High Dimensional Model Explanations: An Axiomatic Approach

Neel Patel, Martin Strobel and Yair Zick

Abstract

Complex black-box machine learning models are regularly used in critical decision-making domains. This has given rise to several calls for algorithmic explainability. Many explanation algorithms proposed in literature assign importance to each feature individually. However, such explanations fail to capture the joint effects of sets of features. Indeed, few works so far formally analyze *high dimensional model explanations*. In this paper, we propose a novel high dimension model explanation method that captures the joint effect of feature subsets.

We propose a new axiomatization for a generalization of the Banzhaf index; our method can also be thought of as an approximation of a black-box model by a higher-order polynomial. In other words, this work justifies the use of the generalized Banzhaf index as a model explanation by showing that it uniquely satisfies a set of natural desiderata and that it is the optimal local approximation of a black-box model.

Our empirical evaluation of our measure highlights how it manages to capture desirable behavior, whereas other measures that do not satisfy our axioms behave in an unpredictable manner.

1 Introduction

Machine learning models are applied in a variety of high-stakes domains, such as healthcare, insurance, credit decisions and more. In order to offer high prediction accuracy over high-dimensional data, one must adopt increasingly complex models. Complex models are, however, more difficult to understand. Explainability has been recently identified as an important design criterion, and has received widespread attention within the ML community; of particular note is research into generating *model explanations*. Broadly speaking, model explanations are based on a labeled dataset as well as other potential sources of information. Most model explanation techniques focus on feature-based explanations: for each feature i, the model explanation outputs a value ϕ_i which signifies the importance of the i-th feature in determining model predictions.

The value ϕ_i can be thought of as a score — 'Alice's high income is highly indicative of her receiving a loan' — or as a recourse — 'had Alice's income been lower by \$10,000/year her loan would have been rejected'. Either way, the basic premise of attribute-based model explanations is to explain complex model decisions via a list of n numerical values, where n is the number of data features. Crucially, this approach fails to capture *feature interactions*. Features are often strongly interdependent, especially in *complex ML models*. For example: consider a black-box model that predicts sentiments associated with a paragraph of text. In such texts; there can be a high negative interaction effect between "not" and "bad", which attribute-based model explanations will fail to capture: assigning influence to "bad" and "not" individually can be misleading. Consider the following review:

> It isn't the greatest scifi flick I've every seen, but it is not a
> bad movie

The above review is classified as a positive sentence. `LIME` Ribeiro et al. (2016) — a type of feature-based explanation — assigns high negative influence to the words "bad" and

"not". However, reviews omitting the word "bad" or "not" will be classified as negative. Explanation methods that indicate feature interactions, rather than just standalone scores, offer a much clearer picture of how the model makes its decisions.

The current model explanation landscape offers a wide variety of model explanation methods; how should we pick the right one? In this paper, we propose an *axiomatic approach*: rather than starting with a candidate solution and then offering post-hoc justifications for using it, start with the properties one would like to have in a high-dimensional model explanation, and derive a solution that satisfies these properties. This is the approach we take in this paper.

We axiomatize a high dimensional model explanation method, called the Banzhaf Interaction Index, or BII. This method faithfully captures how feature interactions influence model decision-making. Not only does our proposed model explanation satisfy a set of desirable properties, it is the *only* model explanation that satisfies all these properties.

1.1 Our Contributions

In this work, we propose a method for high-dimensional explanations for black-box models. Our main goal is to axiomatically capture higher-order feature interaction. Our main contribution is twofold: first, we propose a high-dimensional black-box model explanation method, axiomatically capturing higher-order feature interaction. Our characterization uses properties that are very natural in the ML context, such as feature symmetry and effect monotonicity (Section 3). we obtain a new axiomatization of the *Banzhaf interaction index* which uniquely satisfies symmetry, limit condition, Generalized-2-efficiency, and a newly proposed axiom (in the context of Banzhaf indices): monotonicity. Monotonicity is a rather general property which essentially means that the model explanation should change in a manner faithful to the underlying data. This is very fundamental property for an interaction measure: which states that the net contribution of the subset of features for the machine learning model f is more than that for the model g; then the interaction measure for those features for model f should be more than the interaction measure for those features for model g.

Second, we extend the idea of feature-based model explanations, which can be thought of as a local linear approximations of black-box models, to higher-order polynomial approximations. Especially, we show that our proposed measure can be obtained by approximating the black-box model by a higher-order polynomial (Section 4).

In the appendix, we evaluate our approach on real datasets. In particular, our experiments show that feature-based model explanations may fail to capture feature synergy; moreover, we show that other model explanations which fail to satisfy our axioms tend to behave in an undesirable manner (Appendix E).

1.2 Related Work

While some model explanations provide record-based explanations Koh and Liang (2017), or generate explanations from source code Datta et al. (2017), the bulk of the literature on the topic focuses on feature-based model explanations Ancona et al. (2017); Baehrens et al. (2010); Datta et al. (2015, 2016); Sliwinski et al. (2019); Sundararajan et al. (2017). Ancona et al. (2018) offer an overview of feature based model explanations for deep neural networks. The connection to *cooperative game theory* has been widely discussed and exploited in order to generate model explanations Ancona et al. (2017); Datta et al. (2015, 2016), with a particular focus on the Shapley value and its variants Shapley (1953).

Interaction Indices for Cooperative Games: Two widely accepted measures of marginal influence from cooperative game theory the Shapley value Shapley (1953) and the Banzhaf value Banzhaf III (1964), are uniquely derived from a set of natural axioms (see also Young (1985)). These measures do not capture player interactions; rather, they assign weights to individual players. Owen (1972) proposes the first higher-order solution for a cooperative game for pairwise players. Grabisch and Roubens (1999) extend it to interaction between arbitrary subsets, and axiomatically derive the Shapley and the Banzhaf interaction indices. In a recent paper, Agarwal et al. (2019) propose a new axiomatization for interaction among players which is inspired by the Taylor approximation of a Boolean function.

Feature Interactions: Feature interaction has been studied by different communities. Statistics offers classic ANOVA based feature interaction analysis Fisher (1992); Gelman (2005). Some recent work in the deep learning literature discusses feature interaction: Tsang et al. (2018) learns interactions by inspecting the inter-layer weight matrices of a neural network, Tsang et al. (2017) construct a generalized additive model that contains feature interaction information. In another line of work, Greenside et al. (2018); Cui et al. (2019) compute the interaction among features by computing the (expected) Hessian; this can be thought as an extension of gradient-based influence measures for neural networks Ancona et al. (2018). Datta et al. (2016) also propose an influence measure for a set of features called Set-QII. It essentially measures the change in model output when we randomly change a fixed set of features. Lundberg et al. (2018) propose the Shapley interaction index, specifically for tree-based models.

2   Preliminaries

We denote sets with capital letters $A, B, \ldots$ and use lowercase letters for functions, scalars, and features. To minimize notation clutter, we try to omit braces for singletons, pairs etc., e.g. we write $f(i), S \cup i$ instead of $f(\{i\}), S \cup \{i\}$ and $S \cup ij$ instead of $S \cup \{i, j\}$.

Let $N = \{1, \ldots, n\}$ be the set of *features*. A *black-box model* is a function mapping a set of n-dimensional input vectors $\mathcal{X} \subseteq \mathbb{R}^n$ to $\mathbb{R}$. Our objective is to generate a *model explanation* for a given *point of interest* (POI) $\vec{x} \in \mathbb{R}^n$; this explanation should (ideally) offer stakeholders some insight into the underlying decision-making process that ultimately resulted in the outcome they receive. In this work, we are interested in measuring the extent to which features, and *their high-order interactions* affect model decisions. In order to measure feature interaction effects, we adopt the *baseline comparison* approach Lundberg and Lee (2017); Sundararajan and Najmi (2019); in other words, we assume the existence of a *baseline vector* $\vec{x}'$, to which we compare an input vector $\vec{x}$, in order to generate a model explanation. For example, in the automatic loan acceptance/rejection domain $\vec{x}'$ could correspond to an all-zero vector (e.g. measuring the effect of an applicant having no money in their bank account, as opposed to the true amount they have), or a vector of mean values (e.g. comparing the applicant's true income to the average population income).

In order to generate model explanations, we need to formally reason about the effect of changing features in the POI $\vec{x}$ to their baseline values. Generally speaking, changing a single feature may have no significant effect on the model prediction. For example, if $f(income = 20k, debt = 90k) = 1$, it may well be the case that changing either the applicant's low income ($20k$) or their high debt ($90k$) would not result in them receiving the loan, however, it is unreasonable to claim that neither had an effect on the outcome. To formally reason about the joint effect of features, we define a function measuring their value as a set. Given a set of features $S \subseteq N$, a point of interest $\vec{x}$, a baseline $\vec{x}'$ and a model f, we define a set

function as

$$v(S, \vec{x}, \vec{x}', f) = f(\vec{x}_S, \vec{x}'_{N \setminus S}) - f(\vec{x}'). \tag{1}$$

In other words, the value we assign to a set of features S is the extent to which they cause the model prediction to deviate from the baseline prediction; as a sanity check, we note that $v(\emptyset, \vec{x}, \vec{x}', f) = 0$, and $v(N, \vec{x}, \vec{x}', f) = f(\vec{x}) - f(\vec{x}')$. This formulation induces a *cooperative game*, where features correspond to players. We refer to the game defined in (1) as the *feature effect game*. We omit $\vec{x}$, $\vec{x}'$ and f (they will normally be fixed), focusing solely on the set of features S.

We often replace sets of features with a single feature demarcating the entire set: given a set $T \subseteq N$, $[T]$ denotes a *single* feature corresponding to the set. The *reduced game* w.r.t. the nonempty $T \subseteq N$ is defined on the features $(N \setminus T) \cup [T]$ with the characteristic function $v_{[T]}(\cdot, f) : 2^{N \setminus T \cup \{[T]\}} \to \mathbb{R}$:

$$v_{[T]}(S) = \begin{cases} v(S \cup T) & [T] \in S \\ v(S) & \text{otherwise} \end{cases}$$

Our objective is to generate *high dimensional model explanations*, i.e. functions that assign a value to every subset of features $S \subseteq N$. To do so, we define a *feature interaction index* $S \subseteq N$ for v as $I^v(S)$. In other words, under the game v — namely the game defined in (1) — $I^v(S)$ should roughly capture the overall effect that the set of features S has on the value of v. Going back to the feature effect game defined in (1), $I^v(S)$ should measure the degree to which switching the value of features in S back to their baseline values affects the prediction for $\vec{x}$. A key idea in our analysis is *marginal effect*: consider a single feature $i \in N$. Its marginal effect on a set $T \subseteq N \setminus \{i\}$ equals $v(T \cup i) - v(T)$: the extent to which the value of $v(T)$ changes when i joins T. The marginal effect of i on T is denoted $m_i(T)$. In the context of the feature importance game, $m_i(T, v)$ is the marginal effect of knowing the feature i, given that the values of features in the set T are known. This is similar to other definitions considered in the literature Datta et al. (2016); Lundberg and Lee (2017). When considering a pair of features $i, j \in N$, how would one define their marginal effect on a coalition $T \subseteq N$? One very natural definition is to offset the effect of adding both features to T by the marginal effects of adding i and j separately, i.e.

$$m_{ij}(T, v) = (v(T \cup ij) - v(T)) - (v(T \cup i) - v(T)) - (v(T \cup j) - v(T))$$
$$= v(T \cup ij) - v(T \cup i) - v(T \cup j) + v(T)$$

We can define the marginal contribution of a general $S \subseteq N$, in a similar manner. Let $m_S(T, v)$ be:

$$m_S(T, v) = \sum_{L \subseteq S} (-1)^{|S| - |L|} v(T \cup L).$$

This is also known as the S *discrete derivative of v at T* (echoing the inclusion-exclusion principle). We note that when $T = \emptyset$, $m_S(\emptyset, v)$ is the *Harsanyi dividend* of S, a well-known measure of the synergy (or surplus) generated by S Peleg and Sudhölter (2007). For $T \neq \emptyset$, $m_S(T, v)$ can be thought of as the added value of having the coalition S form, given that the players in T have already committed to joining. More generally, $m_S(T, v)$ represents the marginal interaction between features in S within the set of features $T \cup S$.

3 Axiomatic Characterization of Good Model Explanations

As previously discussed, our objective is to identify *high-quality* model explanations. When pursuing quality metrics for a model explanation, one can take one of two approaches: either show that the model explanation is the optimal solution to some target (e.g. minimizes some loss function) Ribeiro et al. (2016); Sliwinski et al. (2019), or that it satisfies a set of desirable properties Datta et al. (2015, 2016); Sliwinski et al. (2019). We take both approaches in this work, resulting in a *unique* and *optimal* high-dimensional explanation; this is the *Banzhaf interaction index* (BII) Grabisch and Roubens (1999). Given a cooperative game v, the Banzhaf interaction index for a subset $S \subseteq N$ is

$$
\begin{aligned}
I_{\text{BII}}^v(S) &= \frac{1}{2^{n-|S|}} \sum_{T \subseteq N \setminus S} \sum_{L \subseteq S} (-1)^{|S|-|L|} v(L \cup T) \\
&= \frac{1}{2^{n-|S|}} \sum_{T \subseteq N \setminus S} m_S(T, v)
\end{aligned}
\tag{2}
$$

In words, $I_{\text{BII}}^v(S)$ equals S's expected marginal contribution to a set $T \subseteq N \setminus S$, sampled uniformly at random. Grabisch and Roubens (1999) show that BII uniquely satisfies *Linearity*, *Symmetry*, *Dummy*, the *Recursive Property*, *Generalized 2-Efficiency*, and the *Limit Condition*. Theorem 1 proposes a 'leaner' axiomatization of BII , which, as we argue, is more sensible in the model explanation setting. We show that BII is the unique measure which satisfies four natural axioms: *Symmetry*, *Generalized 2-Efficiency*, the *Limit Condition*, and *Monotonicity*. The first three axioms are fairly standard assumptions in identifying 'good' solutions, generalized to interaction indices by Grabisch and Roubens (1999).

Symmetry (S): for any permutation π over N we have that $I^v(S) = I^{\pi v}(\pi S)$. Here, πS equals $\{\pi(i) : i \in S\}$, and πv is the game where the value of a coalition T equals $v(\pi^{-1}T)$. Symmetry is a natural property for any interaction measure: intuitively, it simply stipulates that features' interaction value is independent of their identity, and depends only on their intrinsic coalitional worth.

Generalized 2-Efficiency (GE): for any $i, j \in N$, and for any $S \subseteq N \setminus ij$:

$$
I^{v_{[ij]}}(S \cup [ij]) = I^v(S \cup i) + I^v(S \cup j).
$$

Intuitively, merging two features into one feature encoding the same information results in no additional influence. Generalized 2-Efficiency extends the 2-Efficiency axiom proposed by Lehrer (1988) to characterize the Banzhaf value.

Limit Condition (L): if N is the set of players of the game v then $I^v(N) = m_N(\emptyset, v) = \sum_{L \subseteq N} (-1)^{n-|L|} v(L)$. In other words, the interaction value of the set N equals exactly the added value of it forming, given that no subsets of players have pre-committed themselves to joining.

Next, let us introduce the notion of monotonicity for interaction indices.

Monotonicity (M): If $\forall T \subseteq N \setminus S$, $m_S(T, v_1) \geq m_S(T, v_2)$ and for some $T^* \subseteq N \setminus S$ strict inequality holds then $I^{v_1}(S) > I^{v_2}(S)$. Moreover, if $\forall T \subseteq N \setminus S$, $m_S(T, v_1) = m_S(T, v_2)$ then $I^{v_1}(S) = I^{v_2}(S)$. This idea extends the strong monotonicity axiom proposed by Young (1985), which states that if $m_i(T, v_1) \geq m_i(T, v_2)$ for all $T \subseteq N \setminus i$ then $I^{v_1}(i) \geq I^{v_2}(i)$.

Datta et al. (2016) argue that the monotonicity axiom is better suited for charactering model explanations than the more 'standard' linearity axiom used in the classic characterization of the Shapley value Shapley (1953), as well as the original `BII` characterization Grabisch and Roubens (1999).

3.1   Characterization Result

First, we show that the four axioms we propose imply a generalized version of the dummy property in Lemma 1. The main result of this section is Theorem 1 that characterizes our explanation method uniquely.

Lemma 1. *If I^v satisfies (S), (GE), (L) and (M) then if $m_S(T, v) = 0 \ \forall T \subseteq N \setminus S$ then $I^v(S) = 0$.*

Proof. We begin by showing a weaker claim: if g is a *null game* where $g(S) = 0$ for all $S \subseteq N$, then $I^g(S) = 0$ for all $S \subseteq N$. For a given null game g, we have $m_S(T, g) = 0$ for all $S \subseteq N$ and $T \subseteq N \setminus S$. By symmetry, $I^g(S_1) = I^g(S_2)$ for all $S_1, S_2 \subseteq N$ with $|S_1| = |S_2|$, because for any permutation π, $g = \pi g$. Also, by the Limit Condition(L), $I^g(N) = m_N(\emptyset, g) = 0$. Similarly, for all $i_1 \neq i_2 \in N$, $I^{g_{[i_1, i_2]}}(N \setminus i_1 i_2 \cup [i_1, i_2]) = 0$. In fact, this property holds for all $V \subset N$: $I^{g_{[V]}}(N \setminus V \cup [V]) = 0$. Now we use (GE) property for $I^{g_{[V]}}(N \setminus V \cup [V])$ by sequentially removing all $k \in V$ until it becomes a singleton. First, for all $k_1 \in V$;

$$0 = I^{g_{[V]}}(N \setminus V \cup [V]) = I^{g_{[V \setminus k_1]}}(N \setminus V \cup [V \setminus k_1]) +$$
$$I^{g_{[V \setminus k_1]}}(N \setminus V \cup k_1) = 2 I^{g_{[V \setminus k_1]}}(N \setminus V \cup [V \setminus k_1])$$

The second equality holds because of symmetry (S) property for the game $g_{[V \setminus k_1]}$. Now for $k_2 \in V \setminus k_1$; we similarly use the (GE) property to obtain $I^{g_{[V \setminus k_1]}}(N \setminus V \cup [V \setminus k_1]) = 2 I^{g_{[V \setminus k_1]}}(N \setminus V \cup [V \setminus k_1 k_2])$. We repeat this argument until only one element is left. We get $0 = I^{g_{[V]}}(N \setminus V \cup [V]) = 2^{|V|-1} I^g(N \setminus V \cup k)$. This equality holds for all $V \subseteq N$ and all $k \in V$. Which shows that for all $S \subseteq N$, $I^g(S) = 0$.

Now, to prove the second part of the lemma, consider any game v. If $m_S(T, v) = 0$ for all $T \subseteq N \setminus S$, then $m_S(T, v) = m_S(T, g)$ for all $T \subseteq N \setminus S$, therefore by the Monotonicity property(M), $I^v(S) = I^g(S) = 0$, which concludes the proof. □

Before we prove the main theorem, we define important terms which will be helpful to prove the theorem.

Given a set of features $R \subseteq N$, we define the *primitive game* p^R as:

$$p^R(S) = \begin{cases} 1, & \text{if } R \subseteq S \\ 0, & \text{else} \end{cases}$$

The set $\mathcal{B}^N$ of all Boolean functions forms a *vector space*, and the set of primitive games $\mathcal{P}^N = \{p^R : R \subseteq N\}$ of all primitive games is an *orthonormal basis* of the vector space $\mathcal{B}^N$. Therefore, any Boolean function v (In this context a cooperative game v) is uniquely represented by a linear combination of primitive games. In other words, given a cooperative game $v : 2^N \to \mathbb{R}$, we can uniquely write as a linear combination of primitive games:

$$v(\cdot) = \sum_{R \subseteq N} c_R p^R(\cdot) \tag{3}$$

The unique decomposition of cooperative games is useful for our characterization of high-dimensional model explanations. We also require some technical claims which will be useful for our characterization result (for proofs see Appendix).

Theorem 1. *BII is the only high-dimensional model explanation satisfying (S),(GE),(L) and (M).*

Proof. We first show that BII satisfies all the properties. BII trivially satisfies (S), (L) and (M). To show that it satisfies (GE), take any $S \subseteq N \setminus ij$

$$
\begin{aligned}
I_{\texttt{BII}}^v (S \cup i) =& \frac{1}{2^{n-s-1}} \sum_{T \subseteq N \setminus (S \cup i)} m_{S \cup i}(T, v) \\
=& \frac{1}{2^{n-s-1}} \sum_{T \subseteq N \setminus (S \cup ij)} [m_S(T \cup i, v) - m_S(T, v)] \\
&+ \frac{1}{2^{n-s-1}} \sum_{T \subseteq N \setminus (S \cup ij)} [m_S(T \cup ij, v) - m_S(T \cup j, v)]
\end{aligned}
$$

A similar calculation for j shows that

$$
\begin{aligned}
I_{\texttt{BII}}^v (S \cup j) =& \frac{1}{2^{n-s-1}} \sum_{T \subseteq N \setminus (S \cup ij)} [m_S(T \cup j, v) - m_S(T, v)] \\
&+ \frac{1}{2^{n-s-1}} \sum_{T \subseteq N \setminus (S \cup ij)} [m_S(T \cup ij, v) - m_S(T \cup i, v)]
\end{aligned}
$$

Thus, $I_{\texttt{BII}}^v (S \cup i) + I_{\texttt{BII}}^v (S \cup j)$ equals

$$
\frac{1}{2^{n-s-2}} \times \sum_{T \subseteq N \setminus (S \cup ij)} [m_S(T \cup ij, f) - m_S(T, f)] \tag{4}
$$

Equation (4) shows that $I^{v_{[ij]}}(S \cup [ij]) = I^v(S \cup i) + I^v(S \cup j)$.

BII satisfies the four axioms; to show that it *uniquely* satisfies them, we use the fact that v can be uniquely expressed as the sum of primitive games;

$$
v = \sum_{R \subseteq N} C_R p^R \tag{5}
$$

We define the *index* Γ of a cooperative game v to be the minimum number of terms in the expression of the form (5). We prove the theorem by induction on Γ. For $\Gamma = 0$, in Lemma 1, $I^v(S) = 0$ for all $S \subseteq N$, which coincides with the Banzhaf interaction index.

If $\Gamma = 1$ then $v = C_R p^R$ for some $R \subseteq N$. Consider $S \not\subseteq R$; Proposition 1 in the Appendix implies that $m_S(T, v) = 0$ for all $T \subseteq N$, which in turn implies $I^v(S) = 0 = I_{\texttt{BII}}^v (S)$. By Lemma 2 in the Appendix, $I^v(R) = C_R$, which equals $I_{\texttt{BII}}^v (R)$ by Proposition 2 in the appendix. To complete the proof for the first inductive step, we need to show that for all $S \subseteq R$, $I^v(S) = I_{\texttt{BII}}^v (S) = \frac{C_R}{2^{|R|-|S|}}$. If $S_1, S_2 \subseteq R$ and $s_1 = s_2$ then by symmetry, $I^v(S_1) = I^v(S_2)$: we can define a permutation π over N such that S_2 bijectively maps to some S_1, and all $i \notin S_1 \cup S_2$: are invariant. By the Symmetry property $I^v(S_2) = I^{\pi v}(S_1)$; however, $\pi v = v$ because $v = C_R p^R$.

Now, consider any $i_1 \neq i_2 \in R$ and define $S := R \setminus \{i_1, i_2\}$; by the GE property, we can write

$$
I^{v_{[i_1 i_2]}}(S \cup [i_1 i_2]) = I^v(S \cup i_1) + I^v(S \cup i_2)
$$

which implies for any $Q \subset R$ with $|Q| = |R| - 1$, $I^v(Q) = \frac{1}{2} I^{v_{[i_1 i_2]}}(S \cup [i_1 i_2])$. The reduced game $v_{[i_1 i_2]}$ is

$$
v_{[i_1 i_2]}(S') = \begin{cases} C_R, & \text{if } S \cup [i_1 i_2] \subseteq S' \\ 0, & \text{else} \end{cases}
$$

By Lemma 2 in the Appendix, $I^{v_{[i_1 i_2]}}(S \cup [i_1 i_2]) = C_R$ and $I^v(Q) = \frac{C_R}{2}$. This property holds for all $T \subset N$, $v_{[T]}$; $I^{v_{[T]}}(N \setminus T \cup [T]) = C_R$. By inductively using the (GE) property, in a manner similar to Lemma 2 in the Appendix, we show that $I^v(Q) = \frac{C_R}{2^{|R|-|Q|}}$. By Proposition 2 in the Appendix, this coincides with Banzhaf interaction index concluding the first inductive step.

To complete the proof, assume that $I^v(S)$ coincides with the Banzhaf interaction index whenever the index of the game v is at most $\Gamma = \gamma$. Suppose that v has an index $\gamma + 1$, and expressed as

$$v = \sum_{k=1}^{\gamma+1} C_{R_k} p^{R_k}$$

Let $R = \bigcap_{k=0}^{\gamma+1} R_k$, and suppose that $S \nsubseteq R$. We define another game w:

$$w = \sum_{k:S \subseteq R_k} C_{R_k} p^{R_k}$$

Since $S \nsubseteq R$, the index of w is strictly smaller than $\gamma + 1$. We claim that for all $T \subseteq N \setminus S$; $m_S(T, v) = m_S(T, w)$. Indeed, consider any $T \subseteq N \setminus S$; $m_S(T, v)$ equals

$$\sum_{L \subseteq S} (-1)^{|S|-|L|} v(T \cup L) = \sum_{L \subseteq S} (-1)^{|S|-|L|} \sum_{k=1}^{\gamma+1} C_{R_k} p^{R_k}(T \cup L) =$$

$$\sum_{k=1}^{\gamma+1} \sum_{L \subseteq S} (-1)^{|S|-|L|} C_{R_k} p^{R_k}(T \cup L) = \sum_{k=1}^{\gamma+1} C_{R_k} m_S(T, p^{R_k}) =$$

$$\sum_{k:S \subseteq R_k} C_{R_k} m_S(T, p^{R_k}) = m_S(T, w)$$

The second-last equality holds by Proposition 1 in the Appendix, hence by induction on Γ and monotonicity(M) $I^v(S)$ coincides with BII for all $S \nsubseteq R$.

It remains to show that $I^v(S)$ coincides with BII when $S \subseteq R$. for any $S \subseteq R$, consider any $i \in S$ and define $S' := S \setminus i$. Take any $j \in N$ such that $j_1 \in R_1 \setminus R_2 \cup R_2 \setminus R_1$. By the (GE) property, we can write

$$I^v(S) = I^{v_{[ij_1]}}(S' \cup [ij_1]) - I^v(S' \cup j_1) \tag{6}$$

In Equation (6), $S' \cup j_1 \nsubseteq R$, therefore as previously shown, $I^v(S' \cup j_1)$ coincides with BII for the game v. Consider the restricted game $v_{[ij_1]}$:

$$v_{[ij]} = \sum_{k=0}^{\gamma+1} C_{R_k} p_{[ij_1]}^{R_k \setminus ij_1 \cup [ij_1]}$$

Consider $j_2 \neq j_1 \in R_1 \setminus R_2 \cup R_2 \setminus R_1$. By the (GE) property,

$$I^{v_{[ij_1 j_2]}}(S' \cup [ij_1 j_2]) = I^{v_{[ij_1]}}(S' \cup [ij_1]) + I^{v_{[ij_1]}}(S' \cup j_2) \tag{7}$$

In Equation (7), $S' \cup j_2 \nsubseteq \bigcap_{k=1}^{I+1} R_k \setminus ij_1 \cup [ij_1]$, therefore as we have shown before, $I^{v_{[ij_1]}}(S' \cup j_2) = I_{\text{BII}}^{v_{[ij_1]}}(S' \cup j_2)$. Let us denote $T = R_1 \setminus R_2 \cup R_2 \setminus R_1$ and $T' = i \cup T$. By repeating this

argument for all $j_3, \ldots, j_t \in T$ and exploiting the (GE) property for each j_ℓ, we can write $I^v(S)$ as:

$$I^v(S) = I^{v_{[T']}}(S' \cup [T']) - I^v(S' \cup j_1) - \sum_{l=1}^{t} I^{v_{[ij_1 \ldots j_\ell]}}(S' \cup j_\ell) \qquad (8)$$

All of the summands in (8) coincide with `BII`, because $S' \cup j_\ell \not\subseteq \bigcap_{k=1}^{I+1} R_k \setminus ij_1 \ldots j_{\ell-1} \cup [ij_1 \ldots j_{\ell-1}]$ for all $\ell = 1, \ldots, t$. We can write the reduced game $v_{[T']}$ as

$$v_{[T']} = (C_{R_1} + C_{R_2}) p_{[T']}^{(R_1 \cap R_2 \setminus i) \cup [T']} + \sum_{k=3}^{\gamma+1} C_{R_k} p_{[T']}^{(R_k \setminus T') \cup [T']}$$

Thus, the index of the reduced game $v_{[T']}$ is strictly smaller than $\gamma + 1$. By induction, $I^{v_{[T']}}(S' \cup [V])$ coincides with `BII` for the reduced game $v_{[T']}$. $I^v(S)$ can be written as

$$I^v(S) = I^{v_{[T']}}(S' \cup [T']) - \sum_{l=1}^{t} I^{v_{[ij_1 \ldots j_\ell]}}(S' \cup j_\ell)$$

By using the (GE) property inductively, $I_{\text{BII}}^v(S)$ can also be written in the same form, which implies that $I^v(S)$ coincides with the Banzhaf interaction index for all, $S \subseteq R$. $\square$

3.2 Explaining Our Model Explanations

Does the `BII` measure make sense in the model explanation domain? This is purely a function of the strength of the axioms we set forth. Symmetry is natural enough: if a model explanation depends on the indices of its features then it fails a basic validity test. The index in which a feature appears has no bearing on the underlying trained model (ideally), nor does it affect the outcome.

Recall that Generalized Efficiency requires that model explanations should be invariant under feature merging. In other words—artificially treating a pair of features as a single entity (while maintaining the same underlying model) should not have any effect on how feature behaviors are explained. Interestingly, Shapley values are not invariant under feature merging, a result shown by Lehrer (1988). The following examples illustrate what this entails in actual applications.

Example 1. *Consider an sentiment analysis task where a model predict if a movie review was positive. In a preliminary step the text is parsed by a parser to be machine readable. This can be done in many different ways. For example the sentence "This isn't an absolutely terrible movie" Can be parsed as*

$$| \textit{ This } | \textit{ isn't } | \textit{ an } | \textit{ absolutely } | \textit{ terrible } | \textit{ movie } | \textit{ . } |$$

or as

$$| \textit{ This } | \textit{ is } | \textit{ n't } | \textit{ an } | \textit{ absolut } | \textit{ ely } | \textit{ terrible } | \textit{ movie } | \textit{ . } |.$$

Generalized Efficiency ensures that the influences of "is" and "n't" in the second version add up to the influence of "isn't" in the first. In other words, Generalized Efficiency ensures that the influence of features generated through different parsers behaves in a sensible manner.

Example 2. *Features might be "merged" in another situation when features that were readily available during the training of a model end up being costly to obtain during its deployment. If additionally these features are highly correlated with other features they might just be coupled. E.g. generally birds can fly, so the features* IS_BIRD *and* CAN_FLY *may simply be merged at prediction time[1], to make it easier to enter information into a classifier. Again, Generalized Efficiency ensures that the influence of the merged feature relates in a natural way to the influence of the original features.*

The *Limit condition* normalizes the overall influence to be the discrete derivative of $v(\cdot, f)$ with respect to N. In other words, the total influence distributed to sets of features equals the total marginal effect of reverting features to their baseline values. This is an interesting departure from other efficiency measures. Shapley-based measures require efficiency with respect to $f(\vec{x})$ (or variants thereof), i.e. the total amount of influence should equal the total value the classifier takes at the point of interest (or the difference between the classifier and the baseline value). We require that the total influence equals the (discretized) rate in which features change the outcome. This makes BII more similar in spirit to gradient based model explanations, which are often used for generating model explanations in several application domains Simonyan et al. (2013).

Monotonicity is a very natural property in the model explanation domain: if a set of features has a greater effect on the value $f(\vec{x})$, this should be reflected in the amount of influence one attributes to it. This has already been established in prior works, for Shapley-based measures Datta et al. (2016). However, this property does not naturally generalize when using Shapley-based high-dimensional model explanations. Agarwal et al. (2019) propose a novel generalization of the Shapley value to high-dimensional model explanations, which fails monotonicity for smaller interactions (size of $< k$) for k-th order explanations, however, interactions of size k follow monotonicity. Example 3 highlights issues that may arise when monotonicity is not preserved.

Example 3. *Given a function $f_c(x_1, x_2, x_3) = c x_1 x_2 x_3$ with $c > 0$ defined on binary input space (for example, f is the result of an image classification task where x_i denotes the presence/absence of particular super-pixel). We assume that the baseline is $\vec{x}' = (0, 0, 0)$. Thus, $v(S, \vec{x}, \vec{x}', f_c) = 0$ if $\{2, 3\} \nsubseteq S$, resulting in $v(\{2, 3\}, f_c) = 0$ and $v(N, f_c) = v(\{1, 2, 3\}, f_c) = c x_1 x_2 x_3$. What is the interaction value between 1 and 2? Intuitively $\{1, 2\}$ offer some degree of interaction that monotonically grows as c increases. Moreover, it is easy to see that $v(\cdot, f_c) \geq v(\cdot, f_{c'})$ whenever $c \geq c'$. Set-QII fails to satisfies the monotoncity, and fails to capture the interaction between $\{1, 2\}$ for any c. Set-QII$(\{1, 2\}, S, f_c) = 0$ for all c.*

Similarly, the Shapley-Taylor interaction index for $S = \{1, 2\}$ and $k = 3$ is 0 as it does not follow the monotonicity property, however for $k = 2$ it satisfies monotonicity and interaction value for $\{1, 2\}$ is $\frac{c}{3}$. The BII value for $\{1, 2\}$ is c.

In Section 4, we show that BII can be interpreted as a polynomial approximation, offering additional intuition as to why our explanation method is *optimal*.

4 Geometrical Interaction and BII

The geometry of model explanations is relatively well understood for attribute-based methods Ribeiro et al. (2016); Sliwinski et al. (2019); Lundberg and Lee (2017); Ribeiro et al. (2016) view linear explanation methods as local linear approximations of $f(\cdot)$ around a point of interest $\vec{x}$. Linear model explanations can be thought of as functions taking the following

---

[1]The authors are aware of the existence of ostriches, emus, penguins and the fearsome cassowary.

form: $g(\vec{x}) = \phi_0 + \sum_{i=1}^{N} \phi_i x_i$, where ϕ_i captures the importance of feature i. Taking this interpretation, linear model explanations can be 'objectively' bad explanations: they are a poor local approximation of the underlying model f as black-box model can be highly non-linear around $\vec{x}$ Sliwinski et al. (2019); Tsang et al. (2017). In order to better capture the behavior of a black-box model f, we approximate it using higher-order polynomials. For better visualization, we first assume that the black-box model $f : \{0,1\}^N \to \mathbb{R}$ takes a binary input vector mainly referred to as the humanly understandable feature representation Lundberg and Lee (2017); Ribeiro et al. (2016). First, we start with quadratic approximation of the black-box model $f(\cdot)$,

$$g(\vec{x}) = \phi_0 + \sum_{i=1}^{N} \phi_i x_i + \sum_{i<j} \phi_{i,j} x_i x_j$$

In the above equation, $\phi_{i,j}$ captures the interaction between features i and j; ϕ_i, ϕ_j capture the importance of i and j. Thus, it is not unreasonable to assume that $\phi_{i,j}$ captures the pure interaction effect of i and j: we can delegate the singular effects to ϕ_i, having the resultant coefficient of $x_i x_j$ capture the 'pure' interaction between i and j. This also connects with the idea of the *Statistical Interactions* Fisher (1992); Gelman (2005). For instance, consider a sentiment analysis problem, both the tokens "bad" and "not" have negative influence on the prediction. However, when they are present together as "not bad", their influence is positive. In this simple example, it would be desirable to have $\phi_{\{\text{"not"}, \text{"bad"}\}} > 0$. The idea of higher order interactions can be extended similarly.

Consider a global polynomial approximation of $f(\cdot)$ by a k-degree polynomial:

$$g^k(\vec{x}) = \phi_0 + \sum_{S' \subset N : |S'| < k} \left(\phi_{S'} \prod_{j \in S'} x_j \right) + \sum_{S \subseteq N ; |S| = k} \left(\phi_S \prod_{j \in S} x_j \right). \qquad (9)$$

Again, to capture interaction among the set of features S such that $|S| = k$, we should remove all internal interaction effects captured by $\phi_{S'}$ for $S' \subset S$ as we intend to capture the The polynomial g^k is meant to locally approximate f around the POI $\vec{x}$; what is the best approximation? Finding the best fitting polynomial of the highest possible degree seems like a natural objective. However, taking this approach runs the risk of ignoring lower order feature interactions and their possible effects; this is shown in Example 4.

Example 4. *Consider the degree 3 polynomial studied in Example 3, $f(x_1, x_2, x_3) = cx_1 x_2 x_3$ with the baseline set to $(1,1,1)$ (rather than $(0,0,0)$ as was the case in Example 3). The best approximation to f is clearly itself. However, if we do so, then the interaction coefficients for variable pairs will be zero. This is arguably undesirable: for example, if $\vec{x} = (0,0,1)$, then x_3 has virtually no impact (it is already set at the baseline). Similarly, x_1 and x_2 have little individual effect, but do have significant joint effect - it is only when both are set to 1 that we observe any change in label.*

Now we formally define the optimization problem to find the "best" k-degree polynomial approximation of the black-box model $f(\cdot)$ globally. Let $\mathcal{P}^k$ be the set of k-degree polynomials. We are interested in finding a polynomial $g_f^k(\cdot)$ which globally minimizes the quadratic loss between $f(\vec{x})$ and $g(\vec{x}) \in \mathcal{P}^k$ for all $\vec{x} \in 2^N$, i.e.

$$g_f^k(\vec{x}) = \underset{g(\cdot) \in \mathcal{P}^k}{\arg\min} \sum_{\vec{x} \in 2^N} [f(\vec{x}) - g(\vec{x})]^2. \qquad (10)$$

The interaction among features in S with $|S| = k$ is measured as the coefficient of $\prod_{i \in S} x_i$ in the least square approximation of $f(\cdot)$ with a polynomial of degree k. Theorem 2 shows

that this geometrical definition of feature interaction coincides with the Banzhaf interaction index.

Theorem 2. *Let $g_f^k(\vec{x})$ be the k-degree solution of the optimization problem in Equation (10). Then the coefficients of $\prod_{i \in S} x_i$ for $|S| = k$ in $g_f^k(\vec{x})$ is given by $I_{BII}^v(S)$.*

The proof of the Theorem is a direct corollary of (Hammer and Holzman, 1992, Theorem 4.2). Theorem 2 shows that k-th order interaction obtained via least square approximation of the black-box model using k-th order polynomial coincides with `BII` and extend a geometrical argument for "good" feature-based explanation to feature interactions. Moreover, the least square approximation of the model via polynomial also resonates with the Statistical interaction among features as we already discussed. Therefore `BII` is not only obtained from strong axiomatic but also derived from an intuitive optimization problem.

5 Discussion and Future Work

Designing *provably sound* higher-order explanations for machine learning models in high stake domains is an important problem. Ideally, we want to design explanations we can trust. We offer a variety of reasons to trust in `BII` : it uniquely satisfies a set of natural properties — if we believe that these are sensible then our job is done. That said, the authors believe that a normative approach is critical in the design of fair, transparent AI solutions. Rather than debating approaches, one should debate the fundamental properties they are guaranteed to satisfy. Deriving model explanations from norms offers an intuitive justification for the chosen interaction measure, which fosters trust in the explanation method. As shown in Appendix E, some of our fundamental properties are absolutely critical for the design of sensible explanations. Even if one adheres to an optimization-based approach to explanation design, we show that `BII` is the optimal solution to a natural objective.

While `BII` certainly satisfies several design criteria, it is not without issues. The first challenge is computational. Feature-based explanations are computationally intensive as it is, so it should be no surprise that computing higher-order interactions comes at a higher cost. There are, however, good reasons to believe that `BII` can be computed efficiently on simpler ML model classes, and that low error approximations can be found quickly. This, we believe, is an important direction for future work.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 308–318.

Ashish Agarwal, Kedar Dhamdhere, and Mukund Sundararajan. 2019. A New Interaction Index inspired by the Taylor Series. *arXiv preprint arXiv:1902.05622* (2019). arXiv:1902.05622

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2017. A unified view of gradient-based attribution methods for deep neural networks. In *NIPS 2017-Workshop on Interpreting, Explaining and Visualizing Deep Learning*. 1–11.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. 1–16.

Yoram Bachrach, Evangelos Markakis, Ariel D Procaccia, Jeffrey S Rosenschein, and Amin Saberi. 2008. Approximating power indices. In *Proceedings of the 7th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 943–950.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11 (2010), 1803–1831.

John F Banzhaf III. 1964. Weighted voting doesn't work: A mathematical analysis. *Rutgers Law Review* 19 (1964), 317.

Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. 2011. Computational aspects of cooperative game theory. *Synthesis Lectures on Artificial Intelligence and Machine Learning* 5, 6 (2011), 1–168.

Tianyu Cui, Pekka Marttinen, and Samuel Kaski. 2019. Recovering Pairwise Interactions Using Neural Networks. *arXiv preprint arXiv:1901.08361* (2019). arXiv:1901.08361

Amit Datta, Anupam Datta, Ariel D Procaccia, and Yair Zick. 2015. Influence in classification via cooperative game theory. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*. 511–517.

Anupam Datta, Matthew Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. 2017. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 1193–1210.

Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *Proceedings of the 37th IEEE Symposium on Security and Privacy (Oakland)*. 598–617.

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

Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint arXiv:1702.08608* (2017).

Dheeru Dua and Casey Graff. 2019. UCI Machine Learning Repository. (2019). `http://archive.ics.uci.edu/ml`

Ronald Aylmer Fisher. 1992. Statistical methods for research workers. In *Breakthroughs in statistics*. Springer, 66–70.

Andrew Gelman. 2005. Analysis of variance—why it is more important than ever. *Annals of Statistics* 33, 1 (2005), 1–53.

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

Peyton Greenside, Tyler Shimko, Polly Fordyce, and Anshul Kundaje. 2018. Discovering epistatic feature interactions from neural network models of regulatory DNA sequences. *Bioinformatics* 34, 17 (2018), i629–i637.

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

Harmanpree Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 39th CHI Conference on Human Factors in Computing Systems (CHI)*. 1–14.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 1885–1894.

Ehud Lehrer. 1988. An axiomatization of the Banzhaf value. *International Journal of Game Theory* 17, 2 (1988), 89–99.

Zachary C. Lipton. 2016. The Mythos of Model Interpretability. *arXiv preprint arXiv:1606.03490* (2016). arXiv:1606.03490

Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. 2018. Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv preprint arXiv:1802.03888* (2018). arXiv:1802.03888

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS)*. 4765–4774.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (HLT)*. 142–150.

Smitha Milli, Ludwig Schmidt, Anca D. Dragan, and Moritz Hardt. 2019. Model Reconstruction from Model Explanations. In *Proceedings of the 2nd Conference on Fairness, Accountability, and Transparency (FAT*)*. 1–9.

Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *Proceedings of the 40th IEEE Symposium on Security and Privacy (Oakland)*. 739–753.

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

Neel Patel, Martin Strobel, and Yair Zick. [n. d.]. Online Supplementary Material: High Dimensional Model Explanation: An Axiomatic Approach. ([n. d.]). https://github.com/DataDrivenStrategicCollaborationGroup/High-Dimensional-Model-Explanations-An-Axiomatic-Approach

Bezalel Peleg and Peter Sudhölter. 2007. *Introduction to the Theory of Cooperative Games*. Springer.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (KDD)*. 1135–1144.

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

Reza Shokri, Martin Strobel, and Yair Zick. 2019. Privacy Risks of Explaining Machine Learning Models. *arXiv preprint arXiv:1907.00164* (2019). arXiv:1907.00164

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013). arXiv:1312.6034

Jakub Sliwinski, Martin Strobel, and Yair Zick. 2019. A Characterization of Monotone Influence Measures for Data Classification. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*. 718–725.

Mukund Sundararajan and Amir Najmi. 2019. The many Shapley values for model explanation. *arXiv preprint arXiv:1908.08474* (2019). arXiv:1908.08474

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 3319–3328.

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

Michael Tsang, Hanpeng Liu, Sanjay Purushotham, Pavankumar Murali, and Yan Liu. 2018. Neural interaction transparency (NIT): disentangling learned interactions for improved interpretability. In *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS)*. 5804–5813.

Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 38th CHI Conference on Human Factors in Computing Systems (CHI)*. 1–15.

H Peyton Young. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory* 14, 2 (1985), 65–72.

A Useful Technical Results

Proposition 1. *Given a primitive game p^R, for any $S \subseteq N$: if $S \not\subseteq R$ then $\forall T \subseteq N \setminus S$, $m_S(T, p^R) = 0$. In particular, $I_{BII}^{p^R}(S) = 0$.*

Proof. Suppose that $S \not\subseteq R$. We distinguish between two cases.

Case 1: $R \not\subseteq T \cup S$. In this case, $\forall L \subseteq S$, $T \cup L$ does not contain R, thus $p^R(T \cup L) = 0$ which implies $m_S(T, p^R) = 0$.

Case 2: $R \subseteq T \cup S$. In this case, $m_S(T, p^R)$ equals

$$\sum_{L \subseteq S} (-1)^{|S|-|L|} p^R(L \cup T) = \sum_{L \subseteq S : S \cap R \subseteq L} (-1)^{|S|-|L|} =$$
$$\sum_{L \subseteq S \setminus R} (-1)^{|S|-|S \cap R|-|L|} = \sum_{L \subseteq S \setminus R} (-1)^{|S \setminus R|-|L|} =$$
$$\sum_{k=0}^{|S \setminus R|} (-1)^k \binom{|S \setminus R|}{k} = 0$$

Thus in either case $m_S(T, p^R) = 0$, and we are done. $\qquad \square$

We note that Proposition 1 immediately holds for any game that is a scalar multiple of a primitive game, i.e. for any $v = c \times p^R$, and any $S \not\subseteq R$, $I_{\texttt{BII}}^v (S) = 0$.

Proposition 2. *Given a primitive game p^R, for any $S \subseteq N$: If $v = c \times p^R$, and $S \subseteq R$ then*

$$I_{BII}^v (S) = \frac{c}{2^{|R|-|S|}}.$$

Proof. Since $S \subseteq R$, then for any $L \subset$ and any $T \subseteq N \setminus S$, $v(L \cup T) = 0$. Therefore,

$$
\begin{aligned}
I_{\texttt{BII}}^v (S) &= \frac{1}{2^{n-|S|}} \sum_{T \subseteq N \setminus S} \sum_{L \subseteq S} (-1)^{|S|-|L|} v(L \cup T) \\
&= \frac{1}{2^{n-|S|}} \sum_{T \subseteq N \setminus S} v(S \cup T)
\end{aligned}
\tag{11}
$$

Next, if $T \subseteq N \setminus S$ does not contain $R \setminus S$, then $v(S \cup T) = 0$. Therefore, the summand in (11) equals

$$\sum_{T \subseteq N \setminus R} v(R \cup T) = c \times \sum_{T \subseteq N \setminus R} 1 = c \times 2^{n-|R|}$$

Plugging this back into (11), we obtain the desired result. $\qquad \square$

Next, let us characterize how influence measures satisfying our axioms behave on primitive games. Note that Lemma 2 offers a special case of Proposition 1 for *any* influence measure, rather than just for `BII` .

Lemma 2. *If I^v satisfies (S), (GE), (L) and (M) then for $v = c \times p^R$, $I^v(R) = c$*

Proof. We prove this lemma by inductively removing a feature $k \in N \setminus R$ and using the (GE) property at each step. Take any feature $i \in R$ and remove it from R and define $S := R \setminus \{i\}$. Now for any feature $j_1 \in N \setminus R \neq i$, by (GE) property we can write,

$$I^{v_{[ij_1]}}(S \cup [ij_1]) = I^v(S \cup i) + I^v(S \cup j_1)$$

Since $S \cup j_1 \not\subseteq R$, by Proposition 1, $m_{S \cup j_1}(T, v) = 0$ for all $T \subseteq N \setminus \{S \cup j_1\}$. Therefore by Lemma 1, $I^v(S \cup j_1) = 0$, which yields $I^v(R) = I^{v_{[ij_1]}}(S \cup [ij_1])$. We next remove $j_2 \neq j_1 \in N \setminus R$, and again invoke the (GE) property:

$$I^{v_{[ij_1j_2]}}(S \cup [ij_1j_2]) = I^{v_{[ij_1]}}(S \cup [ij_1]) + I^{v_{[ij_1]}}(S \cup j_2)$$

It is easy to check that $I^{v_{[ij_1]}}(S \cup j_2) = 0$ by Proposition 1 and Lemma 1 for the reduced game $v_{[ij_1]}$.

$$v_{[ij_1]}(S') = \begin{cases} c, & \text{if } S \cup [ij_1] \subseteq S' \\ 0, & \text{else} \end{cases}$$

$m_{S \cup j_2}(T, v_{[ij_1]}) = 0$ for any $T \subseteq \{N \setminus \{i, j_1\} \cup [ij_1]\} \setminus \{S \cup j_2\}$. This implies that $I^v(R) = I^{v_{[ij_1j_2]}}(S \cup [ij_1j_2])$. By repeating this argument for all $j \in N \setminus R$, we will have $I^v(R) = I^{v_{[N \setminus S]}}(S \cup [N \setminus S])$. We can write the reduced game $v_{[N \setminus S]}$ as

$$v_{[N \setminus S]}(S') = \begin{cases} c, & \text{if } S' = S \cup [N \setminus S] \\ 0, & \text{else.} \end{cases}$$

By the Limit (L) property,

$$I^{v_{[N \setminus S]}}(S \cup [N \setminus S]) = m_{S \cup [N \setminus S]}(\emptyset, v_{[N \setminus S]}) = c,$$

which concludes the proof $\qquad \square$

B   Additional Example

In the following example, we demonstrate that the Shapley interaction index can be misleading in simple situations. We consider the general majority classification function which exhibits pairwise feature interaction. Shapley interaction indices fail to capture these interactions. Moreover, Shapley-Taylor interaction indices fail to capture the sign of pairwise interactions for the same function.

Example 5. *Agarwal et al. (2019) Consider a classification function whose input space is binary. Let the classification function f be $f(x_1, \ldots, x_n) = 1$ iff $\sum_i x_i \geq k$ and 0 elsewhere with the baseline vector $\vec{x}' = \vec{0}$. Thus, $v(S, \vec{1}, \vec{x}') = 1$ iff $|S| \geq k$ and 0 otherwise. For $k = \frac{n}{2}$, the function coincides with the majority function discussed in the cooperative game theory literature. Clearly, there exists pairwise interaction among features, however, the Shapley interaction value for each pairwise feature is 0. In contrast, the pairwise Banzhaf interaction index for any feature pair $\{i, j\}$ is $c_k(2k - (n + 3))$ where $c_k > 0$. Pairwise interaction is negative when $2k < n + 3$. This can be explained by the following argument: the number of winning coalitions containing $\{i, j\}$ is $\binom{n-2}{k-2}$, and the number of winning coalitions that do not contain $\{i, j\}$ $\binom{n-2}{k}$, which is higher for smaller k. This shows that $\{i, j\}$ has more interaction effect for the output 0. On the other hand, the Shapley-Taylor pairwise interaction index is $\frac{2}{n(n-1)}$, and fails to capture the sign of the interaction index.*

C   Computational Complexity

Computing `BII` exactly — even for individual features, where it coincides with the Banzhaf index — is intractable (see (Chalkiadakis et al., 2011, Chapter 4) for a general overview); however, due to its probabilistic nature, `BII` can be well-approximated via random sampling for individual features Bachrach et al. (2008). The sampling approach is widely used for approximating cooperative game theoretic influence measures for the individual features Datta et al. (2016); Lundberg and Lee (2017).

To approximate the pairwise interaction for `BII`, we can decompose pairwise `BII` as following:

$$I_{\texttt{BII}}^{v}(ij) = \frac{1}{2^{n-2}} \sum_{T \subseteq N \setminus ij} (v(T \cup ij) - v(T \cup i) - v(T \cup j) + v(T))$$

$$= I_{\texttt{BII}}^{v_{[ij]}}([ij]) - I_{\texttt{BII}}^{v(N \setminus i)}(j) - I_{\texttt{BII}}^{v(N \setminus j)}(i) \tag{12}$$

This shows that there exists an efficient sampling scheme which approximates the pairwise `BII`: each term in Equation 12 can be approximated by polynomially many samples similar to the Datta et al. (2016); Lundberg and Lee (2017). Moreover, we further note that the approximation algorithms discussed for approximating Shapley interaction index in Lundberg et al. (2018) and polynomial algorithms for the tree-based decision model can be implemented for `BII` as well.

D   Broader Impact

The recent wave of research on explainable machine learning is motivated by the disparity between the success of machine learning, and its pronounced lack of public accountability. This work takes the next natural step towards a more general theory of model explanations. Our work treats explanation complexity as a *tunable parameter*: we can generate

explanations of varying degree of complexity, as is appropriate for different contexts and stakeholders. The authors view this as a *positive* effect on society.

Machine Learning models have been shown to potentially leak information about the training data Nasr et al. (2019) and there is some recent work that shows how explanations can be used to reconstruct models Milli et al. (2019) or recover user data Shokri et al. (2019). There is a growing body of evidence that explanations make models more prone to leak private training data; such leakage could undermine public trust in machine learning models, the very same trust that model explanations try to establish. It is only natural to assume that our high-dimensional model explanations are more vulnerable to adversarial exploitation, as they release more information per query. This is a potential societal *risk*. Advances in developing differentially private machine learning models Abadi et al. (2016) offer some hope that privacy preserving model explanations can be developed. In fact, QII is shown to be differentially private (with respect to the sample set) Datta et al. (2016).

Finally, our work offers several *mathematical* guarantees of explanation quality, but does not focus on its *actual adoptability*. There is currently little evidence that mathematically justified model explanations, as elegant as they may be, are acceptable by human stakeholders, or actually help humans understand model decisions. In fact, recent work suggests that model explanations can result in overconfidence of software developers in their understanding of models Kaur et al. (2020). We believe that an axiomatic approach, like the one chosen in the paper helps ensure that stakeholders understand the extent to which an explanation can be useful. This makes explanation methods such as ours more *trustworthy*. However, we also believe that before model explanation methods are deployed, the research community needs to clearly specify which tools can be used for which tasks, and that end users understand their limitations. Some parts of the community are already moving towards a more user-centric direction and believe that this can help overcome existing shortcomings Doshi-Velez and Kim (2017); Lipton (2016); Wang et al. (2019).

E Experimental Evaluation

In this section, we experimentally illustrate the need for high dimensional model explanations and the axioms discussed above. We consider two-well known benchmark datasets; the *UCI Adult dataset* Dua and Graff (2019) tries to predict whether an individual's income exceeds \$50,000/year based on census data, and the *IMDB movie review dataset* Maas et al. (2011) tries to predict the positive/negative sentiment of movie reviews based on their word content.

For the Adult dataset we train a 3-level decision three (see Figure 1) which achieves an 84% test accuracy. We use this easily comprehensible model to show the shortcomings of existing methods and the importance of our axioms. We also train a random forest with 50 estimators and a maximum depth of 10 that achieves a slightly higher accuracy of 86%[2].

For the IMDB dataset we train the BERT language model[3] Devlin et al. (2018), which achieves 89% evaluation accuracy; the BERT-Base model we use has 110M parameters, making it virtually incomprehensible to humans.

Like many other model explanation methods, `BII` assumes the existence of a baseline, i.e. a vector having "typical" values which can be used as a reference for the point of interest when creating the explanation. In the Adult dataset, we assume the baseline for continuous features to be their median value; categorical features are one-hot encoded, and the baseline is assumed to be zero. For the IMDB dataset, each word is a distinct feature, and the

---

[2]We use keras library to train decision tree and random forest `https://github.com/keras-team/keras`.
[3]We use pertained BERT model for the IMDB dataset available at `https://github.com/artemisart/bert-sentiment-IMDB`.
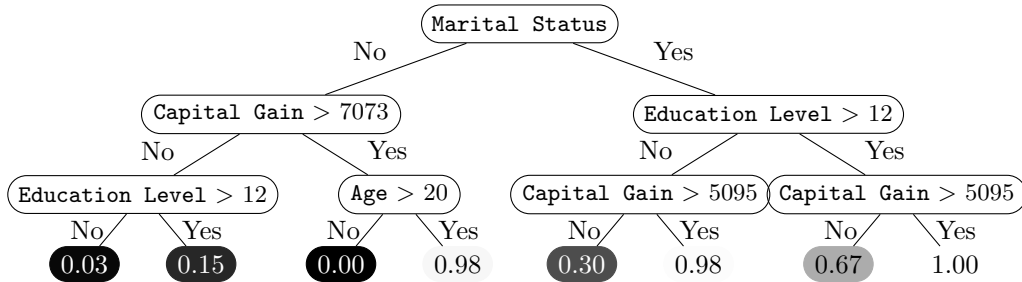
Figure 1: A three-level decision tree trained on the adult dataset, achieving 84% test accuracy.

baseline is an empty sentence (i.e. all features set to 0).

E.1 Baseline Explanation Frameworks

We compare `BII` with an interaction measure based on the Shapley interaction index (SII) Grabisch and Roubens (1999) and a metric that falls into the framework proposed in Datta et al. (2016) which we call Set-QII. For all methods we consider a version using the baseline i.e. the `BII` and SII values are the respective game-theoretic measures when "playing" the game of setting features to the baseline. Set-QII measures the influence of a (set of) features by measuring the difference of the function value on the point of interest and a single another point where the values of this (set of) feature(s) is set to the baseline. For all three metrics, we consider model explanations capturing single feature influences and pairwise feature interactions. We briefly discuss the computational complexity in the supplementary material Patel et al. ([n. d.]).

To further illustrate the importance of measuring interactions we also study the well known `LIME` method Ribeiro et al. (2016) which only generates individual feature importance.

We also consider a naive approach to create feature interactions which we call the *additive expansion*. The additive expansion of a feature-based explanation methods is computing joint influence of feature i and j (interaction among features i and j) by simply adding individual influence of features i and j, $I(ij) = I(i) + I(j)$. This is a naive way to compute the joint effect of features when only individual feature importance is provided. Via additive expansion we can create a feature-interaction measure from `LIME`, as well as ad-hoc feature-interaction measures from the feature-importance versions of Banzhaf, Shapley and QII.

E.2 The Importance of Interactions

While seemingly obvious, we begin our empirical investigation with a simple question: are there instances where high-dimensional model explanations are needed? In other words, is it possible that feature-wise explanations offer sufficient explanations in practice? Our first experiment indicates that high-dimensional explanations matter. In Figure 2, we visualize interactions for the movie review described in the Introduction (Section 1). We compute pairwise interactions using `BII` among features for the BERT model. We represent the explanations generated by `BII` as a symmetric matrix, where the i-th diagonal element represents the influence of feature i in the decision tree and (i, j)-th element of the matrix for $i < j$ represents the interaction between i and j.

Comparing Figures 2 and 3, it is immediately clear why feature importance offers additional insights. The values generated by `LIME` are indicative of a negative review (whereas
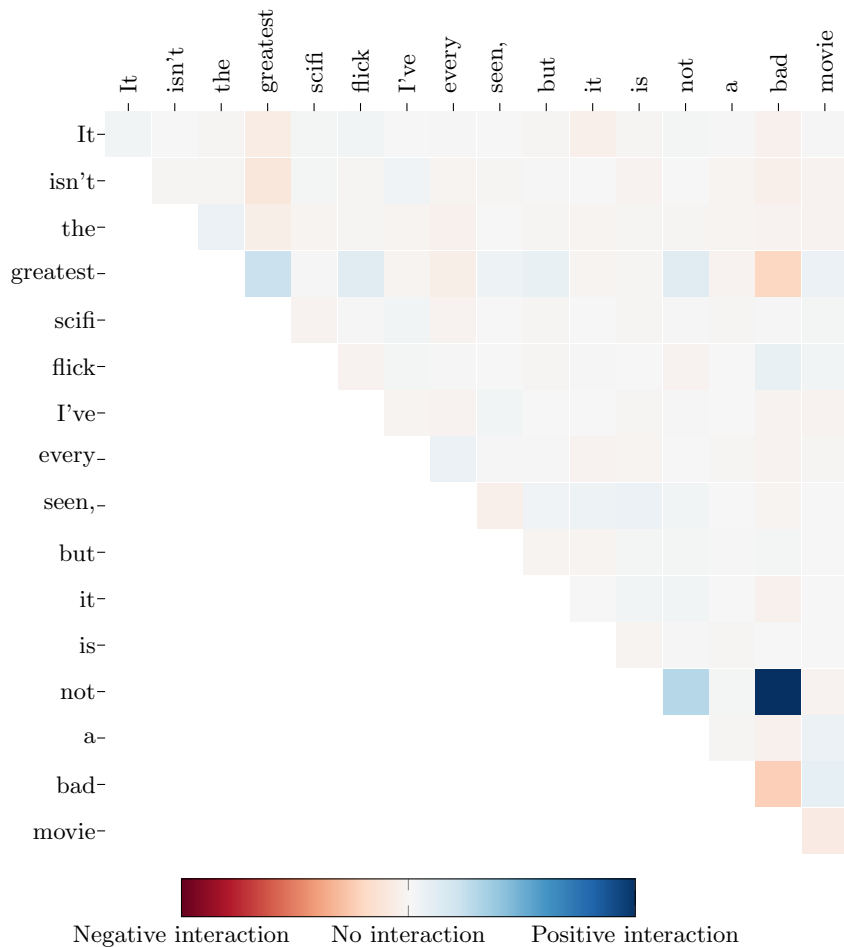
Figure 2: Feature importance and interactions for the movie review given in the introduction generated via `BII`. `BII` is able to pick up on the strong positive interaction between "not" and "bad" that leads to a positive prediction for this review. Note that the individual importance of "not" and "bad" is rather weak - their synergy is assigned importance.

the review we examine is positive), and look somewhat odd: 'not' has a strong positive influence. On the other hand, the interactions make it clear that BERT has picked up on the fact that 'not bad' indicates a good review. We compare additional `BII` feature interactions for the BERT model and the IMDB classification task in the supplementary material Patel et al. ([n. d.]) (Figures 1 to 5). `LIME` and `BII` mostly agree for the individual feature importance; however, `BII` captures non-trivial interactions among pairs of words that are impossible to observe only from feature-based explanation.

In Figure 4, we plot the explanation generated by `BII` and `LIME` for individuals with `Marital Status` =yes, `Capital Gain` $\in \{5095, 7073\}$ and `Education Level` < 12 from the adult dataset for the simple tree classifier (see Figure 1). `LIME` suggests that `Capital Gain` is the most important feature and the other two features have minor importance. However, for any coalition of features (activated features are set to their POI value, and the others to their baseline value: `Marital Status` = No and `Capital Gain` = 0), changing only `Marital Status` or `Capital Gain` to their POI value does not significantly change the
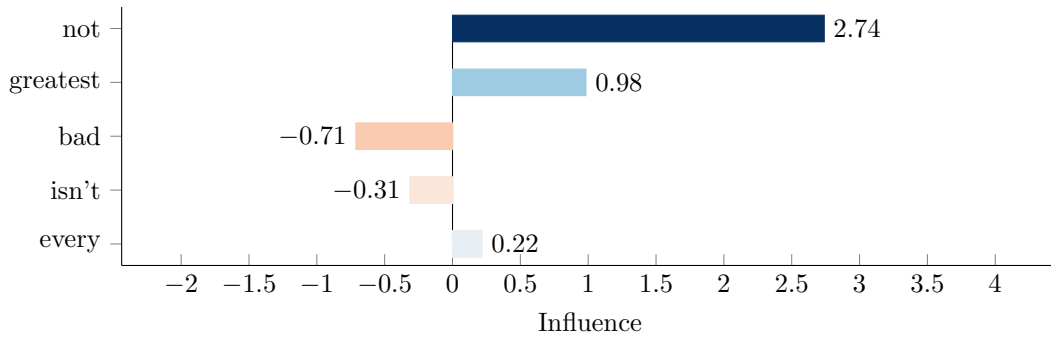
Figure 3: `LIME` feature importance for the movie review given in Section 1. The measure assigns significant weight to the word "not", but little weight to "bad".
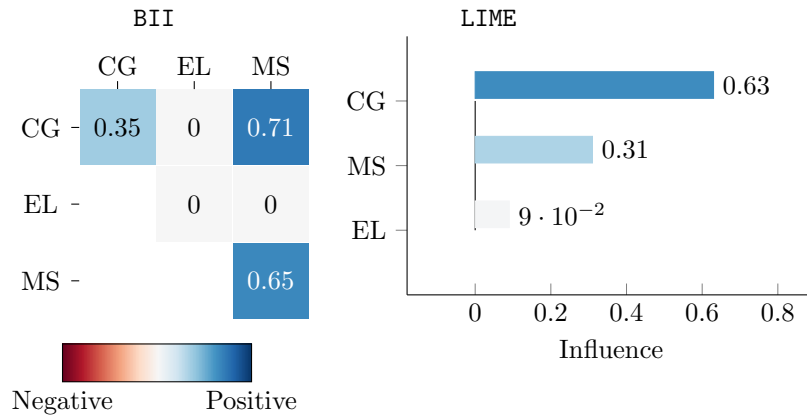

Figure 4: The explanation generated by `BII` and `LIME` for another point in the dataset. `LIME` fails to capture interactions.

prediction value; but, changing both of them concurrently to their POI value results in a significant increase in prediction probability (see Table 1). `BII` successfully captures these insights in Figure 4, and suggests that the high positive interaction between `Capital Gain` and `Marital Status` is responsible for the positive label. When considering *feature interaction*, there is high positive interaction between `Capital Gain` and `Marital Status`, an insight that linear explanations do not capture, as seen in Table 1.

Figure 5 summarizes the explanations generated by `BII` , `SII` , `LIME` and `QII` for a married woman with a `Capital Gain` (CG) of 5178 with `Education Level` = 13 in the adult test data for the simple tree classifier (Figure 1). Column (a) is the additive expansion of the feature-importance versions of these measures and Column (b) shows the methods that explicitly consider interactions. Both `BII` and `SII` indicate a negative synergy between `Education Level` and `Capital Gain`, whereas all methods indicate positive individual importance. Hence, additive expansions fail to capture negative interactions, even for relatively simple models.
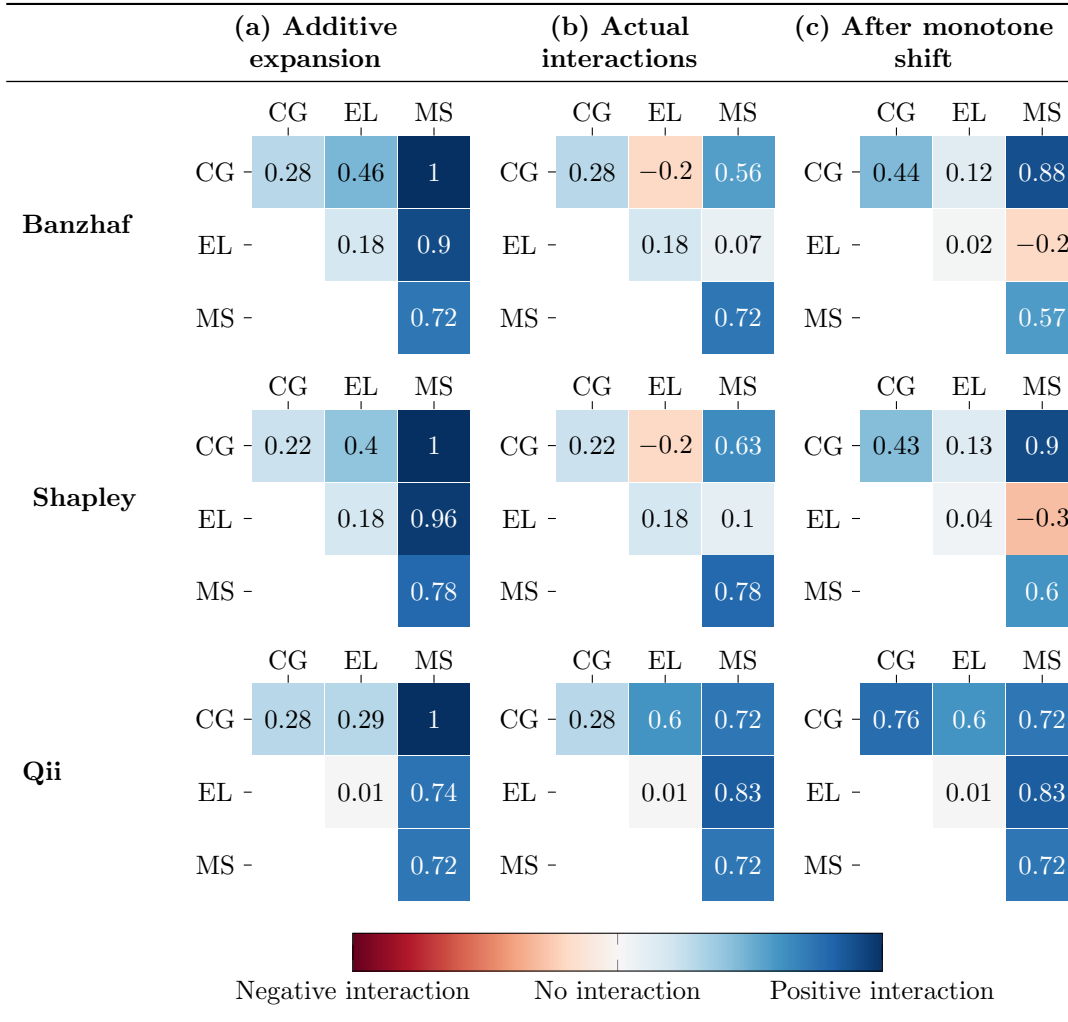
(a) Additive expansion | **(b) Actual interactions** | **(c) After monotone shift**

Banzhaf

	CG	EL	MS
CG	0.28	0.46	1
EL		0.18	0.9
MS			0.72

	CG	EL	MS
CG	0.28	−0.2	0.56
EL		0.18	0.07
MS			0.72

	CG	EL	MS
CG	0.44	0.12	0.88
EL		0.02	−0.2
MS			0.57

Shapley

	CG	EL	MS
CG	0.22	0.4	1
EL		0.18	0.96
MS			0.78

	CG	EL	MS
CG	0.22	−0.2	0.63
EL		0.18	0.1
MS			0.78

	CG	EL	MS
CG	0.43	0.13	0.9
EL		0.04	−0.3
MS			0.6

Qii

	CG	EL	MS
CG	0.28	0.29	1
EL		0.01	0.74
MS			0.72

	CG	EL	MS
CG	0.28	0.6	0.72
EL		0.01	0.83
MS			0.72

	CG	EL	MS
CG	0.76	0.6	0.72
EL		0.01	0.83
MS			0.72

Negative interaction — No interaction — Positive interaction

Figure 5: The explanations for a married woman with a `Capital Gain` (CG) of 5178 with `Education Level` = 13 (i.e. Masters) in the adult test data for the above tree. (a) Linear explanations highlight the importance of `Marital Status` (MS) and `Capital Gain` (CG), they fail to show how these to features interact. (b) The positive interaction between `Capital Gain` (CG) and `Marital Status` (MS) highlights that both features a needed for a positive outcome. (c) After the tree gets modified to strengthen the interaction, the explanation proposed in Set-QII does not change, failing monotonicity.

Table 1: This example shows the joint importance of the interactions between `Capital Gain` and `Marital Status` in the coalition of {`Age`, `Education Level`}, and the monotonicity property in the tree T (Figure 1). The prediction probability becomes 1.00 when `Marital Status` and `Capital Gain` both join the coalition {`Age`, `Education Level`}; however, when the `Marital Status` or `Capital Gain` join individually in the coalition {`Age`, `Education Level`}, the prediction probability becomes 0.67 and 0.15 respectively. This shows that the features have synergistic effect, which is not captured by linear explanations (this holds for other coalitions, not just {`Age`, `Education Level`}).

Features	POI	Base	Coalition			
			{A,EL}	{A,EL,MS}	{A,EL,CG}	{A,EL,MS,CG}
`Marital Status` (MS)	Yes	No	No	Yes	No	Yes
`Capital Gain` (CG)	6021	0	0	0	6021	6021
`Age` (A)	32	37	32	32	32	32
`Education Level` (EL)	13	10	13	13	13	13
Prediction	1.00	0.03	0.15	0.67	0.15	1.00

E.3 Monotonicity Matters

Having established the importance of measuring feature synergy, let us turn to evaluating the importance of our proposed axioms. To do so, we consider a modified version of the tree depicted in Figure 1. The output in the second-to-last leaf is changed from 0.67 to 0.01, i.e. married people with high `Education Level` but low `Capital Gain` are now classified as not having a high paying job. This change does not affect the prediction of the two points discussed above. However, this increases the interaction between `Capital Gain` and `Marital Status`: having a lower `Capital Gain` (or being unmarried) both lead to a negative outcome. More formally, we denote the cooperative game induced by the original tree by f and the one induced by the modified tree by f'. Under f', the joint marginal gains increase:

$$m_{\mathrm{MS,CG}}(\{\texttt{Age}, \texttt{Education Level}\}, f') >$$
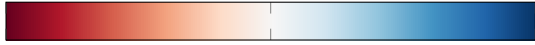$$m_{\mathrm{MS,CG}}(\{\texttt{Age}, \texttt{Education Level}\}, f),$$

and $m_{\mathrm{MS,CG}}(T, f') = m_{\mathrm{MS,CG}}(T, f)$ for any other coalition T. According to the *monotonicity* property, the interaction index value among {`Age`, `Education Level`} should be higher under f' than f. This change is reflected in Shapley interactions and `BII` as both measures satisfy monotonicity (see Figure 5, column (b) for actual interaction and column (c) for interaction after the modification). Set-QII, however, is unchanged; it fails to satisfy monotonicity, and even for this small example cannot distinguish between the two trees. Similarly, the interaction between `Capital Gain` and `Education Number` should also increase which is again reflected by `SII` and `BII` ; Set-QII fails to show this (see Figure 5).

E.4 The Effects of 2-Efficiency

Table 2 highlights how failure to satisfy generalized 2-efficiency can lead to counterintuitive interactions when features are merged. We calculated the `SII` values for `Education Level` and `Education Number` for a random forest with `Marital Status` (left) and `Capital Gain` (right) on some sampled points; we recalculated them after merging the features to [`Education Level`; `Education Number`]. These features can be naturally merged as they are in fact identical: `Education Number` is just a numerical representation of `Education Level`.

Table 2: Failure to satisfy generalized 2-efficiency leads to counterintuitive interactions after features are merged. We merge `Education Level` and `Education Number` (which are actually identical); Shapley interaction values with `Marital Status` (left) and `Capital Gain` (right) before and after the features are merged tend to overestimate the expected merged value (as compared to the expected value under generalised 2-efficiency).

	Interaction with `Marital Status`					Interaction with `Capital Gain`				
	Pt. 1	Pt. 2	Pt. 3	Pt. 4	Pt. 5	Pt. 1	Pt. 2	Pt. 3	Pt. 4	Pt. 5
Education Level (EL)	-.068	-.032	-.028	-.066	-.044	-.017	0	.016	.015	-.002
Education Number (EN)	.05	.123	.118	.041	.17	.016	-.007	-.009	-.006	-.011
[EL; EN]	-.077	.161	.154	-.077	.23	.013	-.013	.017	.019	.014
Expected [EL;EN]	-.018	.091	.090	-.025	.126	-.001	-.007	.007	.009	.009
Difference	-.059	.070	.064	-.052	-.103	.014	-.006	.010	.010	.005

Negative interaction   No interaction   Positive interaction

In some instances, the interaction values of `Education Level` and `Education Number` with other features differ in sign (i.e. one has positive interaction and another has a negative interaction). In this case, under 2-efficiency, the interaction value of the merged feature [`Education Level`; `Education Number`] should lie somewhere between the two. In Table 2, we show some of the points where the Shapley interaction of `Marital Status` and `Capital Gain` with merged feature deviates from the expected interactions. For example, for Point 2, the SII of `Education Level` with `Marital Status` is -0.32 and `Education Number` with `Marital Status` is 0.123, therefore the interaction after merging `Education Number` and `Education Level` should be less than 0.123; however, the SII interaction value between `Marital Status` and the merged feature is 0.161 — more than 0.123! In other words, SII placed far too much influence on two identical merged features. We show additional examples in Table 2. Since BII satisfies generalized 2-efficiency, influence is preserved under merging and thus matches the expected outcome.

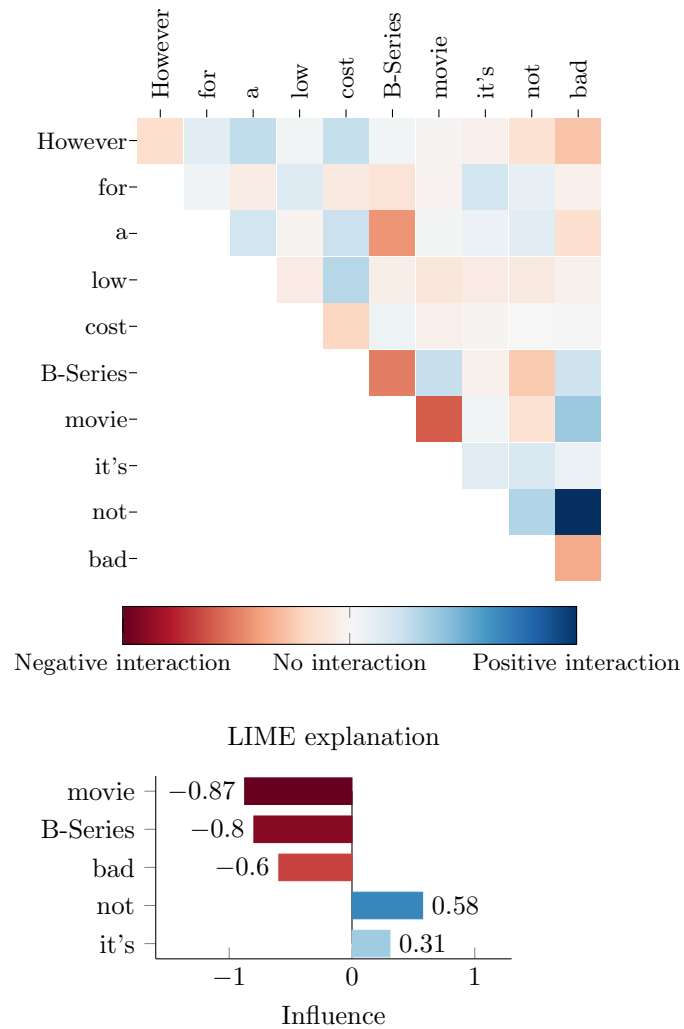# F Additional Example Explanations for BERT

Figure 6: While most individual features have negative influence for both `BII` and `LIME`, `BII` is able to pick up on the strong positive interaction between "not" and "bad" that leads to a positive prediction of this review.
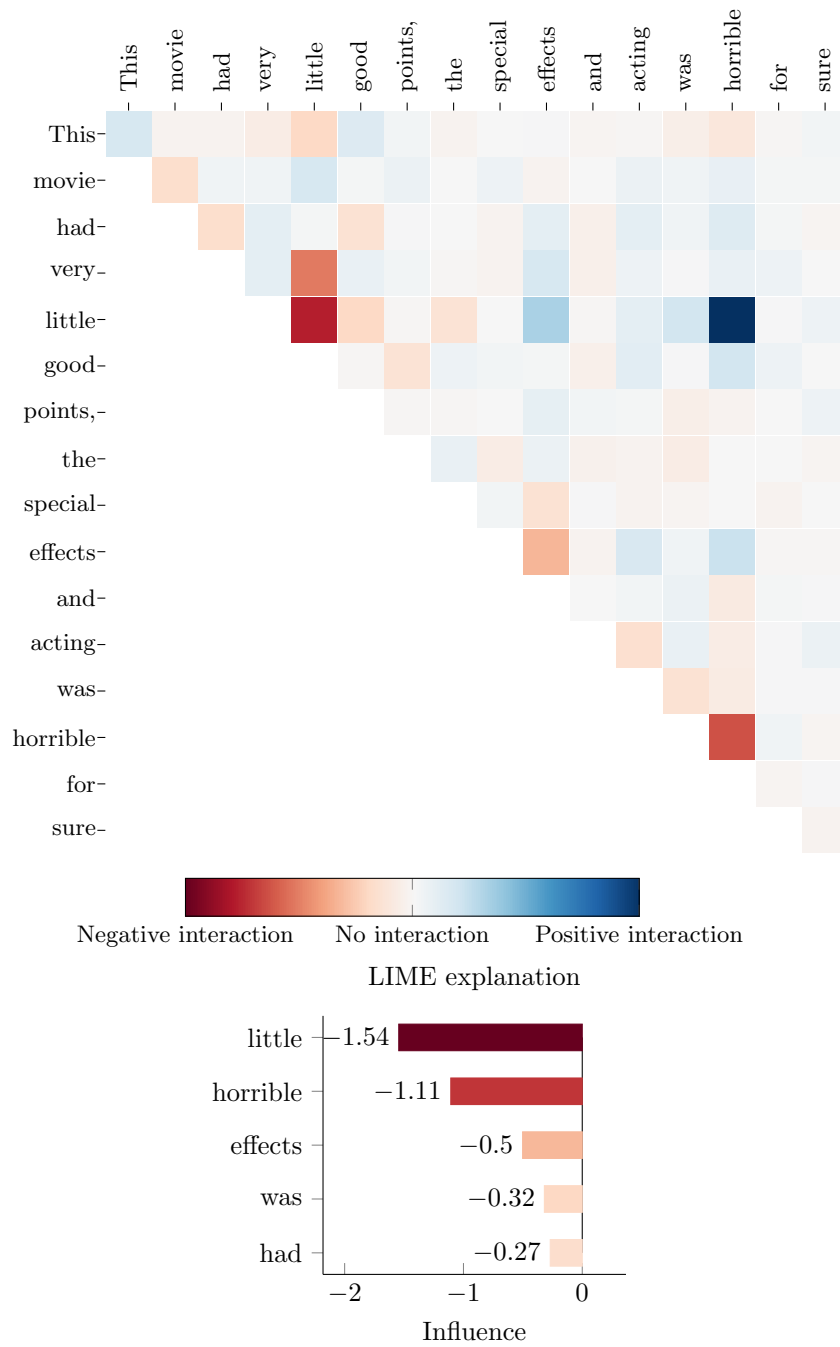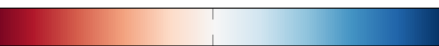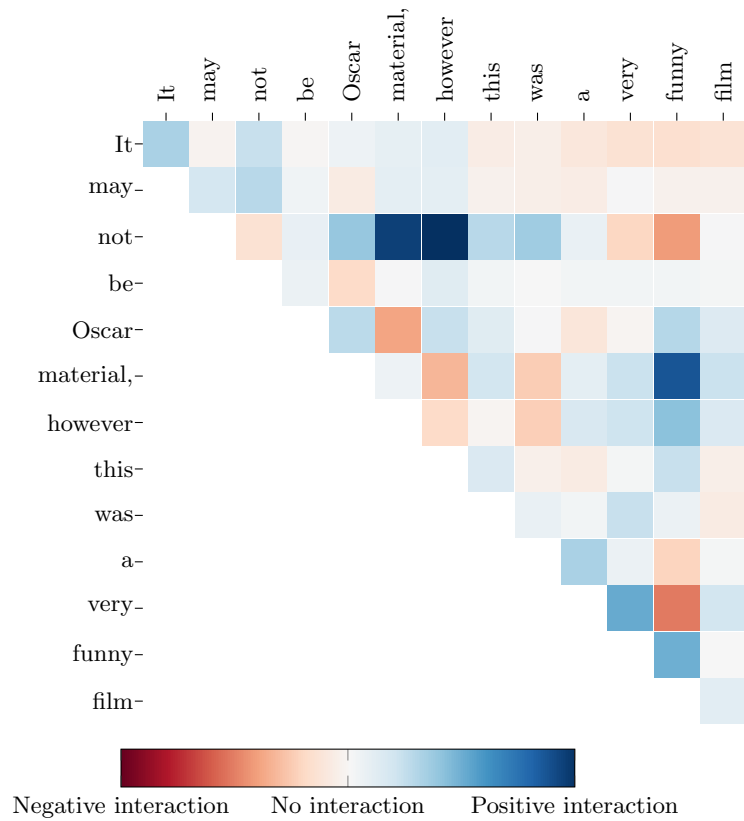
Figure 7: While most individual features have negative influence suggested by `BII` and `LIME`, `BII` is able to pick up on the strong positive interaction between "little" and "horrible" that leads to less confidence on the negative prediction of the review.
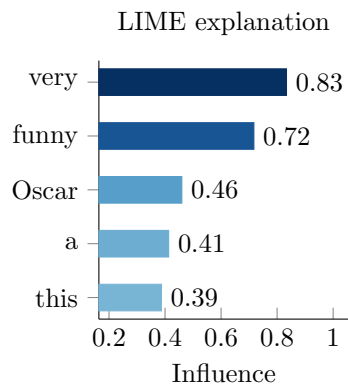
Figure 8: While most individual features have positive influence for `BII` and `LIME` both, `BII` is able to pick up some interesting interactions among words. For example the strong negative interaction between "not" and "funny" and "very" and "funny" that leads to less confidence on the negative prediction of the review.
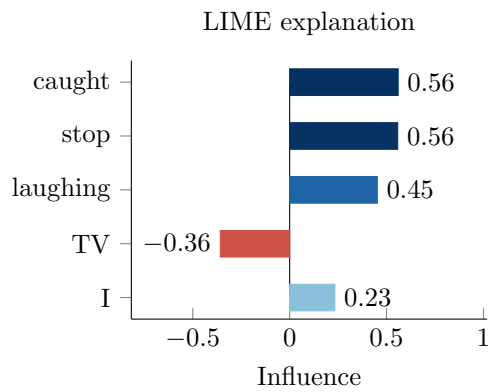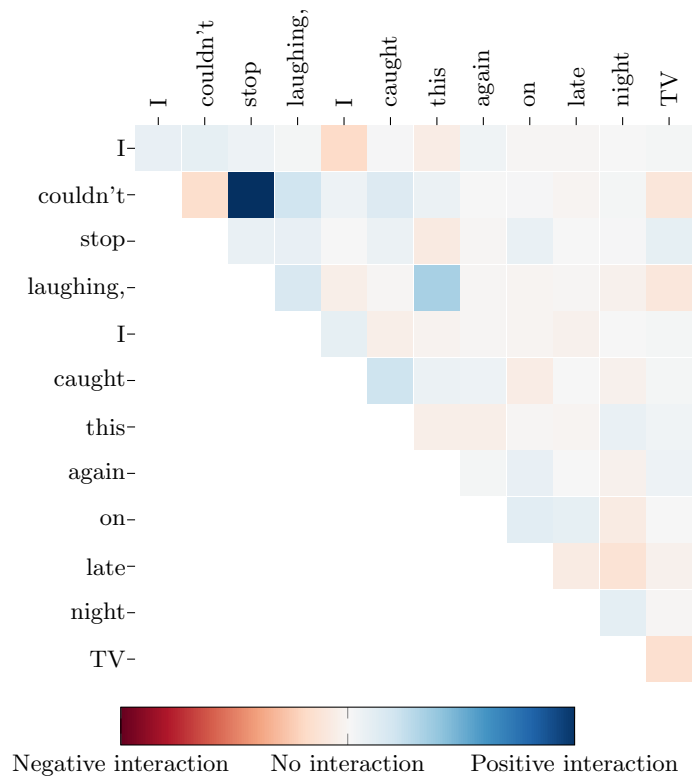
Figure 9: `BII` assigns almost negligible importance to individual word and picks up on the strong positive interaction between "couldn't" and "stop" that leads to a positive prediction of this review. However, `LIME` or any other feature based model explanation fails to capture such interactions.
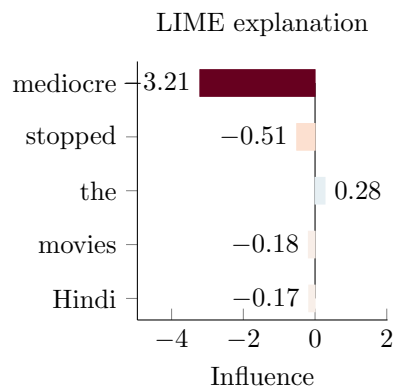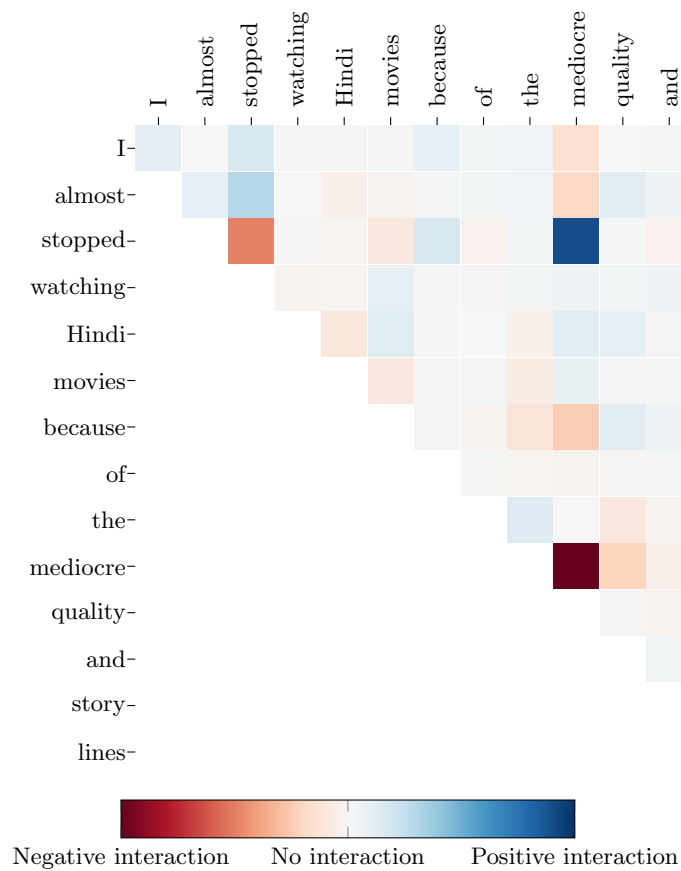
Figure 10: While `BII` and `LIME` agrees on the feature importance level and suggests that "stopped" and "mediocre" has high negative influence. However, `BII` points out that only one of the words "stopped" and "mediocre" has more negative influnce and they both together has positive interactions that leads to less confidence on the negative prediction of the review.

Contact

Neel Patel
University of Southern California
Los Angeles, USA
Email: `neelbpat@usc.edu`

Martin Strobel
National University of Singapore
Singapore, Singapore
Email: `mstrobel@comp.nus.edu.sg`

Yair Zick
University of Massachusetts, Amherst
Amherst, USA
Email: `yzick@umass.edu`