ROUBENS, M. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory 28* (11 1999), 547–565.

[10] HARRISON, D., AND RUBINFELD, D. L. Hedonic prices and the demand for clean air. *J. Environ. Economics & Management 5* (1978), 81–102.

[11] KIM, Y. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL* (2014), A. Moschitti, B. Pang, and W. Daelemans, Eds., ACL, pp. 1746–1751.

[12] LUNDBERG, S., AND LEE, S.-I. A unified approach to interpreting model predictions. In *NIPS* (2017).

[13] LUNDBERG, S. M., ERION, G. G., AND LEE, S. Consistent individualized feature attribution for tree ensembles. *CoRR abs/1802.03888* (2018).

[14] LUNDBERG, S. M., AND LEE, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 4768–4777.

[15] MALEKI, S., TRAN-THANH, L., HINES, G., RAHWAN, T., AND ROGERS, A. Bounding the estimation error of sampling-based shapley value approximation with/without stratifying. *ArXiv abs/1306.4265* (2013).

[16] MUDRAKARTA, P. K., TALY, A., SUNDARARAJAN, M., AND DHAMDHERE, K. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers* (2018), I. Gurevych and Y. Miyao, Eds., Association for Computational Linguistics, pp. 1896–1906.

[17] OWEN, A. B. Sobol' indices and Shapley value. *SIAM Journal of Uncertainty Quantification 2*, 1 (???? 2014), 245–251.

[18] OWEN, G. Multilinear extensions of games. *Management Science 18*, 5-part-2 (Jan. 1972), 64–79.

[19] PRECUP, D., AND TEH, Y. W., Eds. *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* (2017), vol. 70 of *Proceedings of Machine Learning Research*, PMLR.

[20] RAJPURKAR, P., ZHANG, J., LOPYREV, K., AND LIANG, P. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016* (2016), pp. 2383–2392.

[21] SHAPLEY, L. S. A value of n-person games. *Contributions to the Theory of Games* (1953), 307–317.

[22] SHRIKUMAR, A., GREENSIDE, P., AND KUNDAJE, A. Learning important features through propagating activation differences. In Precup and Teh (19), pp. 3145–3153.

[23] SIMONYAN, K., VEDALDI, A., AND ZISSERMAN, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR* (2013).

[24] SINGH, C., MURDOCH, W. J., AND YU, B. Hierarchical interpretations for neural network predictions. In *International Conference on Learning Representations* (2019).

[25] SLIWINSKI, J., STROBEL, M., AND ZICK, Y. Axiomatic characterization of data-driven influence measures for classification. In *The Thirty-Third AAAI Conference on Artificial Intelligence* (2019), pp. 718–725.

[26] SPRINGENBERG, J. T., DOSOVITSKIY, A., BROX, T., AND RIEDMILLER, M. A. Striving for simplicity: The all convolutional net. *CoRR* (2014).

[27] STRUMBELJ, E., AND KONONENKO, I. An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res. 11* (Mar. 2010), 1–18.

[28] SUNDARARAJAN, M., TALY, A., AND YAN, Q. Axiomatic attribution for deep networks. In Precup and Teh (19), pp. 3319–3328.

[29] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological) 58*, 1 (1996), 267–288.

[30] TSANG, M., CHENG, D., AND LIU, Y. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations* (2018).

[31] TSANG, M., LIU, H., PURUSHOTHAM, S., MURALI, P., AND LIU, Y. Neural interaction transparency (nit): Disentangling learned interactions for improved interpretability. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds. Curran Associates, Inc., 2018, pp. 5804–5813.

[32] YU, A. W., DOHAN, D., LE, Q., LUONG, T., ZHAO, R., AND CHEN, K. Fast and accurate reading comprehension by combining self-attention and convolution. In *International Conference on Learning Representations* (2018).

7. Appendix

7.1. Proof of Theorem 1

Proof. For a set $T \subseteq N$, define the *Möbius coefficients* as:

$$a(T) = \sum_{S \subseteq T} (-1)^{t-s} F(S), \qquad T \subseteq N$$

Decomposition of any function F into unanimity functions can be written in terms of $a(T)$'s as follows:

$$F(S) = \sum_{T \subseteq N} a(T) u_T(S)$$

Using the linearity axiom, we can extend $\mathcal{I}_S^k$ from unanimity functions to F as follows:

$$\mathcal{I}_S^k(v) = \sum_{T \subseteq N} a(T) \mathcal{I}_S^k(u_T)$$

Using the Shapley-Taylor indices for unanimity functions from Proposition 4, we get:

$$\begin{aligned}
\mathcal{I}_S^k(v) &= \sum_{T \subseteq N, S \subseteq T} a(T) \frac{1}{\binom{t}{k}} \\
&= \sum_{W \subseteq N \setminus S} a(W \cup S) \frac{1}{\binom{w+k}{k}} \qquad \text{where } W = T \setminus S \\
&= \sum_{W \subseteq N \setminus S} \sum_{U \subseteq W} (-1)^{u-w} \delta_S F(U) \frac{1}{\binom{w+k}{k}} \\
&\qquad \text{(using Lemma 2 below)} \\
&= \sum_{U \subseteq N \setminus S} \delta_S F(U) \sum_{W \supseteq U, W \subseteq N \setminus S} \frac{(-1)^{u-w}}{\binom{w+k}{k}}
\end{aligned}$$

Now we analyze the inner sum. We use the following identity: $\binom{w+k}{k} = 1/(k \cdot B(w+1,k))$, where $B()$ is the Beta function.

$$\begin{aligned}
&\sum_{U \subseteq W \subseteq N \setminus S} \frac{(-1)^{u-w}}{\binom{w+k}{k}} \\
&= \sum_{W \supseteq U, W \subseteq N \setminus S} (-1)^{u-w} k \cdot B(w+1,k) \\
&= k \sum_{w=u}^{n-k} \binom{n-k-u}{w-u} (-1)^{u-w} B(w+1,k) \\
&= k \sum_{w=u}^{n-k} \binom{n-k-u}{w-u} (-1)^{u-w} \int_0^1 x^w (1-x)^{k-1} dx \\
&= k \int_0^1 \sum_{w=u}^{n-k} \binom{n-k-u}{w-u} (-1)^{u-w} x^w (1-x)^{k-1} dx \\
&\qquad \text{(exchanging sum \& integral)} \\
&= k \int_0^1 x^u (1-x)^{k-1} \sum_{w'}^{n-k-u} \binom{n-k-u}{w'} (-1)^{w'} x^{w'} dx \\
&\qquad \text{(setting } w' = w-u\text{)} \\
&= k \int_0^1 x^u (1-x)^{k-1} (1-x)^{n-k-u} dx \\
&= kB(u+1, n-u) \qquad \text{(definition of Beta function)} \\
&= \frac{k}{n} \frac{1}{\binom{n-1}{u}}
\end{aligned}$$

We use this expression for the inner sum in the above equation to get:

$$\mathcal{I}_S^k(v) = \frac{k}{n} \sum_{U \subseteq N \setminus S} \delta_S F(U) \frac{1}{\binom{n-1}{u}}$$

This finishes the proof. □

The next Lemma provides a relation between the Möbius coefficients $a(T)$ and the discrete derivatives.

Lemma 2. *Möbius coefficients and discrete derivatives are related by following relation:*

$$a(T \cup S) = \sum_{W \subseteq T} (-1)^{t-w} \delta_S F(W)$$

for S and T such that $S \cap T = \emptyset$.

Proof. For two sets, S and T with $S \cap T = \emptyset$:

$$\begin{aligned} a(T \cup S) &= \sum_{W' \subseteq T \cup S} (-1)^{t+s-w'} F(W') \\ &= \sum_{W \subseteq T} \sum_{U \subseteq S} (-1)^{t+s-w-u} F(U \cup W) \\ &\quad \text{(where } W = W' \cap T \text{ and } U = W' \cap S) \\ &= \sum_{W \subseteq T} (-1)^{t-w} \sum_{U \subseteq S} (-1)^{s-u} F(U \cup W) \\ &= \sum_{W \subseteq T} (-1)^{t-w} \delta_S F(W) \end{aligned}$$

□

7.2. Proof of Theorem 3

Proof. Recall that $g(t) = f(t, \dots, t)$ for $t \in [0, 1]$. First, we derive the multivariate expression in Equation 10.

$$\frac{g^{(j)}(0)}{j!} = \sum_{S \subseteq N, |S|=j} \Delta_S f(0, \dots, 0)$$

Consider the expansion of $g^{(j)}(0)$:

$$\frac{g^{(j)}(0)}{j!} = \frac{1}{j!} \sum_{i_1} \sum_{i_2} \cdots \sum_{i_j} \frac{\partial^j f(0, \dots, 0)}{\partial x_{i_1} \dots \partial x_{i_j}}$$

Notice that f is a multilinear function. Therefore, only the mixed partial terms survive. Furthermore, all $j!$ mixed partials wrt $x_{i_1}, \dots, x_{i_j}$ are identical. Hence, we can simplify the above equation to:

$$\begin{aligned} \frac{g^{(j)}(0)}{j!} &= \sum_{i_1 < i_2 < \dots < i_j} \frac{\partial^j f(0, \dots, 0)}{\partial x_{i_1} \dots \partial x_{i_j}} \\ &= \sum_{S \subseteq N, |S|=j} \Delta_S f(0, \dots, 0) \end{aligned}$$

Similarly, consider the multivariate Lagrange remainder term in Equation 10:

$$\int_{t=0}^{1} \frac{(1-t)^{k-1}}{(k-1)!} g^{(k)}(t) dt$$

As before, in the derivative of g, only the mixed partial terms are left:

$$\frac{g^{(k)}(t)}{k!} = \sum_{S \subseteq N, |S|=k} \Delta_S f(t, \dots, t)$$

We use this expression in Lagrange remainder term and interchanging the order of integral and summation. Note that there is an extra factor of k that survives on the right side.

$$\begin{aligned} &\int_{t=0}^{1} \frac{(1-t)^{k-1}}{(k-1)!} g^{(k)}(t) dt = \\ &\sum_{S \subseteq N, |S|=k} \int_{t=0}^{1} k(1-t)^{k-1} \Delta_S f(t, \dots, t) dt \end{aligned} \tag{11}$$

For the rest of the proof of the theorem, we consider the special case unanimity functions u_W for a set $W \subseteq N$ and the corresponding multilinear extension:

$$f_W(x) = \prod_{i \in W} x_i \tag{12}$$

We prove the theorem for the unanimity functions. Since unanimity functions form the basis, the general case follows from linearity axiom.

Recall that a set $S \subseteq N$, $|S| = j$:

$$\Delta_S f_W(x) := \frac{\partial^j f_W}{\partial x_{i_1} \dots \partial x_{i_j}} \quad \text{where} \quad S = \{i_1, \dots, i_j\}$$

Using Equation 12, we get

$$\begin{aligned} \Delta_S f_W(x) &= \prod_{i \in W \setminus S} x_i \quad \text{iff } S \subseteq W \\ &= 1 \quad \text{when } S = W \text{ and } 0 \text{ otherwise.} \end{aligned}$$

Hence $\Delta_S f_W(0) = 1$ iff $S = W$ and 0 otherwise. This gives us:

$$\Delta_S f_W(0, \dots, 0) = \mathcal{I}_S^k(u_W) \tag{13}$$

This proves the result for $j < k$.

Next we analyze the Lagrange Remainder term. Consider a set $S \subseteq N, |S| = k$. We use Eqaution 12 to evaluate $\Delta_S f_W(t, \dots, t)$ using the fact that $\Delta_S f_W(\cdot) = 0$ if $S \not\subseteq W$:

$$\int_{t=0}^{1} k(1-t)^{k-1} \Delta_S f_W(t,\dots,t)dt \qquad (14)$$

$$= \int_{t=0}^{1} k(1-t)^{k-1} \left(\prod_{i \in W \setminus S} t \right) dt$$

$$= \int_{t=0}^{1} k(1-t)^{k-1} \left(t^{w-k} \right) dt \qquad |W| = w \text{ and } |S| = k$$

$$= k \int_{t=0}^{1} t^{w-k}(1-t)^{k-1} dt$$

$$= k \cdot B(w+1, k) \qquad B(\cdot,\cdot) \text{ is the Beta function}$$

$$= 1 / \binom{w}{k}$$

$$= \mathcal{I}_S^k(u_W) \qquad \text{from Equation 8}$$

This finishes the proof for the remainder term. □