Influence in Classification via Cooperative Game Theory

Amit Datta, Anupam Datta, Ariel D. Procaccia and Yair Zick

May 1, 2015

CMU-CyLab-15-001

CyLab
Carnegie Mellon University
Pittsburgh, PA 15213

Influence in Classification via Cooperative Game Theory

Amit Datta, Anupam Datta, Ariel D. Procaccia and Yair Zick
Carnegie Mellon University
amitdatta,danupam@cmu.edu
arielpro,yairzick@cs.cmu.edu

Abstract

A dataset has been classified by some unknown classifier into two types of points. What were the most important factors in determining the classification outcome? In this work, we employ an axiomatic approach in order to uniquely characterize an influence measure: a function that, given a set of classified points, outputs a value for each feature corresponding to its influence in determining the classification outcome. We show that our influence measure takes on an intuitive form when the unknown classifier is linear. Finally, we employ our influence measure in order to analyze the effects of user profiling on Google's online display advertising.

1 Introduction

A recent white house report [Podesta *et al.*, 2014] highlights some of the major risks in the ubiquitous use of big data technologies. According to the report, one of the major issues with large scale data collection and analysis is a glaring lack of transparency. For example, a credit reporting company collects consumer data from third parties, and uses machine learning analysis to estimate individuals' credit score. On the one hand, this method is "impartial": an emotionless algorithm cannot be accused of being malicious (discriminatory behavior is not hard-coded). However, it is hardly transparent; indeed, it is difficult to tease out the determinants of one's credit score: it depends on the user's financial activities, age, address, the behavior of similar users and many other factors. This is a major issue: big-data analysis does not intend to discriminate, but inadvertent discrimination does occur: treating users differently based on unfair criteria (e.g. online retailers offering different discounts or goods based on place of residence or past purchases).

In summary, big data analysis leaves users vulnerable. They may be discriminated against, and no one (including the algorithm's developers!) may even know why; what's worse, traditional methods for preserving user anonymity (e.g. by "opting out" of data collection) offer little protection; big data techniques allow companies to infer individuals' data based on similar users [Barocas and Nissenbaum, 2014]. Since it is

often difficult to "pop the hood" and understand the inner workings of classification algorithms, maintaining transparency in classification is a major challenge. In more concrete terms, transparency can be interpreted as understanding what influences the decisions of a black-box classifier. This is where our work comes in.

Suppose that we are given a dataset B of users; here, every user $\mathbf{a} \in B$ can be thought of as a vector of features (e.g. $\mathbf{a} = (\text{age, gender, IP address} \dots)$), where the i-th coordinate of $\mathbf{a}$ corresponds to the state of the i-th feature. Each $\mathbf{a}$ has a value $v(\mathbf{a})$ (say, the credit score of $\mathbf{a}$). We are interested in the following question: *given a dataset B of various feature vectors and their values, how influential was each feature in determining these values?*

In more detail, given a set $N = \{1, \dots, n\}$ of features, a dataset B of feature profiles, where every profile $\mathbf{a}$ has a value $v(\mathbf{a})$, we would like to compute a measure $\phi_i(N, B, v)$ that corresponds to feature i's importance in determining the labels of the points in B. We see this work as an important first step towards a concrete methodology for transparency analysis of big-data algorithms.

Our Contribution: We take an axiomatic approach — which draws heavily on cooperative game theory — to define an influence measure. The merit of our approach lies in its independence of the underlying structure of the classification function; all we need is to collect data on its behavior.

We show that our influence measure is the unique measure satisfying some natural properties (Section 2). As a case study, we show that when the input values are given by a linear classifier, our influence measure has an intuitive geometric interpretation (Section 3). Finally, we show that our axioms can be extended in order to obtain other influence measures (Section 4). For example, our axioms can be used to obtain a measure of *state influence*, as well as influence measures where a prior distribution on the data is assumed, or a measure that uses pseudo-distance between user profiles to measure influence.

We complement our theoretical results with an implementation of our approach, which serves as a proof of concept (Section 5). Using our framework, we identify ads where certain user features have a significant influence on whether the ad is shown to users. Our experiments show that our influence measures behave in a desirable manner. In particular, a Spanish language ad — clearly biased towards Spanish speakers — demonstrated the highest influence of any feature among all ads.

1.1 Related Work

Axiomatic characterizations have played an important role in the design of provably fair revenue divisions [Shapley, 1953; Young, 1985; Banzhaf, 1965; Lehrer, 1988]. Indeed, one can think of the setting we describe as a generalization of cooperative games, where agents can have more than one state — in cooperative games, agents are either present or absent from a coalition. Some papers extend cooperative games to settings where agents have more than one state, and define influence measures for such settings [Chalkiadakis *et al.*, 2010; Zick *et al.*, 2014]; however, our setting is far more general.

Our definition of influence measures the ability of a feature to affect the classification outcome if changed (e.g. how often does a change in gender cause a change in the display frequency of an ad); this idea is used in the analysis of cause [Halpern and Pearl, 2005; Tian and Pearl, 2000], and responsibility [Chockler and Halpern, 2004]; our influence measure can be seen as an application of these ideas to a classification setting.

Influence measures are somewhat related to *feature selection* [Blum and Langley, 1997]. Feature selection is the problem of finding the set of features that are most relevant to the classification task, in order to improve the performance of a classifier on the data; that is, it is the problem of finding a subset of features, such that if we train a classifier using just those features, the error rate is minimized. Some of the work on feature selection employs feature ranking methods; some even use the Shapley value as a method for selecting the most important features [Cohen *et al.*, 2005]. Our work differs from feature selection both in its objectives and its methodology. Our measures can be used in order to rank features, but we are not interested in training classifiers; rather, we wish to decide which features influence the decision of an unknown classifier. That said, one can certainly employ our methodology in order to rank features in feature selection tasks.

When the classifier is linear, our influence measures take on a particularly intuitive interpretation as the aggregate volume between two hyperplanes [Marichal and Mossinghoff, 2006].

Recent years have seen tremendous progress on methods to enhance fairness in classification [Dwork *et al.*, 2012; Kamishima *et al.*, 2011], user privacy [Balebako *et al.*, 2012; Pedreschi *et al.*, 2008; Wills and Tatar, 2012] and the prevention of discrimination [Kamiran and Calders, 2009; Calders and Verwer, 2010; Luong *et al.*, 2011]. Our work can potentially inform all of these research thrusts: a classifier can be deemed fair if the influence of certain features is low; for example, high gender influence may indicate discrimination against a certain gender. In terms of privacy, if a hidden feature (i.e. one that is not part of the input to the classifier) has high influence, this indicates a possible breach of user privacy.

2 Axiomatic Characterization

We begin by briefly presenting our model. Given a set of *features* $N = \{1, \ldots, n\}$, let A_i be the set of possible *values*, or *states* that feature i can take; for example, the i-th feature could be gender, in which case $A_i = \{\text{male}, \text{female}, \text{other}\}$. We are given *partial* outputs of a function over a dataset containing feature profiles. That is, we are given a subset B of $A = \prod_{i \in N} A_i$, and a valuation $v(\mathbf{a})$ for every $\mathbf{a} \in B$. By given, we mean that we do not know the actual structure of v, but we know what values it takes over the dataset B. Formally, our input is a tuple $\mathcal{G} = \langle N, B, v \rangle$, where $v : A \to \mathbb{Q}$ is a function assigning a value of $v(\mathbf{a})$ to each data point $\mathbf{a} \in B$. We refer to $\mathcal{G}$ as the *dataset*. When $v(\mathbf{a}) \in \{0, 1\}$ for all $\mathbf{a} \in B$, v is a *binary classifier*. When $B = A$ and $|A_i| = 2$ for all $i \in N$, the dataset corresponds to a standard TU cooperative game [Chalkiadakis *et al.*, 2011] (and is a simple game if $v(\mathbf{a}) \in \{0, 1\}$).

We are interested in answering the following question: *how influential is feature*

i? Our desired output is a measure $\phi_i(\mathcal{G})$ that will be associated with each feature i. The measure $\phi_i(\mathcal{G})$ should be a good metric of the importance of i in determining the values of v over B.

Our goal in this section is to show that there exists a unique influence measure that satisfies certain natural axioms. We begin by describing the axioms, starting with symmetry.

Given a dataset $\mathcal{G} = \langle N, B, v\rangle$ and a bijective mapping σ from N to itself, we define $\sigma\mathcal{G} = \langle \sigma N, \sigma B, \sigma v\rangle$ in the natural way: σN has all of the features relabeled according to σ (i.e. the index of i is now $\sigma(i)$); σB is $\{\sigma\mathbf{a} \mid \mathbf{a} \in B\}$, and $\sigma v(\sigma\mathbf{a}) = v(\mathbf{a})$ for all $\sigma\mathbf{a} \in \sigma B$. Given a bijective mapping $\tau : A_i \to A_i$ over the states of some feature $i \in N$, we define $\tau\mathcal{G} = \langle N, \tau B, \tau v\rangle$ in a similar manner.

Definition 2.1. An influence measure ϕ satisfies the *feature symmetry* property if it is invariant under relabelings of features: given a dataset $\mathcal{G} = \langle N, B, v\rangle$ and some bijection $\sigma : N \to N$, $\phi_i(\mathcal{G}) = \phi_{\sigma(i)}(\sigma\mathcal{G})$ for all $i \in N$. A influence measure ϕ satisfies the *state symmetry* property if it is invariant under relabelings of states: given a dataset $\mathcal{G} = \langle N, B, v\rangle$, some $i \in N$, and some bijection $\tau : A_i \to A_i$, $\phi_j(\mathcal{G}) = \phi_j(\tau\mathcal{G})$ for all $j \in N$. Note that it is possible that $i \neq j$. A measure satisfying both state and feature symmetry is said to satisfy the *symmetry* axiom (Sym).

Feature symmetry is a natural extension of the symmetry axiom defined for cooperative games (see e.g. [Banzhaf, 1965; Lehrer, 1988; Shapley, 1953]). However, state symmetry does not make much sense in classic cooperative games; it would translate to saying that for any set of players $S \subseteq N$ and any $j \in N$, the value of i is the same if we treat S as $S \setminus \{j\}$, and $S \setminus \{j\}$ as S. While in the context of cooperative games this is rather uninformative, we make non-trivial use of it in what follows.

We next describe a sufficient condition for a feature to have no influence: a feature should not have any influence if it does not affect the outcome in any way. Formally, a feature $i \in N$ is a *dummy* if $v(\mathbf{a}) = v(\mathbf{a}_{-i}, b)$ for all $\mathbf{a} \in B$, and all $b \in A_i$ such that $(\mathbf{a}_{-i}, b) \in B$.

Definition 2.2. An influence measure ϕ satisfies the *dummy* property if $\phi_i(\mathcal{G}) = 0$ whenever i is a dummy in the dataset $\mathcal{G}$.

The dummy property is a standard extension of the dummy property used in value characterizations in cooperative games. However, when dealing with real datasets, it may very well be that there is no vector $\mathbf{a} \in B$ such that $(\mathbf{a}_{-i}, b) \in B$; this issue is discussed further in Section 6.

Cooperative game theory employs a notion of value additivity in the characterization of both the Shapley and Banzhaf values. Given two datasets $\mathcal{G}_1 = \langle N, B, v_1\rangle, \mathcal{G}_2 = \langle N, B, v_2\rangle$, we define $\mathcal{G} = \langle N, A, v\rangle = \mathcal{G}_1 + \mathcal{G}_2$ with $v(\mathbf{a}) = v_1(\mathbf{a}) + v_2(\mathbf{a})$ for all $\mathbf{a} \in B$.

Definition 2.3. An influence measure ϕ satisfies additivity (AD) if $\phi_i(\mathcal{G}_1 + \mathcal{G}_2) = \phi_i(\mathcal{G}_1) + \phi_i(\mathcal{G}_2)$ for any two datasets $\mathcal{G}_1 = \langle N, B, v_1\rangle, \mathcal{G}_2 = \langle N, B, v_2\rangle$.

The additivity axiom is commonly used in the axiomatic analysis of revenue division in cooperative games (see [Lehrer, 1988; Shapley, 1953]); however, it fails to

capture a satisfactory notion of influence in our more general setting. We now show that any measure that satisfies additivity, in addition to the symmetry and dummy properties, must evaluate to zero for all features. To show this, we first define the following simple class of datasets.

Definition 2.4. Let $\mathcal{U}_{\mathbf{a}} = \langle N, A, u_{\mathbf{a}} \rangle$ be the dataset defined by the classifier $u_{\mathbf{a}}$, where $u_{\mathbf{a}}(\mathbf{a}') = 1$ if $\mathbf{a}' = \mathbf{a}$, and is 0 otherwise. The dataset $\mathcal{U}_{\mathbf{a}}$ is referred to as the *singleton dataset* over $\mathbf{a}$.

It is an easy exercise to show that additivity implies that for any scalar $\alpha \in \mathbb{Q}$, $\phi_i(\alpha\mathcal{G}) = \alpha\phi_i(\mathcal{G})$, where the dataset $\alpha\mathcal{G}$ has the value of every point scaled by a factor of α.

Proposition 2.5. *Any influence measure that satisfies the (Sym), (D) and (AD) axioms evaluates to zero for all features.*

Proof. First, we show that for any $\mathbf{a}, \mathbf{a}' \in A$ and any $b \in A_i$, it must be the case that $\phi_i(\mathcal{U}_{(\mathbf{a}_{-i},b)}) = \phi_i(\mathcal{U}_{(\mathbf{a}'_{-i},b)})$. This is true because we can define a bijective mapping from $\mathcal{U}_{(\mathbf{a}_{-i},b)}$ to $\mathcal{U}_{(\mathbf{a}'_{-i},b)}$: for every $j \in N \setminus \{i\}$, we swap a_j and a'_j. By state symmetry, $\phi_i(\mathcal{U}_{(\mathbf{a}_{-i},b)}) = \phi_i(\mathcal{U}_{(\mathbf{a}'_{-i},b)})$.

Next, if ϕ is additive, then for any dataset $\mathcal{G} = \langle N, B, v \rangle$, $\phi_i(\mathcal{G}) = \sum_{\mathbf{a} \in B} v(\mathbf{a})\phi_i(\mathcal{U}_{\mathbf{a}})$. That is, the influence of a feature must be the sum of its influenceover singleton datasets, scaled by $v(\mathbf{a})$.

Now, suppose for contradiction that there exists some singleton dataset $\mathcal{U}_{\bar{\mathbf{a}}}$ ($\bar{\mathbf{a}} \in B$) for which some feature $i \in N$ does not have an influence of zero. That is, we assume that $\phi_i(\mathcal{U}_{\bar{\mathbf{a}}}) \neq 0$. We define a dataset $\mathcal{G} = \langle N, A, v \rangle$ in the following manner: for all $\mathbf{a} \in A$ such that $\mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}$, we set $v(\mathbf{a}) = 1$, and $v(\mathbf{a}) = 0$ if $\mathbf{a}_{-i} \neq \bar{\mathbf{a}}_{-i}$. In the resulting dataset, $v(\mathbf{a})$ is solely determined by the values of features in $N \setminus \{i\}$; in other words $v(\mathbf{a}) = v(\mathbf{a}_{-i}, b)$ for all $b \in A_i$, hence feature i is a dummy. According to the dummy axiom, we must have that $\phi_i(\mathcal{G}) = 0$; however,

$$\begin{aligned} 0 = \phi_i(\mathcal{G}) &= \sum_{\mathbf{a}: v(\mathbf{a})=1} \phi_i(\mathcal{U}_{\mathbf{a}}) = \sum_{b \in A_i} \phi_i(\mathcal{U}_{(\bar{\mathbf{a}}_{-i},b)}) \\ &= \sum_{b \in A_i} \phi_i(\mathcal{U}_{\bar{\mathbf{a}}}) = |A_i|\phi_i(\mathcal{U}_{\bar{\mathbf{a}}}) > 0, \end{aligned}$$

where the first equality follows from the decomposition of $\mathcal{G}$ into singleton datasets, and the third equality holds by Symmetry. This is a contradiction. □

As Proposition 2.5 shows, the additivity, symmetry and dummy properties do not lead to a meaningful description of influence. A reader familiar with the axiomatic characterization of the Shapley value [Shapley, 1953] will find this result rather disappointing: the classic characterizations of the Shapley and Banzhaf values assume additivity (that said, The axiomatization by Young [1985] does not assume additivity).

We now show that there is an influence measure uniquely defined by an alternative axiom, which echoes the union intersection property described by Lehrer [1988]. In what follows, we assume that all datasets are classified by a binary classifier. We write

$W(B)$ to be the set of all profiles in B such that $v(\mathbf{a}) = 1$, and $L(B)$ to be the set of all profiles in B that have a value of 0. We refer to $W(B)$ as the *winning profiles* in B, and to $L(B)$ as the *losing profiles* in B. We can thus write $\phi_i(W(B), L(B))$, rather than $\phi_i(\mathcal{G})$. Given two disjoint sets $W, L \subseteq A$, we can define the dataset as $\mathcal{G} = \langle W, L\rangle$, and the influence of i as $\phi_i(W, L)$, without explicitly writing N, B and v. As we have seen, no measure can satisfy the additivity axiom (as well as symmetry and dummy axioms) without being trivial. We now propose an alternative influence measure, captured by the following axiom:

Definition 2.6. An influence measure ϕ satisfies the *disjoint union (DU)* property if for any $Q \subseteq A$, and any disjoint $R, R' \subseteq A \setminus Q$, $\phi_i(Q, R) + \phi_i(Q, R') = \phi_i(Q, R \cup R')$, and $\phi_i(R, Q) + \phi_i(R', Q) = \phi_i(R \cup R', Q)$.

An influence measure ϕ satisfying the (DU) axiom is additive with respect to independent observations *of the same type*. Suppose that we are given the outputs of a binary classifier on two datasets: $\mathcal{G}_1 = \langle W, L_1\rangle$ and $\mathcal{G}_2 = \langle W, L_2\rangle$. The (DU) axiom states that the ability of a feature to affect the outcome on $\mathcal{G}_1$ is independent of its ability to affect the outcome in $\mathcal{G}_2$, if the winning states are the same in both datasets.

Replacing additivity with the disjoint union property yields a unique influence measure, with a rather simple form.

$$\chi_i(\mathcal{G}) = \sum_{\mathbf{a}\in B} \sum_{b\in A_i:(\mathbf{a}_{-i},b)\in B} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})| \tag{1}$$

χ measures the number of times that a change in the state of i causes a change in the classification outcome. If we normalize χ and divide by $|B|$, the resulting measure has the following intuitive interpretation: pick a vector $\mathbf{a} \in B$ uniformly at random, and count the number of points in A_i for which $(\mathbf{a}_{-i}, b) \in B$ and i changes the value of $\mathbf{a}$. We note that when all features have two states and $B = A$, χ coincides with the (raw) Banzhaf power index [Banzhaf, 1965].

We now show that χ is a unique measure satisfying (D), (Sym) and (DU). We begin by presenting the following lemma, which characterizes influence measures satisfying (D), (Sym) and (DU) when dataset contains only a single feature.

Lemma 2.7. *Let ϕ be an influence measure that satisfies state symmetry, and let $\mathcal{G}_1 = \langle\{i\}, A_i, v_1\rangle$ and $\mathcal{G}_2 = \langle\{i\}, A_i, v_2\rangle$ be two datasets with a single feature i; if the number of winning states under $\mathcal{G}_1$ and $\mathcal{G}_2$ is identical, then $\phi_i(\mathcal{G}_1) = \phi_i(\mathcal{G}_2)$.*

Proof Sketch. We simply construct a bijective mapping from the winning states of i under $\mathcal{G}_1$ and its winning states in $\mathcal{G}_2$. By state symmetry, $\phi_i(\mathcal{G}_1) = \phi_i(\mathcal{G}_2)$. □

Lemma 2.7 implies that for single feature games, the value of a feature only depends on the number of winning states, rather than their identity.

We are now ready to show the main theorem for this section: χ is the unique influence measure satisfying the three axioms above, up to a constant factor.

Theorem 2.8. *An influence measure ϕ satisfies (D), (Sym) and (DU) if and only if there exists a constant C such that for every dataset $\mathcal{G} = \langle N, B, v\rangle$*

$$\phi_i(\mathcal{G}) = C \cdot \chi_i(\mathcal{G}).$$

Proof. It is an easy exercise to verify that χ satisfies the three axioms, so we focus on the "only if" direction.

We present our proof assuming that we are given the set A as data; the proof goes through even if we assume that we are presented with some arbitrary $B \subseteq A$. Let us write $W = W(A)$ and $L = L(A)$. Given some $\mathbf{a}_{-i} \in A_{-i}$, we write $L_{\mathbf{a}_{-i}} = \{\bar{\mathbf{a}} \in L \mid \mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}\}$, and $W_{\mathbf{a}_{-i}} = \{\bar{\mathbf{a}} \in W \mid \mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}\}$.

Using the disjoint union property, we can decompose $\phi_i(W, L)$ as follows:

$$\phi_i(W, L) = \sum_{\mathbf{a}_{-i} \in A_{-i}} \sum_{\bar{\mathbf{a}}_{-i} \in A_{-i}} \phi_i(W_{\mathbf{a}_{-i}}, L_{\bar{\mathbf{a}}_{-i}}). \tag{2}$$

Now, if $\bar{\mathbf{a}}_{-i} \neq \mathbf{a}_{-i}$, then feature i is a dummy given the dataset provided. Indeed, state profiles are either in $W_{\mathbf{a}_{-i}}$ or in $L_{\bar{\mathbf{a}}_{-i}}$; that is, if $v(\mathbf{a}_{-i}, b) = 0$, then $(\mathbf{a}_{-i}, b)$ is unobserved, and if $v(\bar{\mathbf{a}}_{-i}, b) = 1$, then $(\bar{\mathbf{a}}_{-i}, b)$ is unobserved. We conclude that

$$\phi_i(W, L) = \sum_{\mathbf{a}_{-i} \in A_{-i}} \phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}}). \tag{3}$$

Let us now consider $\phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}})$. Since ϕ satisfies state symmetry, Lemma 2.7 implies that ϕ_i can only possibly depend on $\mathbf{a}_{-i}$, $|W_{\mathbf{a}_{-i}}|$ and $|L_{\mathbf{a}_{-i}}|$. Next, for any $\mathbf{a}_{-i}$ and $\mathbf{a}'_{-i}$ such that $|L_{\mathbf{a}_{-i}}| = |L_{\mathbf{a}'_{-i}}|$ and $|W_{\mathbf{a}_{-i}}| = |W_{\mathbf{a}'_{-i}}|$, so by Lemma 2.7 $\phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}}) = \phi_i(W_{\mathbf{a}'_{-i}}, L_{\mathbf{a}'_{-i}})$. In other words ϕ_i only depends on $|W_{\mathbf{a}_{-i}}|, |L_{\mathbf{a}_{-i}}|$, and not on the identity of $\mathbf{a}_{-i}$.

Thus, one can see ϕ_i for a single feature as a function of two parameters, w and l in $\mathbb{N}$, where w is the number of winning states and l is the number of losing states. According to the dummy property, we know that $\phi_i(w, 0) = \phi_i(0, l) = 0$; moreover, the disjoint union property tells us that $\phi_i(x, l) + \phi_i(y, l) = \phi_i(x + y, l)$, and that $\phi_i(w, x) + \phi_i(w, y) = \phi_i(w, x + y)$. We now show that $\phi_i(w, l) = \phi_i(1, 1)wl$.

Our proof is by induction on $w + l$. For $w + l = 2$ the claim is clear. Now, assume without loss of generality that $w > 1$ and $l \geq 1$; then we can write $w = x + y$ for $x, y \in \mathbb{N}$ such that $1 \leq x, y < w$. By our previous observation,

$$\begin{aligned}\phi_i(w, l) &= \phi_i(x, l) + \phi_i(y, l) \\ &\overset{i.h.}{=} \phi_i(1, 1)xl + \phi_i(1, 1)yl = \phi_i(1, 1)wl.\end{aligned}$$

Now, $\phi_i(1, 1)$ is the influence of feature i when there is exactly one losing state profile, and one winning state profile. We write $\phi_i(1, 1) = c_i$.

Let us write $W_i(\mathbf{a}_{-i}) = \{b \in A_i \mid v(\mathbf{a}_{-i}, b) = 1\}$ and $L_i(\mathbf{a}_{-i}) = A_i \setminus W_i(\mathbf{a}_{-i})$. Thus, $|W_{\mathbf{a}_{-i}}| = |W_i(\mathbf{a}_{-i})|$, and $|L_{\mathbf{a}_{-i}}| = |L_i(\mathbf{a}_{-i})|$. Putting it all together, we get that

$$\phi_i(\mathcal{G}) = c_i \sum_{\mathbf{a}_{-i} \in A_{-i}} |W_i(\mathbf{a}_{-i})| \cdot |L_i(\mathbf{a}_{-i})| \tag{4}$$

We just need to show that the measure given in (4) equals χ_i (modulo c_i). Indeed, (4) equals $\sum_{\mathbf{a} \in A:\ v(\mathbf{a})=0} |W_i(\mathbf{a}_{-i})|$, which in turn equals $\sum_{\mathbf{a} \in A:\ v(\mathbf{a})=0} \sum_{b \in A_i} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|$. Similarly, (4) equals

$$\sum_{\mathbf{a} \in A: v(\mathbf{a})=1} \sum_{b \in A_i} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|.$$

Thus,

$$\sum_{\mathbf{a}_{-i} \in A_{-i}} |W_i(\mathbf{a}_{-i})| \cdot |L_i(\mathbf{a}_{-i})| = \frac{1}{2} \sum_{\mathbf{a} \in A} \sum_{b \in A_i} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|;$$

in particular, for every dataset $\mathcal{G} = \langle N, A, v \rangle$ and every $i \in N$, there is some constant C_i such that $\phi_i(\mathcal{G}) = C_i \chi_i(\mathcal{G})$. To conclude the proof, we must show that $C_i = C_j$ for all $i, j \in N$. Let $\sigma : N \to N$ be the bijection that swaps i and j; then $\phi_i(\mathcal{G}) = \phi_{\sigma(i)}(\sigma\mathcal{G})$. By feature symmetry, $C_i \chi_i(\mathcal{G}) = \phi_i(\mathcal{G}) = \phi_{\sigma(i)}(\sigma\mathcal{G}) = \phi_j(\sigma\mathcal{G}) = C_j \chi_j(\sigma\mathcal{G}) = C_j \chi_i(\mathcal{G})$, thus $C_i = C_j$. □

3 Case Study: Influence for Linear Classifiers

To further ground our results, we now present their application to the class of linear classifiers. For this class of functions, our influence measure takes on an intuitive interpretation.

A *linear classifier* is defined by a hyperplane in $\mathbb{R}^n$; all points that are on one side of the hyperplane are colored blue (in our setting, have value 1), and all points on the other side are colored red (have a value of 0). Formally, we associate a weight $w_i \in \mathbb{R}$ with every one of the features in N (we assume that $w_i \neq 0$ for all $i \in N$); a point $\mathbf{x} \in \mathbb{R}^n$ is blue if $\mathbf{x} \cdot \mathbf{w} \geq q$, where $q \in \mathbb{R}$ is a given parameter. The classification function $v : \mathbb{R}^n \to \{0, 1\}$ is given by

$$v(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \cdot \mathbf{w} \geq q \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Fixing the value of x_i to some $b \in \mathbb{R}$, let us consider the set $W_i(b) = \{\mathbf{x}_{-i} \in \mathbb{R}^{n-1} \mid v(\mathbf{x}_{-i}, b) = 1\}$; we observe that if $b < b'$ and $w_i > 0$, then $W_i(b) \subset W_i(b')$ (if $w_i < 0$ then $W_i(b') \subset W_i(b)$). Given two values $b, b' \in \mathbb{R}$, we denote by

$$D_i(b, b') = \{\mathbf{x}_{-i} \in \mathbb{R}^{n-1} \mid v(\mathbf{x}_{-i}, b) \neq v(\mathbf{x}_{-i}, b')\}.$$

By our previous observation, if $b < b'$ then $D_i(b, b') = W_i(b') \setminus W_i(b)$, and if $b > b'$ then $D_i(b, b') = W_i(b) \setminus W_i(b')$.

Suppose that rather than taking points in $\mathbb{R}^n$, we only take points in $[0, 1]^n$; then we can define $|D_i(b, b')| = Vol(D_i(b, b'))$, where

$$Vol(D_i(b, b')) = \int_{\mathbf{x}_{-i} \in [0,1]^{n-1}} |v(\mathbf{x}_{-i}, b') - v(\mathbf{x}_{-i}, b)| \partial \mathbf{x}_{-i}.$$

In other words, in order to measure the total influence of setting the state of feature i to b, we must take the total volume of $D_i(b, b')$ for all $b' \in [0, 1]$, which equals $\int_{b'=0}^{1} Vol(D_i(b, b')) \partial b$. Thus, the total influence of setting the state of i to b is $\int_{\mathbf{x} \in [0,1]^n} |v(\mathbf{x}_{-i}, b) - v(\mathbf{x})| \partial \mathbf{x}$. The total influence of i would then be naturally the total influence of its states, i.e.

$$\int_{b=0}^{1} \int_{\mathbf{x} \in [0,1]^n} |v(\mathbf{x}_{-i}, b) - v(\mathbf{x})| \partial \mathbf{x} \partial b. \tag{6}$$

The formula in Equation (6) is denoted by $\chi_i(\mathbf{w}; q)$. Equation (1) is a discretized version of Equation (6); the results of Section 2 can be extended to the continuous setting, with only minimal changes to the proofs.

We now show that the measure given in (6) agrees with the weights in some natural manner. This intuition is captured in Theorem 3.1 (proof omitted).

Theorem 3.1. *Let v be a linear classifier defined by $\mathbf{w}$ and q; then $\chi_i(\mathcal{G}) \geq \chi_j(\mathcal{G})$ if and only if $|w_i| \geq |w_j|$.*

Given Theorem 3.1, one would expect the following to hold: suppose that we are given two weight vectors, $\mathbf{w}, \mathbf{w}' \in \mathbb{R}^n$ such that $w_j = w'_j$ for all $j \neq i$, but $w_i < w'_i$. Let v be the linear classifier defined by $\mathbf{w}$ and q and v' be the linear classifier defined by $\mathbf{w}'$ and q. Is it the case that feature i is more influential under v' than under v? In other words, does influence monotonicity hold when we increase the weight of an individual feature? The answer to this is negative.

Example 3.2. Let us consider a single feature game where $N = \{1\}$, $A_1 = [0, 1]$, and $v(x) = 1$ if $wx \geq q$, and $v(x) = 0$ if $wx < q$ for a given $w > q$. The fraction of times that 1 is pivotal is

$$|Piv_1| = \int_{b=0}^{1} \int_{x=0}^{1} \mathbb{I}(v(b){=}1 \wedge v(x){=}0) \partial x \partial b;$$

simplifying, this expression is equal to $\left(1 - \frac{q}{w}\right) \frac{q}{w}$. We can show that $\chi_1 = 2|Piv_1|$, we have that χ_1 is maximized when $q = 2w$; in particular, χ_1 is monotone increasing when $q < w \leq 2q$, and it is monotone decreasing when $w \geq 2q$.

Example 3.2 highlights the following phenomenon: fixing the other features to be $\mathbf{a}_{-i}$, the influence of i is maximized when $|L_{\mathbf{a}_{-i}}| = |W_{\mathbf{a}_{-i}}|$. This can be interpreted probabilistically: we sample a random feature from B, and assume that for any fixed $\mathbf{a}_{-i} \in A_{-i}$, $\Pr[v(\mathbf{a}_{-i}, b) = 1] = \frac{1}{2}$. The better a feature i agrees with our assumption, the more i is rewarded. More generally, an influence measure satisfies the *agreement with prior assumption* (APA) axiom if for any vector $(p_1, \ldots, p_n) \in [0, 1]^n$, and any fixed $\mathbf{a}_{-i} \in A_{-i}$, i's influence increases as $|\Pr[v(\mathbf{a}_{-i}, b) = 1] - p_i|$ decreases. A variant of the symmetry axiom (that reflects changes in probabilities when labels change), along with the dummy and disjoint union axioms can give us a weighted influence measure as described in Section 4.2, that also satisfies the (APA) axiom.

4 Extensions of the Feature Influence Measure

Section 2 presents an axiomatic characterization of feature influence, where the value of each feature vector is either zero or 1. We now present a few possible extensions of the measure, and the variations on the axioms that they require.

4.1 State Influence

Section 2 provided an answer to questions of the following form: what is the impact of gender on classification outcomes? The answer provided in previous sections was

that influence was a function of the feature's ability to change outcomes by changing its state.

It is also useful to ask a related question: what is the impact of the gender feature being set to "female" on classification outcomes? In other words, rather than measuring feature influence, we are measuring the influence of feature i being in a certain *state*. The results described in Section 2 can be easily extended to this setting. Moreover, the impossibility result described in Proposition 2.5 no longer holds when we measure state — rather than feature — influence: we can replace the disjoint union property with additivity to obtain an alternative classification of state influence.

4.2 Weighted Influence

Suppose that in addition to the dataset B, we are given a weight function $w : B \to \mathbb{R}$. $w(\mathbf{a})$ can be thought of as the number of occurrences of the vector $\mathbf{a}$ in the dataset, the probability that $\mathbf{a}$ appears, or some intrinsic importance measure of $\mathbf{a}$. Note that in Section 2 we implicitly assume that all points occur at the same frequency (are equally likely) and are equally important. A simple extension of the disjoint union and symmetry axioms to a weighted variant shows that the only weighted influence measure that satisfies these axioms is

$$\chi_i^w(B) = \sum_{\mathbf{a}\in B} \sum_{b\in A_i:(\mathbf{a}_{-i},b)\in B} w(\mathbf{a})|v(\mathbf{a}_{-i},b) - v(\mathbf{a})|.$$

4.3 General Distance Measures

Suppose that instead of a classifier $v : A \to \{0,1\}$ we are given a pseudo-distance measure: that is, a function $d : A \times A \to \mathbb{R}$ that satisfies $d(\mathbf{a},\mathbf{a}') = d(\mathbf{a}',\mathbf{a})$, $d(\mathbf{a},\mathbf{a}) = 0$ and the triangle inequality. Note that it is possible that $d(\mathbf{a},\mathbf{a}') = 0$ but $\mathbf{a} \neq \mathbf{a}'$. An axiomatic analysis in such general settings is possible, but requires more assumptions on the behavior of the influence measure. Such an axiomatic approach leads us to show that the influence measure

$$\chi_i^d(B) = \sum_{\mathbf{a}\in B} \sum_{b\in A_i:(\mathbf{a}_{-i},b)\in B} d((\mathbf{a}_{-i},b),\mathbf{a})$$

is uniquely defined via some natural axioms. The additional axioms are a simple extension of the disjoint union property, and a minimal requirement stating that when $B = \{\mathbf{a},(\mathbf{a}_{-i},b)\}$, then the influence of a feature is $\alpha d((\mathbf{a}_{-i},b),\mathbf{a})$ for some constant α independent of i. The extension to pseudo-distances proves to be particularly useful when we conduct empirical analysis of Google's display ads system, and the effects user metrics have on display ads.

5 Implementation

We implement our influence measure to study Google's display advertising system. Users can set demographics (like gender, age) on the Google Ad Settings page[1]. These

[1] `google.com/settings/ads`

settings are used by the Google ad serving algorithm to determine which ads to serve to a user. We apply our influence measure to study how demographic settings influence the targeted ads served by Google. We use the open-source AdFisher tool [Datta *et al.*, 2014] for automating browser instances to perform online actions and collect ads.

For this study, we pick the set of features: $N = \{\text{gender, age, language}\}$. We use two genders: {male, female}, three age groups: {18–24, 35–44, 55–64}, and two languages: {English, Spanish}. There are $2 \times 3 \times 2 = 12$ combinations of these attributes. We use AdFisher to launch twelve fresh browser instances, and randomly assign each of them to one of these combinations. AdFisher drives each browser to apply the corresponding settings on the Ad Settings page, then collect ads served by Google on the BBC news page `bbc.com/news`. For each browser, the news page is reloaded 10 times with 5 second intervals.

To eliminate ads differing due to random chance, we collect ads over 100 iterations, each comprising of 12 browser instances, thereby obtaining data for 1200 simulated users. All browser instances were run from the same Ubuntu machine, so that no confounding factors introduce additional differences; for example, different locations, indicated by IP address, may affect the ads served by Google. The 1200 browsers received a total of $32,451$ ads (763 unique, as determined by the combination of title and display URL). Each feature vector $\mathbf{a}$ thus has a frequency vector of all ads $v'(\mathbf{a})$. In order to reduce the amount of noise, we remove all ads that were served fewer than 100 times in the course of the experiment, which left 55 ads of the original 763. Thus, $v'(\mathbf{a})$ is a vector with 55 coordinates, whose k^{th} coordinate is the number of times ad k appeared for a user whose profile is given by $\mathbf{a}$. We normalize $v'(\mathbf{a})$ for each ad by the total number of times that ad appeared. Thus we obtain the final value-vectors by computing $v_k(\mathbf{a}) = \frac{v'_k(\mathbf{a})}{\sum_{\mathbf{a}} v'_k(\mathbf{a})}, \forall \mathbf{a}, \forall k \in [1, 55]$.

Since there is no function that assigns a value to each feature vector, we use the general distance influence measure described in Section 4.3. The pseudo-distance we use is $cosd(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{||\mathbf{x}|| \cdot ||\mathbf{y}||}$. Cosine similarity has been used by Tschantz *et al.* [2014] and Guha *et al.* [2010] to measure similarity between display ads. The influence measure for gender, age, and language were 0.124, 0.120, and 0.141 respectively. We note that language has the highest influence on ads, however, this measure does not provide additional insight into how the influence was reflected in the ads.

To better understand the influence of features on display ads, we observe influence in individual ads. Fixing an ad k, we define the value of a feature vector to be the number of times that ad k was displayed for users with that feature vector, and use χ to measure influence.

We compare the influence measures for each attribute across all the ads and identify the top ads that demonstrate high influence. The ad for which language had the highest influence was a Spanish language ad, which was served only to browsers that set 'Spanish' as their language on the Ad Settings page. In fact, the language influence for this ad (0.167) was more than twice the next highest influence (0.075) for any ad across all attributes.

To conclude, using a general distance measure between two value-vectors, we identify that language has the highest influence on ads. By using a more fine-grained distance function, we can single out one ad which demonstrates high influence for

language. While in this case the bias is acceptable, the experiment suggests that our framework is effective in pinpointing biased or discriminatory ads.

6 Conclusions and Future Work

In this work, we analyze influence measures for classification tasks. Our influence measure is uniquely defined by a set of natural axioms, and is easily extended to other settings. The main advantage of our approach is the minimal knowledge we have of the classification algorithm. We show the applicability of our measure by analyzing the effects of user features on Google's display ads, despite having no knowledge of Google's classification algorithm (which, we suspect, is quite complex).

Dataset classification is a useful application of our methods; however, our work applies to extensions of TU cooperative games where agents have more than two states (e.g. OCF games [Chalkiadakis *et al.*, 2010]).

The measure χ is trivially hard to compute exactly, since it generalizes the raw Banzhaf power index, for which this task is known to be hard [Chalkiadakis *et al.*, 2011]. That said, both the Shapley and Banzhaf values can be approximated via random sampling [Bachrach *et al.*, 2010]. It is straightforward to show that random sampling provides good approximations for χ as well, assuming a binary classifier.

Our results can be extended in several ways. The measure χ is the number of times a change in a feature's state causes a change in the outcome. However, a partial dataset of observations may not contain any pair of vectors $\mathbf{a}, \mathbf{a}' \in B$, such that $\mathbf{a}' = (\mathbf{a}_{-i}, b)$. In Section 5, we control the dataset, so we ensure that all feature profiles appear. However, other datasets would not be as well-behaved. Extending our influence measure to accommodate non-immediate influence is an important step towards implementing our results to other classification domains. Indeed, the next step of our work is analyzing large-scale datasets, in order to better understand the ideas behind our influence measure.

Finally, our experimental results, while encouraging, are illustrative rather than informative: they tell us that Google's display ads algorithm is clever enough to assign Spanish ads to Spanish speakers. Our experimental results enumerate the number of *displayed ads*; this is not necessarily indicative of users' clickthrough rates. Since our users are virtual entities, we are not able to measure their clickthrough rates; a broader experiment, where user profiles correspond to actual human subjects, would provide better insights into the effects user profiling has on display advertising.

References

Y. Bachrach, E. Markakis, E. Resnick, A.D. Procaccia, J.S. Rosenschein, and A. Saberi. Approximating power indices: theoretical and empirical analysis. *Autonomous Agents and Multi-Agent Systems*, 20(2):105–122, 2010.

R. Balebako, P. Leon, R. Shay, B. Ur, Y. Wang, and L. Cranor. Measuring the effectiveness of privacy tools for limiting behavioral advertising. In *Web 2.0 Workshop on Security and Privacy*, 2012.

J.F. Banzhaf. Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review*, 19:317–343, 1965.

S. Barocas and H. Nissenbaum. Big data's end run around procedural privacy protections. *Communications of the ACM*, 57(11):31–33, October 2014.

A.L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1):245–271, 1997.

T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

G. Chalkiadakis, E. Elkind, E. Markakis, M. Polukarov, and N.R. Jennings. Cooperative games with overlapping coalitions. *Journal of Artificial Intelligence Research*, 39:179–216, 2010.

G. Chalkiadakis, E. Elkind, and M. Wooldridge. *Computational Aspects of Cooperative Game Theory*. Morgan and Claypool, 2011.

H. Chockler and J.Y. Halpern. Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, 22:93–115, 2004.

S. Cohen, E. Ruppin, and G. Dror. Feature selection based on the shapley value. In *Proceedings of the 19th international joint conference on Artificial intelligence (IJCAI'05)*, pages 665–670, 2005.

Amit Datta, Michael Carl Tschantz, and Anupam Datta. Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. Technical Report arXiv:1408.6491v1, ArXiv, August 2014.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. S. Zemel. Fairness through awareness. In *Proceedings of the 3rd Conference on Innovations in Theoretical Computer Science (ITCS'12)*, pages 214–226, 2012.

S. Guha, B. Cheng, and P. Francis. Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet measurement (IMC'10)*, pages 81–87, 2010.

Joseph Y Halpern and Judea Pearl. Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, 56(4):843–887, 2005.

F. Kamiran and T. Calders. Classifying without discriminating. In *Proceedings of the 2nd International Conference on Computer, Control and Communication (IC4'09)*, pages 1–6, 2009.

T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *Proceedings of the 11th IEEE International Conference on Data Mining Workshops (ICDMW'11)*, pages 643–650, 2011.

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

B.T. Luong, S. Ruggieri, and F. Turini. k-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'11)*, pages 502–510, 2011.

J.L. Marichal and M. J. Mossinghoff. Slices, slabs, and sections of the unit hypercube. *arXiv preprint math/0607715*, 2006.

D. Pedreschi, S. Ruggieri, and F. Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD'08)*, pages 560–568, 2008.

J. Podesta, P. Pritzker, E.J. Moniz, J. Holdern, and J. Zients. Big data: Seizing opportunities, preserving values. Technical report, Executive Office of the President - the White House, May 2014.

L.S. Shapley. A value for n-person games. In *Contributions to the Theory of Games, vol. 2*, Annals of Mathematics Studies, no. 28, pages 307–317. Princeton University Press, 1953.

J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Annals of Mathematics and Artificial Intelligence*, 28(1-4):287–313, 2000.

M.C. Tschantz, A. Datta, A. Datta, and J. M. Wing. A methodology for information flow experiments. *CoRR*, abs/1405.2376, 2014.

C.E. Wills and C. Tatar. Understanding what they do with what they know. In *Proceedings of the 2012 ACM workshop on Privacy in the electronic society (WPES'12)*, pages 13–18, 2012.

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

Y. Zick, E. Markakis, and E. Elkind. Arbitration and stability in cooperative games with overlapping coalitions. *Journal of Artificial Intelligence Research*, 50:847–884, 2014.

Appendix:
Influence in Classification

A Proof of Theorem 3.1

We define $Piv_i(b) = \{\mathbf{x} \in [0,1]^n \mid v(\mathbf{x}) = 1, v(\mathbf{x}_{-i}, b) = 0\}$, to be the set of all *pivotal* vectors (w.r.t. b), and *A-Piv*$_i(b) = \{\mathbf{x} \in [0,1]^n \mid v(\mathbf{x}) = 0, v(\mathbf{x}_{-i}, b) = 1\}$ to be the set of all *anti-pivotal vectors*. We write $Piv_i = \big\{(\mathbf{x}, b) \in [0,1]^{n+1} \mid \mathbf{x} \in Piv_i(b)\big\}$ and *A-Piv*$_i = \big\{(\mathbf{x}, b) \in [0,1]^{n+1} \mid \mathbf{x} \in$ *A-Piv*$_i(b)\big\}$. We note that $Vol(Piv_i) = Vol($*A-Piv*$_i)$. Given a point $(\mathbf{x}, b) \in Piv_i$, we know that $v(\mathbf{x}) = 0$ but $v(\mathbf{x}_{-i}, b) = 1$. Therefore, the point $((\mathbf{x}_{-i}, b), x_i)$ is in *A-Piv*$_i$. We conclude that

$$\begin{aligned}\chi_i &= \int_{b=0}^{1} |Piv_i(b)| + |\textit{A-Piv}_i(b)| \partial b \\ &= \int_{b=0}^{1} Vol(Piv_i(b)) \partial b + \int_{b=0}^{1} Vol(\textit{A-Piv}_i(b)) \partial b \\ &= Vol(Piv_i) + Vol(\textit{A-Piv}_i) = 2\, Vol(Piv_i)\end{aligned}$$

We begin by stating a few technical lemmas. Our objective is to establish some volume-preserving transformations between vectors for which j is pivotal, and vectors for which i is pivotal.

Thus, to show that $\chi_i \geq \chi_j$ whenever $w_i \geq w_j > 0$, it suffices to show that $Vol(Piv_i) \geq Vol(Piv_j)$.

Lemma A.1. *Suppose that $w_i > w_j > 0$; if $\mathbf{x} \in Piv_j(b) \setminus Piv_i(b)$ then $x_i > x_j$.*

Proof. First, note that if $v(\mathbf{x}_{-j}, b) = 1$ but $v(\mathbf{x}) = 0$, then $x_j < b$. Now, suppose that $x_i \leq x_j$; we show that $(\mathbf{x}_{-j}, b) \cdot \mathbf{w} \leq (\mathbf{x}_{-i}, b) \cdot \mathbf{w}$. Indeed,

$$\begin{aligned}(\mathbf{x}_{-j}, b) \cdot \mathbf{w} \leq & (\mathbf{x}_{-i}, b) \cdot \mathbf{w} \iff \\ x_i w_i + b w_j \leq & x_j w_j + b w_i \iff \\ x_i w_i - x_j w_j \leq & b(w_i - w_j)\end{aligned}$$

Thus, we just need to show that $x_i w_i - x_j w_j \leq b(w_i - w_j)$. Since $x_i \leq x_j$, $x_i w_i - x_j w_j \leq x_j(w_i - w_j)$, and since $w_i > w_j$, this is at most $b(w_i - w_j)$, as required. This means that if $x_i \leq x_j$ then $\mathbf{x} \in Piv_i(b)$, which concludes the first part of the proof. □

Let $f_{ij} : \mathbb{R}^n \to \mathbb{R}^n$ be the transformation

$$f_{ij}(\mathbf{x})_k = \begin{cases} x_i & \text{if } k = j \\ x_j & \text{if } k = i \\ x_k & \text{otherwise.} \end{cases}$$

Lemma A.2. *If $\mathbf{x} \in Piv_j(b) \setminus Piv_i(b)$ then $f_{ij}(\mathbf{x}) \notin Piv_j(b) \cup$ A-Piv$_j(b)$.*

Proof. First, note that $(b - x_j)(w_i - w_j) > 0$; this is because $b > x_j$ and $w_i > w_j$. This implies that $x_j w_i + b w_j < b w_i + x_j w_j$. Now, since $v(\mathbf{x}_{-i}, b) = 0$, we know that $b w_i + x_j w_j < q - \sum_{k \neq i,j} x_k w_k$; therefore, $\mathbf{w} \cdot (f_{ij}(\mathbf{x})_{-j}, b) = \sum_{k \neq i,j} x_k w_k + x_j w_i + b w_j < q$, and $v(f_{ij}(\mathbf{x})_{-j}, b) = 0$. This implies that $f_{ij}(\mathbf{x}) \notin \mathit{Piv}_j(b)$.

Now, $(x_i - x_j)(w_i - w_j) > 0$ since $x_i > x_j$ by Lemma A.1. Therefore, $x_j w_i + x_i w_j < x_i w_i + x_j w_j < q - \sum_{k \neq i,j} x_k w_k$, which implies that $\mathbf{w} \cdot f_{ij}(\mathbf{x}) < q$, hence $v(f_{ij}(\mathbf{x})) = 0$. In particular, $f_{ij}(\mathbf{x}) \notin \textit{A-Piv}_j(b)$. □

Lemma A.3. *Suppose* $w_i > w_j > 0$ *and that* $\mathbf{x} \in \mathit{Piv}_j(b) \setminus \mathit{Piv}_i(b)$*; if* $f_{ij}(\mathbf{x}) \notin \mathit{Piv}_i(b)$ *then* $x_i \geq b > x_j$.

Proof. Suppose that $b > x_i > x_j$. We note that $(b - x_i)(w_i - w_j) > 0$, which implies that $b w_i + x_i w_j > x_i w_i + b w_j \geq q - \sum_{k \neq i,j} x_k w_k$. Hence, $\mathbf{w} \cdot (f_{ij}(\mathbf{x})_{-i}, b) > q$, which implies that $f_{ij}(\mathbf{x}) \in \mathit{Piv}_i(b)$. Thus, if $f_{ij}(\mathbf{x}) \notin \mathit{Piv}_i(b)$, it must be the case that $x_i \geq b > x_j$. □

Given some $\mathbf{x} \in [0,1]^n$ and some $b \in [0,1]$, we define $g_{ij} : [0,1]^n \times [0,1] \to [0,1]^n$ as follows:

$$g_{ij}(\mathbf{x}, b)_k = \begin{cases} x_j & \text{if } k = i \\ b & \text{if } k = j \\ x_k & \text{otherwise.} \end{cases}$$

Lemma A.4. *If* $\mathbf{x} \in \mathit{Piv}_j(b) \setminus \mathit{Piv}_i(b)$ *and* $f_{ij}(\mathbf{x}) \notin \mathit{Piv}_i(b)$*, then* $g_{ij}(\mathbf{x}, b) \in \mathit{Piv}_i(x_i) \setminus (\mathit{Piv}_j(x_i) \cup \textit{A-Piv}_j(x_i))$.

Proof. First, we observe that $(g_{ij}(\mathbf{x}, b)_{-i}, x_i) = (\mathbf{x}_{-j}, b)$, and that $(g_{ij}(\mathbf{x}, b)_{-j}, x_i) = f_{ij}(\mathbf{x})$. As observed in Lemma A.2, if $x_i > x_j$ then $v(f_{ij}(\mathbf{x})) = 0$. Therefore, $g_{ij}(\mathbf{x}, b) \notin \mathit{Piv}_j(x_j)$. Moreover, since $\mathbf{x} \in \mathit{Piv}_j(b)$, $v(g_{ij}(\mathbf{x}, b)) = 1$, so $g_{ij}(\mathbf{x}, b) \in \mathit{Piv}_i(x_i)$. On the other hand, $(b - x_j)(w_i - w_j) > 0$, so $x_j w_i + b w_j < b w_i + x_j w_j < q - \sum_{k \neq i,j} x_k w_k$, so $g_{ij}(\mathbf{x}, b) \cdot \mathbf{w} < q$. This means that $g_{ij}(\mathbf{x}, b) \notin \textit{A-Piv}_j(x_i)$. □

Given a set $S \subseteq \mathbb{R}^m$ and a function $f : \mathbb{R}^m \to \mathbb{R}^m$, we define $f(S) = \{f(\mathbf{s}) \mid \mathbf{s} \in S\}$. We can extend f_{ij} and g_{ij} defined above to functions from $\mathbb{R}^{n+1}$ to $\mathbb{R}^{n+1}$ as follows. Given a point $(\mathbf{x}, b) \in \mathbb{R}^{n+1}$, we define $F_{ij}(\mathbf{x}, b) = (f_{ij}(\mathbf{x}), b)$, and $G_{ij}(\mathbf{x}, b) = (g_{ij}(\mathbf{x}, b), x_i)$. We note that both F_{ij} and G_{ij} merely swap coordinates in their inputs, thus they preserve distances:

$$d(G_{ij}(\mathbf{x}, b), G_{ij}(\mathbf{y}, c)) = d((\mathbf{x}, b), (\mathbf{y}, c))$$

for any metric d. Isoperimetric transformations are known to preserve volume: if $I : \mathbb{R}^m \to \mathbb{R}^m$ is an isoperimetry, then $\mathit{Vol}(S) = \mathit{Vol}(I(S))$ for any $S \subseteq \mathbb{R}^m$.

Theorem A.5. *If* $w_i \geq w_j > 0$ *then* $\mathit{Vol}(\mathit{Piv}_j) \leq \mathit{Vol}(\mathit{Piv}_i)$.

Proof. We partition Piv_j as follows. We denote

$$\begin{aligned} A_{ij} &= \mathit{Piv}_j \cap \mathit{Piv}_i, \\ B_{ij} &= \{(\mathbf{x}, b) \in \mathit{Piv}_j \setminus \mathit{Piv}_i \mid (f_{ij}(\mathbf{x}), b) \in \mathit{Piv}_i\}, \text{ and} \\ C_{ij} &= \{(\mathbf{x}, b) \in \mathit{Piv}_j \setminus \mathit{Piv}_i \mid (f_{ij}(\mathbf{x}), b) \notin \mathit{Piv}_i\}. \end{aligned}$$

Clearly, A_{ij}, B_{ij} and C_{ij} partition Piv_j.

According to Lemma A.2, $F_{ij}(B_{ij}) \subseteq Piv_i \setminus Piv_j$. Now, let us observe C_{ij}. According to Lemma A.4, $G_{ij}(C_{ij}) \subseteq Piv_i \setminus Piv_j$. It remains to show that $F_{ij}(B_{ij}) \cap G_{ij}(C_{ij}) = \emptyset$. Suppose that there are some $(\mathbf{x}, b) \in B_{ij}, (\mathbf{z}, c) \in C_{ij}$ such that $(f_{ij}(\mathbf{x}), b) = (g_{ij}(\mathbf{z}, c), z_i)$. This means that $(\mathbf{z}, c) = ((\mathbf{x}_{-i}, b), x_i)$. To prove a contradiction, it suffices to show that if $(\mathbf{x}, b) \in B_{ij}$ then we have that $((\mathbf{x}_{-i}, b), x_i) \notin C_{ij}$. In order to be in C_{ij}, it must be the case that $f_{ij}(\mathbf{x}_{-i}, b) \notin Piv_i(x_i)$; we show that $f_{ij}(\mathbf{x}_{-i}, b) \in Piv_i(x_i)$. First, let us write $f_{ij}(\mathbf{x}_{-i}, b) = \mathbf{y}$. We note that $y_k = x_k$ for all $k \neq i, j$, that $y_j = b$, and that $y_i = x_j$. Since $b > x_j$, it must be the case that $(b - x_j)(w_i - w_j) > 0$, hence $bw_i + x_j w_j > x_j w_i + bw_j$. Therefore, $\mathbf{w} \cdot \mathbf{y} < \mathbf{w} \cdot (\mathbf{x}_{-i}, b)$. Now, since $(\mathbf{x}, b) \in Piv_j \setminus Piv_i$, it must be the case that $v(\mathbf{x}_{-i}, b) = 0$, i.e. that $\mathbf{w} \cdot (\mathbf{x}_{-i}, b) < q$. This means that $v(\mathbf{y}) = 0$. We now show that $v(\mathbf{y}_{-i}, x_i) = 1$. Since $y_i = x_j$ and $y_j = b$, $(\mathbf{y}_{-i}, x_i) = (\mathbf{x}_{-j}, b)$. Since $(\mathbf{x}, b) \in Piv_j$, $v(\mathbf{y}_{-i}, x_i) = v(\mathbf{x}_{-j}, b) = 1$. Therefore, $\mathbf{y} \in Piv_i$, and thus $((\mathbf{x}_{-i}, b), x_i) \notin C_{ij}$. We conclude that indeed $F_{ij}(B_{ij}) \cap G_{ij}(C_{ij}) = \emptyset$.

To conclude,

$$\begin{aligned} Vol(Piv_j) &= Vol(A_{ij}) + Vol(B_{ij}) + Vol(C_{ij}) \\ &= Vol(A_{ij}) + Vol(F_{ij}(B_{ij})) + Vol(G_{ij}(C_{ij})) \\ &\leq Vol(Piv_i) \end{aligned}$$

which concludes the proof. □

Corollary A.6. *Let $\mathcal{G} = \langle N, [0,1]^n, v \rangle$ be a game where v is a linear separator given by $\mathbf{w}$ and q. If $w_i \geq w_j > 0$ then $\chi_i(\mathcal{G}) \geq \chi_j(\mathcal{G})$.*

Corollary A.6 shows that χ is monotone in feature weights. a complementary result shows that increasing a feature's weight would result in an increase in influence. Next, we show that Corollary A.6 holds even when weights are negative.

Lemma A.7. *Let $\mathcal{G} = \langle \{1,2\}, [0,1]^2, v \rangle$ be a 2-feature linear separator with $w_1 \geq 0$ and $w_2 < 0$. Then $\chi_1(\mathcal{G}) > \chi_2(\mathcal{G})$ if and only if $|w_1| > |w_2|$.*

Proof. We begin by assuming that $q \geq 0$. First, suppose that $w_1 < q$. In that case, for all $(x_1, x_2) \in [0,1]^2$, we have $x_1 w_1 + x_2 w_2 \leq x_1 w_1 \leq w_1 < q$, so $v(x_1, x_2) = 0$ for all $(x_1, x_2) \in [0,1]^2$. In particular, $\chi_1(\mathcal{G}) = \chi_2(\mathcal{G}) = 0$ and we are done.

We now assume that $w_1 \geq q$. We show that the claim holds by direct computation of χ_1, χ_2. We start by computing $\chi_1(\mathcal{G})$. By definition, $\chi_1(\mathcal{G})$ equals

$$\int_0^1 \left(\int_0^1 \mathbb{I}(v(x_1, x_2) = 1) \partial x_1 \int_0^1 \mathbb{I}(v(y_1, x_2) = 0) \partial y_1 \right) \partial x_2$$

which equals

$$\int_0^1 \left(\int_0^1 \mathbb{I}(x_1 \geq \frac{q - x_2 w_2}{w_1}) \partial x_1 \int_0^1 \mathbb{I}(y_1 < \frac{q - x_2 w_2}{w_1}) \partial y_1 \right) \partial x_2 \qquad (7)$$

The internal integrals in (7) are zero whenever $\frac{q - x_2 w_2}{w_1} \notin [0,1]$. We know that $\frac{q - x_2 w_2}{w_1} \geq 0$ for all $x_2 \in [0,1]$; however, $\frac{q - x_2 w_2}{w_1} \leq 1$ only when $x_2 \leq \frac{w_1 - q}{-w_2}$. This inequality is

non trivial only if $\frac{w_1-q}{-w_2} \leq 1$. This happens only when $q \geq w_1 + w_2$. Therefore, we distinguish between two cases; the first case is when $q \geq w_1 + w_2$, and the second is when $q < w_1 + w_2$. In the second case, since $q > 0$, $w_1 + w_2 > 0$ as well, hence $|w_1| > |w_2|$. In the first case we have:

$$\begin{aligned}\chi_1(\mathcal{G}) =& \int_0^{\frac{w_1-q}{-w_2}} \left(1 - \frac{q - x_2 w_2}{w_1}\right)\left(\frac{q - x_2 w_2}{w_1}\right)\partial x_2 \\ =& \frac{(w_1 - q)^2(2q + w_1)}{6(-w_2)w_1^2}\end{aligned} \tag{8}$$

In the second case we have

$$\begin{aligned}\chi_1(\mathcal{G}) =& \int_0^{1} \left(1 - \frac{q - x_2 w_2}{w_1}\right)\left(\frac{q - x_2 w_2}{w_1}\right)\partial x_2 \\ =& \frac{6q(w_1 + w_2) - 6q^2 - w_2(3w_1 + 2w_2)}{6w_1^2}\end{aligned} \tag{9}$$

Now, let us proceed to compute $\chi_2(\mathcal{G})$. We have that $\chi_2(\mathcal{G})$ equals

$$\int_0^1 \left(\int_0^1 \mathbb{I}(v(x_1, x_2) = 1)\partial x_2 \int_0^1 \mathbb{I}(v(x_1, y_2) = 0)\partial y_2\right)\partial x_1$$

which equals

$$\int_0^1 \left(\int_0^1 \mathbb{I}(x_2 \leq \frac{x_1 w_1 - q}{-w_2})\partial x_2 \int_0^1 \mathbb{I}(y_2 > \frac{x_1 w_1 - q}{-w_2})\partial y_2\right)\partial x_1 \tag{10}$$

Again, the internal integrals in (10) are not zero only if $\frac{x_1 w_1 - q}{-w_2} \in [0, 1]$. $\frac{x_1 w_1 - q}{-w_2} \geq 0$ if and only if $x_1 \geq \frac{q}{w_1}$, and $\frac{x_1 w_1 - q}{-w_2} \leq 1$ if and only if $x_1 \leq \frac{q - w_2}{w_1}$. This inequality is non-trivial only if $\frac{q-w_2}{w_1} < 1$, which happens only when $q < w_1 + w_2$. Thus, we again distinguish between the case when $q \geq w_1 + w_2$ and the case when $q < w_1 + w_2$. In the first case, we have

$$\begin{aligned}\chi_2(\mathcal{G}) =& \int_{\frac{q}{w_1}}^{1} \left(\frac{x_1 w_1 - q}{-w_2}\right)\left(1 - \frac{x_1 w_1 - q}{-w_2}\right)\partial x_2 \\ =& \frac{(w_1 - q)^2(2q - 2w_1 - 3w_2)}{6w_2^2 w_1}\end{aligned} \tag{11}$$

and in the second case, $\chi_2(\mathcal{G})$ equals

$$\int_{\frac{q}{w_1}}^{\frac{q-w_2}{w_1}} \left(\frac{x_1 w_1 - q}{-w_2}\right)\left(1 - \frac{x_1 w_1 - q}{-w_2}\right)\partial x_2 = \frac{-w_2}{6w_1} \tag{12}$$

Let us compare the values when $q \geq w_1 + w_2$.

$$\begin{aligned}
\chi_1(\mathcal{G}) \geq& \chi_2(\mathcal{G}) & \iff \\
\frac{(w_1-q)^2(2q+w_1)}{6(-w_2)w_1^2} \geq& \frac{(w_1-q)^2(2q-2w_1-3w_2)}{6w_2^2w_1} & \iff \\
\frac{2q+w_1}{w_1} \geq& \frac{2q-2w_1-3w_2}{-w_2} & \iff \\
(-w_2)(2q+w_1) \geq& w_1(2q-2w_1-3w_2) & \iff \\
w_1(w_1+w_2) \geq& q(w_1+w_2) & (13)
\end{aligned}$$

Thus, (13) holds with equality if $w_1 = -w_2$, $\chi_1(\mathcal{G}) > \chi_2(\mathcal{G})$ if $w_1 > -w_2$ (since $w_1 > q \geq 0$ by assumption), and $\chi_1(\mathcal{G}) < \chi_2(\mathcal{G})$ otherwise. For the second case, we have

$$\begin{aligned}
\chi_1(\mathcal{G}) \geq& \chi_2(\mathcal{G}) & \iff \\
\frac{6q(w_1+w_2)-6q^2-w_2(3w_1+2w_2)}{6w_1^2} \geq& \frac{-w_2}{6w_1} & \iff \\
\frac{6q(w_1+w_2)-6q^2-w_2(3w_1+2w_2)}{w_1} \geq& -w_2 & \iff \\
6q(w_1+w_2)-6q^2-w_2(3w_1+2w_2) \geq& (-w_2)w_1 & \iff \\
6q(w_1+w_2)-6q^2-2w_2(w_1+w_2) \geq& 0 & \iff \\
(3q-w_2)(w_1+w_2) \geq& 3q^2 & (14)
\end{aligned}$$

Now, (14) holds with equality if $w_1 + w_2 = 0$, since then $q = 0$ as well. Finally, if $w_1 + w_2 > 0$, then it holds with strict inequality since $w_1 + w_2 \geq q$ and $3q - w_2 > 3q$, and we are done.

Next, let us assume that $q < 0$. We again directly compute $\chi_1(\mathcal{G})$ and $\chi_2(\mathcal{G})$. First, if $w_2 \geq q$, then $x_1w_1 + x_2w_2 \geq x_2w_2 \geq w_2 \geq q$ for all $(x_1, x_2) \in [0,1]^2$; hence $\chi_1(\mathcal{G}) = \chi_2(\mathcal{G}) = 0$, and the claim trivially holds. We now assume that $w_2 < q$. Again, we have that $\chi_1(\mathcal{G})$ equals

$$\int_0^1 \left(\int_0^1 \mathbb{I}(x_1 \leq \frac{q - x_2w_2}{w_1}) \partial x_1 \int_0^1 \mathbb{I}(y_1 > \frac{q - x_2w_2}{w_1}) \partial y_1 \right) \partial x_2 \qquad (15)$$

We need to have $\frac{q-x_2w_2}{w_1} \in [0,1]$. $\frac{q-x_2w_2}{w_1} \geq 0$ if and only if $x_2 \geq \frac{q}{w_2}$. Since $w_2 < q$, this value is always less than 1. Moreover, $\frac{q-x_2w_2}{w_1} \leq 1$ if and only if $x_2 \leq \frac{q-w_1}{w_2}$. This inequality is not trivial only if $\frac{q-w_1}{w_2} \leq 1$, which happens whenever $q \geq w_2 + w_1$. Thus, when $q \geq w_1 + w_2$, $\chi_1(\mathcal{G})$ equals

$$\int_{\frac{q}{w_2}}^{\frac{q-w_1}{w_2}} \left(\frac{q - x_2w_2}{w_1} \right) \left(1 - \frac{q - x_2w_2}{w_1} \right) \partial x_2 = \frac{w_1}{-6w_2}$$

and when $q < w_1 + w_2$, $\chi_1(\mathcal{G})$ equals

$$\int_{\frac{q}{w_2}}^{1} \left(\frac{q - x_2 w_2}{w_1}\right)\left(1 - \frac{q - x_2 w_2}{w_1}\right)\partial x_2 = \frac{(q - w_2)^2(2q - 2w_2 - 3w_1)}{6 w_2 w_1^2}$$

For $\chi_2(\mathcal{G})$, we employ a similar reasoning. First, $\chi_2(\mathcal{G})$ equals

$$\int_0^1 \left(\int_0^1 \mathbb{I}(x_2 \leq \frac{x_1 w_1 - q}{-w_2})\partial x_2 \int_0^1 \mathbb{I}(y_2 > \frac{x_1 w_1 - q}{-w_2})\partial y_2\right)\partial x_1 \tag{16}$$

And again, $\frac{x_1 w_1 - q}{-w_2} \in [0, 1]$ if and only if $x_1 \leq \frac{q - w_2}{w_1}$. Note that since $w_2 < q$, $\frac{q - w_2}{w_1} \geq 0$. This constraint is only meaningful when $q < w_1 + w_2$. Thus, when $q \geq w_1 + w_2$, we have that $\chi_2(\mathcal{G})$ equals

$$\int_0^1 \left(\frac{x_1 w_1 - q}{-w_2}\right)\left(1 - \frac{x_1 w_1 - q}{-w_2}\right)\partial x_1 = -\frac{6q^2 - 6q(w_1 + w_2) + w_1(3w_2 + 2w_1)}{6 w_2^2}$$

and equals

$$-\frac{(q - w_2)^2(2q + w_2)}{6 w_2^2 w_1}$$

otherwise.

Next, we compare the values we obtained. When $q \geq w_1 + w_2$, we have that $w_1 + w_2 < 0$, and in particular, $|w_2| > |w_1|$. Moreover,

$$\begin{aligned}
-\frac{6q^2 - 6q(w_1 + w_2) + w_1(3w_2 + 2w_1)}{6 w_2^2} \geq& \frac{w_1}{-6 w_2} && \iff \\
\frac{-6q^2 + 6q(w_1 + w_2) - w_1(3w_2 + 2w_1)}{-w_2} \geq& w_1 && \iff \\
-6q^2 + 6q(w_1 + w_2) - w_1(2w_2 + 2w_1) \geq& 0 && \iff \\
(3q - w_1)(w_2 + w_1) \geq& 3q^2
\end{aligned}$$

Under our assumptions, this inequality holds, and we are done with the first case. For the second case,

$$\begin{aligned}
\frac{(q - w_2)^2(2q - 2w_2 - 3w_1)}{6 w_2 w_1^2} \geq& -\frac{(q - w_2)^2(2q + w_2)}{6 w_2^2 w_1} && \iff \\
\frac{2w_2 + 3w_1 - 2q}{w_1} \geq& -\frac{2q + w_2}{-w_2} && \iff \\
(-w_2)(2w_2 + 3w_1 - 2q) \geq& w_1(-2q - w_2) && \iff \\
(-w_2)(w_1 + w_2) \geq& (-q)(w_1 + w_2)
\end{aligned}$$

Since $w_2 < q$, this inequality holds with equality when $w_1 = -w_2$, it is strict whenever $|w_1| > |w_2|$, and the reverse holds when $|w_1| < |w_2|$. □

We are now ready to complete the proof of Theorem 3.1.

Proof of Theorem 3.1. We have shown the case where $w_i \geq w_j \geq 0$ in Theorem A.5. We have also shown this to be true for two features in Lemma A.7. We just need to show that Lemma A.7 extends to the case of arbitrary players. Suppose that $|w_i| > |w_j|$. Let us write $\chi_i(\langle N, \mathbf{w}; q\rangle)$ to be the influence of i under the linear classifier defined by $\mathbf{w}$ and q. We observe that

$$\begin{aligned}\chi_i(\langle N, \mathbf{w}; q\rangle) =& \int_{\mathbf{x}_{-i,-j}} \chi_i(\langle\{i,j\}, (w_i, w_j); q - \sum_{k\neq i,j} x_k w_k\rangle)\\ \geq& \int_{\mathbf{x}_{-i,-j}} \chi_j(\langle\{i,j\}, (w_i, w_j), q - \sum_{k\neq i,j} x_k w_k\rangle)\\ =&\chi_j(\langle N, \mathbf{w}; q\rangle)\end{aligned}$$

which concludes the proof. □

B Proof that χ satisfies (D), (Sym) and (DU)

We show that χ satisfies the three axioms. If $v(\mathbf{a}_{-i}, b) = v(\mathbf{a})$ for all $\mathbf{a} \in A$ and all $b \in A_i$, then $|v(\mathbf{a}_{-i}, b) - v(\mathbf{a})| = 0$, and in particular, $\chi_i(\mathcal{G}) = 0$; hence, χ satisfies the dummy property. Suppose we are given a bijection $\sigma_i : A_i \to A_i$. We observe that

$$\begin{aligned}\chi_i(\mathcal{G}) =&\frac{1}{|A|}\sum_{\mathbf{a}\in A}\sum_{b\in A_i} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|\\ =&\frac{1}{|A|}\sum_{\mathbf{a}_{-i}\in A_{-i}}\sum_{b'\in A_i}\sum_{b\in A_i} |v(\mathbf{a}_{-i}, \sigma_i(b)) - v(\mathbf{a}_{-i}, \sigma_i(b'))|\\ =&\frac{1}{|A|}\sum_{\mathbf{a}_{-i}\in A_{-i}}\sum_{b'\in A_i}\sum_{b\in A_i} |v_{\sigma_i}(\mathbf{a}_{-i}, b) - v_{\sigma_i}(\mathbf{a}_{-i}, b')|\\ =&\frac{1}{|A|}\sum_{\mathbf{a}\in A}\sum_{b'\in A_i}\sum_{b\in A_i} |v_{\sigma_i}(\mathbf{a}_{-i}, b) - v_{\sigma_i}(\mathbf{a})| = \chi_i(\sigma_i\mathcal{G})\end{aligned}$$

so χ is invariant under permutations of feature states. Similarly, for any bijection $\sigma : N \to N$, $\chi_i(\mathcal{G}) = \chi_{\sigma(i)}(\sigma\mathcal{G})$; therefore, χ satisfies symmetry.

Given a set $B \subseteq A$ and a feature i, let us write $W_{\bar{\mathbf{a}}_{-i}}(B) = \{\mathbf{a} \in B \mid v(\mathbf{a}) = 1, \mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}\}$, and $L_{\bar{\mathbf{a}}_{-i}}(B) = \{\mathbf{a} \in B \mid v(\mathbf{a}) = 0, \mathbf{a}_{-i} = \bar{\mathbf{a}}_{-i}\}$. We observe that $W_{\mathbf{a}_{-i}}(B) \cap W_{\bar{\mathbf{a}}_{-i}}(B) = L_{\mathbf{a}_{-i}}(B) \cap L_{\bar{\mathbf{a}}_{-i}}(B) = \emptyset$; moreover, $L(B) = \bigcup_{\mathbf{a}_{-i}\in A_{-i}} L_{\mathbf{a}_{-i}}(B)$ and $W(B) = \bigcup_{\mathbf{a}_{-i}\in A_{-i}} W_{\mathbf{a}_{-i}}(B)$. Now, given some $B \subseteq A$, let

us take some $W' \subseteq W(A) \setminus W(B)$.

$$\begin{aligned}\chi_i(W(B), L(B)) &= \sum_{\mathbf{a} \in B} \sum_{\substack{b \in A_i: \\ (\mathbf{a}_{-i}, b) \in B}} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})| \\ &= \sum_{\mathbf{a} \in W(B)} \sum_{\substack{b \in A_i: \\ (\mathbf{a}_{-i}, b) \in B}} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})| \\ &\quad + \sum_{\mathbf{a} \in L(B)} \sum_{\substack{b \in A_i: \\ (\mathbf{a}_{-i}, b) \in B}} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|\end{aligned}$$

Next, we observe that the first summand equals

$$\sum_{\mathbf{a} \in W(B)} \sum_{\substack{b \in A_i: \\ (\mathbf{a}_{-i}, b) \in B}} v(\mathbf{a}) - v(\mathbf{a}_{-i}, b),$$

which equals

$$\sum_{\mathbf{a}_{-i} \in A} \sum_{\mathbf{a} \in W_{\mathbf{a}_{-i}}(B)} \sum_{\substack{b \in A_i: \\ (\mathbf{a}_{-i}, b) \in B}} v(\mathbf{a}) - v(\mathbf{a}_{-i}, b) \tag{17}$$

Now, $v(\mathbf{a}) - v(\mathbf{a}_{-i}, b) = 1$ if and only if $v(\mathbf{a}_{-i}, b) = 0$; that is, if $(\mathbf{a}_{-i}, b) \in L_{\mathbf{a}_{-i}}(B)$. Thus, Equation (17) equals

$$\begin{aligned}&\sum_{\mathbf{a}_{-i} \in A} \sum_{\mathbf{a} \in W_{\mathbf{a}_{-i}}(B)} |L_{\mathbf{a}_{-i}}(B)| = \\ &\sum_{\mathbf{a}_{-i} \in A} |W_{\mathbf{a}_{-i}}(B)||L_{\mathbf{a}_{-i}}(B)|\end{aligned} \tag{18}$$

A similar construction with W' shows that

$$\chi_i(W', L(B)) = \sum_{\mathbf{a}_{-i} \in A_{-i}} |W'_{\mathbf{a}_{-i}}| \cdot |L_{\mathbf{a}_{-i}}(B)|;$$

since $W(B)$ and W' are disjoint, χ satisfies the disjoint union property.

C Relation to Classic Values in TU Cooperative Games

Our work generalizes influence measurement in classic TU cooperative games. We recall that a cooperative game with transferrable utility is given by a set of players $N = \{1, \ldots, n\}$, and a function $v : 2^N \to \mathbb{R}$, called the *characteristic function*. A game is defined by the tuple $\mathcal{G} = \langle N, v \rangle$. We say that a game $\mathcal{G}$ is *monotone* if for all $S \subseteq T \subseteq N$, $v(S) \leq v(T)$.

Classic literature identifies two canonical methods of measuring feature influence in cooperative games, the Shapley value [Shapley, 1953], and the Banzhaf value [Banzhaf,

1965]. We begin by providing the following definitions. Given a set $S \subseteq N$ and a player i, we let $m_i(S) = v(S \cup \{i\}) - v(S)$ denote the *marginal contribution* of i to S. The value $m_i(S)$ simply describes the added benefit of having i join the coalition S. Let $\Pi(N)$ be the set of all bijections from N to itself (also called the set of permutations of N); given some $\sigma \in \Pi(N)$ We let $P_i(\sigma) = \{j \in N \mid \sigma(j) < \sigma(i)\}$ be the set of the predecessors of i under σ. We define $m_i(\sigma) = v(P_i(\sigma) \cup \{i\}) - v(P_i(\sigma))$.

Definition C.1. The *Banzhaf value* of a player $i \in N$ is given by

$$\beta_i(\mathcal{G}) = \frac{1}{2^n} \sum_{S \subseteq N} m_i(S).$$

The Banzhaf value takes on a simple probabilistic interpretation: if we choose a set S uniformly at random from N, the Banzhaf value of a player is his expected marginal contribution to that set.

Rather than uniformly sampling sets, the Shapley value is based on uniformly sampling permutations.

Definition C.2. The *Shapley value* of a player $i \in N$ is given by

$$\varphi_i(\mathcal{G}) = \frac{1}{n!} \sum_{\sigma \in \Pi(N)} v(P_i(\sigma) \cup \{i\}) - v(P_i(\sigma)).$$

Intuitively, one can think of the Shapley value as the result of the following process. We randomly pick some order of the players; each player receives a payoff that is equal to his marginal contribution to his predecessors in the ordering. The Shapley value is simply the expected payoff a player receives in this scheme.

When we sample sets uniformly at random from $N \setminus \{i\}$, we are heavily biased towards selecting sets whose size is approximately $n/2$. When measuring influence according to the Shapley value, we are no longer biased towards any set size. One can think of the Shapley value is measuring a player's expected marginal contribution to a set S, where S is chosen according to the following process. First, we pick some $k \in \{0, \ldots, n-1\}$ uniformly at random, and then we pick a set of size k uniformly at random.

We observe that our classification setting is a generalization of TU cooperative games. Think of each player as a feature that can take on two values: 0 (corresponding to "absent"), and 1 (corresponding to "present"). An immediate observation is that ζ coincides with the Banzhaf value for TU cooperative games. Is there some natural extension of the Shapley value for general classification tasks?

Our work provides a negative answer to this question. We observe that Theorem D.1 states that the only value that satisfies the dummy, symmetry and linearity axioms is ζ. When reduced to the cooperative game setting, we obtain axioms that were used to axiomatically characterize both the Shapley and the Banzhaf values [Lehrer, 1988; Shapley, 1953; Young, 1985].

The dummy axiom (Definition 2.2) reduces to the following: a player $i \in N$ is a dummy if for all $S \subseteq N$, $v(S \cup \{i\}) = v(S)$. Thus, the dummy axiom requires that if a player is a dummy, then his value should be zero.

The symmetry axiom (Definition 2.1) reduces to the following: given a game $\mathcal{G} = \langle N, v\rangle$, and some $i, j \in N$, let us define $\mathcal{G}' = \langle N, v'\rangle$ as follows: for all $S \subseteq N \setminus \{i, j\}$, $v'(S) = v(S)$, and $v'(S \cup \{i, j\}) = v(S \cup \{i, j\})$; however, $v'(S \cup \{i\}) = v(S \cup \{j\})$ and $v'(S \cup \{j\}) = v(S \cup \{i\})$. A value ϕ satisfies symmetry if $\phi_i(\mathcal{G}) = \phi_j(\mathcal{G}')$. Symmetry reduces to saying that if we replace $v(S)$ with $v(S \setminus i)$ for all S such that $i \in S$, and replace $v(S)$ with $S \cup \{i\})$ for all S such that $i \notin S$, then the total influence of a player (i.e. his influence when being absent plus his influence when present) does not change.

Additivity as defined in Definition 2.3 is also naturally applied to TU cooperative games and is equivalent to the definition given in other axiomatic treatments of values in cooperative games.

It is well-known that both the Banzhaf and Shapley values satisfy the dummy, symmetry and additivity axioms, and indeed, Proposition 2.5 applies to them both: the Banzhaf value (and Shapley) of a player only measures the effect of player i joining a coalition, but not the effect of him leaving it. These two values, however, sum to 0. Indeed:

$$\begin{aligned}
\beta_{i,1}(\mathcal{G}) + \beta_{i,0}(\mathcal{G}) =& \frac{1}{2^n} \sum_{S \subseteq N} v(S \cup \{i\}) - v(S) \\
&+ \frac{1}{2^n} \sum_{S \subseteq N} v(S \setminus \{i\}) - v(S) \\
=& \frac{1}{2^n} \sum_{S \subseteq N \setminus \{i\}} v(S \cup \{i\}) - v(S) \\
&+ \frac{1}{2^n} \sum_{S \subseteq N \setminus \{i\}} v(S) - v(S \cup \{i\}) \\
=& 0
\end{aligned}$$

Theorem 2.8 characterizes χ as the unique value to satisfy the dummy, symmetry and disjoint union properties.

Going back to the classification setting, it is easy to see that Definition 2.6 implies that for $C \subseteq A$ and any two sets $B, B' \subseteq A \setminus C$, $\phi_i(B, C) + \phi_i(B', C) = \phi_i(B \cup B', C) + phi_i(B \cap B', C)$.

One can directly interpret the DU property in TU cooperative games. Given a game $\mathcal{G} = \langle N, v\rangle$ and a subset $\mathcal{B}$ of 2^N, both the Shapley and Banzhaf values can be defined to ignore any elements that are not contained in $\mathcal{B}$. It is easy to see that Theorem 2.8 implies the uniqueness of χ for TU cooperative games, and that it equals the Banzhaf value. Thus, Theorem 2.8 can be seen as an alternative axiomatization of the Banzhaf value, this time from the binary classification perspective.

D Axiomatic Approach to State Influence

Section 2 provided an answer to questions of the following form: what is the impact of gender on classification. The answer provided in previous sections was that influence was a function of the feature's ability to change outcomes by changing its state.

It is also useful to ask a related question: suppose that a certain search engine user is profiled as a female. What is the influence of this profiling decision? In other words, rather than measuring feature influence, we are measuring the influence of feature i being in a certain *state*.

For a feature $i \in N$ and a state $b \in N$, we can ask what is the influence of the state b, rather than the influence of i. That is, rather than having a value $\phi_i(\mathcal{G})$ for a feature $i \in N$, we now study the influence of the state $b \in A_i$, i.e. a real value $\phi_{i,b}(\mathcal{G})$ for each $i \in N$ and $b \in A_i$.

While Proposition 2.5 implies that any *feature* influence measure that satisfies the dummy, symmetry and additivity axioms must be trivial, this result does not carry through to measures of state influence.

Dummy (D): given $i \in N$ and $b \in A_i$, we say that α satisifies the dummy property if whenever $v(\mathbf{a}_{-i}, b) = v(\mathbf{a})$ for all $\mathbf{a} \in A$, $\alpha_{i,b} = 0$.

Symmetry (Sym): Two states $b, b' \in A_i$ are symmetric if for all $\mathbf{a} \in A$, $v(\mathbf{a}_{-i}, b) = v(\mathbf{a}_{-i}, b')$. A value α satisfies symmetry if $\alpha_{i,b} = \alpha_{i,b'}$ whenever b and b' are symmetric.

Linearity (L): Given games $\mathcal{G}_1 = \langle N, A, v_1 \rangle$ and $\mathcal{G}_2 = \langle N, A, v_2 \rangle$, let us write $\mathcal{G} = \langle N, A, v \rangle$ where $v = v_1 + v_2$. We assume that v_1 and v_2 are such that v is still a function with binary values (i.e. if $v_1(\mathbf{a}) = 1$ then $v_2(\mathbf{a}) = 0$). A value α is linear if $\alpha_{i,b}(\mathcal{G}) = \alpha_{i,b}(\mathcal{G}_1) + \alpha_{i,b}(\mathcal{G}_2)$.

Let us define

$$\zeta_{i,b}(\mathcal{G}) = \frac{1}{|A|} \sum_{\mathbf{a} \in A} v(\mathbf{a}_{-i}, b) - v(\mathbf{a}) \tag{19}$$

We let $\bar{\zeta}$ denote the value ζ without the normalizing factor $\frac{1}{|A|}$. We refer to $\bar{\zeta}$ as the *raw* version of ζ. In Theorem D.1, we show that $\bar{\zeta}$ is the unique (up to a constant) value that satisfies the symmetry, dummy and linearity axioms.

Theorem D.1. *If a value ϕ satisfies the (D), (Sym), and (L), then $\phi = c\zeta$, where c is an arbitrary constant.*

Proof. Let us observe that every game $v : A \to \{0, 1\}$ can be written as the disjoint sum of unanimity games; namely $v = \sum_{\mathbf{a} \in A : v(\mathbf{a}) = 1} u_{\mathbf{a}}$. Thus, it suffices to show that the claim holds for unanimity games.

Let $\mathcal{U}_{\mathbf{a}} = \langle N, A, u_{\mathbf{a}} \rangle$; we show that $\phi_{i,b}(\mathcal{U}_{\mathbf{a}})$ equals $\zeta_{i,b}(\mathcal{U}_{\mathbf{a}})$. First, if $b = a_i$ then $\bar{\zeta}_{i,b}(\mathcal{U}_{\mathbf{a}}) = |A_i| - 1$; if $b \neq a_i$, then $\bar{\zeta}_{i,b}(\mathcal{U}_{\mathbf{a}}) = -1$. Now, by symmetry, we have that $\phi_{i,b}(\mathcal{U}_{\mathbf{a}}) = \phi_{i,b'}(\mathcal{U}_{\mathbf{a}})$ for all $b, b' \neq a_i$. If we write $\phi_{i,b}(\mathcal{U}_{\mathbf{a}}) = y$ for all $b \neq a_i$, and $\phi_{i,a_i}(\mathcal{U}_{\mathbf{a}}) = x$, then according to Proposition 2.5, $\sum_{b \neq a_i} y + x = 0$, which implies that $x = -y(|A_i| - 1)$. Finally, according to feature symmetry, the value of y cannot depend on i, and is equal for all $j \in N$. We conclude that for all $i \in N$ and all $b \in A_i$, $\phi_{i,b}(\mathcal{G}) = \zeta_{i,b}(\mathcal{G})$. □

As a direct corollary of Theorem A.5, we have that the unique (up to a constant) state value to satisfy (Sym), (D) and (DU) axioms (see Definitions 2.1, 2.2 and 2.6 in Section 2) is

$$\chi_{i,b}(\mathcal{G}) = \sum_{\mathbf{a} \in A} |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|.$$

E Influence in Weighted Settings

Unlike previous sections, let us assume that there is some weight function $w : A \to \mathbb{R}$ that assigns a non-negative weight to every state vector. w can be thought of as a prior distribution that governs the likelihood of observing a state vector $\mathbf{a} \in A$. Given $B \subseteq A$, let $w(B)$ denote $\sum_{\mathbf{a} \in B} w(\mathbf{a})$. We also write for a given $b \in A_i$, $w(b \mid \mathbf{a}_{-i}) = \sum_{\mathbf{a}_{-i} \in A_{-i}} w(\mathbf{a}_{-i}, b)$; for a given $\mathbf{a}_{-i} \in A_{-i}$, we write $w(\mathbf{a}_{-i}) = \sum_{b \in A_i} w(\mathbf{a}_{-i}, b)$. Given this definition, let us rethink the disjoint union property. Given a set of winning state vectors $W \subseteq A$ and a set of losing state vectors $L \subseteq A$, we can think of a weighted influence measure as a function ϕ_i of W, L and $w : A \to \mathbb{R}_+$.

Fix some $C \subseteq A$. Given two functions $w, w' : A \to \mathbb{R}_+$ that agree on C (i.e. $w(\mathbf{a}) = w'(\mathbf{a})$ for all $\mathbf{a} \in C$), and some $B \subseteq A \setminus C$, let us write

$$w \oplus_B w'(\mathbf{a}) = \begin{cases} w(\mathbf{a}) & \text{if } \mathbf{a} \in C \\ w(\mathbf{a}) + w'(\mathbf{a}) & \text{if } \mathbf{a} \in B. \end{cases}$$

Definition E.1. We say that an influence measure satisfies *weighted disjoint union* (WDU) if for any disjoint $B, C \subseteq A$ and any two weight functions $w, w' : A \to \mathbb{R}_+$ that agree on C, we have that $\phi_i(B, C, w) + \phi_i(B, C, w') = \phi_i(B, C, w \oplus_B w')$.

Lemma E.2. *Weighted disjoint union implies the disjoint union property.*

We again write $W_{\mathbf{a}_{-i}} = \{(\mathbf{a}_{-i}, b) \in A \mid v(\mathbf{a}_{-i}, b) = 1\}$, and $L_{\mathbf{a}_{-i}} = \{(\mathbf{a}_{-i}, b) \in A \mid v(\mathbf{a}_{-i}, b) = 0\}$.

Given a weight function $w : A \to \mathbb{R}_+$ and a game $\mathcal{G} = \langle N, A, v \rangle$, let

$$\chi^p(\mathcal{G}, w) = \sum_{\mathbf{a} \in A} w(\mathbf{a}) \sum_{b \in A_i} w(b|\mathbf{a}_{-i}) |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|.$$

Let us extend the symmetry axiom (Definition 2.1) to a weighted variant. Given a weight function $w : A \to \mathbb{R}_+$ and a bijection σ over A_i or N, we let $\sigma w(\mathbf{a}) = w(\sigma \mathbf{a})$.

Definition E.3. Given a game $\mathcal{G} = \langle N, A, v \rangle$ and a weight function $w : A \to \mathbb{R}$, we say that an influence measure ϕ is *state-symmetric* with respect to w (Sym-w) if for any permutation $\sigma : A_i \to A_i$, and all $j \in N$, $\phi_j(\sigma \mathcal{G}, \sigma w) = \phi_j(\mathcal{G}, w)$. That is, relabeling the states and letting them keep their original distributions does not change the value of any feature. Similarly, we say that an influence measure ϕ is *feature-symmetric* if for any permutation $\sigma : N \to N$, $\phi_{\sigma(i)}(\sigma \mathcal{G}, \sigma w) = \phi_i(\mathcal{G}, w)$. That is, relabeling the coordinate of a feature does not change its value.

Theorem E.4. *If a probabilistic influence measure ϕ satisfies (D), (Sym) and (DU) with respect to some $\mathcal{D}$, then*

$$\phi_i(\mathcal{G}, \mathcal{D}) = C\chi^p(\mathcal{G}, \mathcal{D}).$$

Before we proceed, we wish to emphasize two important aspects of Theorem E.4. First, if we set $p(\mathbf{a}) = \frac{1}{|A|}$ then we obtain Theorem 2.8. In other words, χ is an influence measure that assumes that all elements in the dataset are equally likely.

Another point of note is the underlying process that the influence measures entail. If we assume that the weight function describes a distribution over A, one can think of the influence measure as the following process. We begin by picking a point from A at random (uniformly at random in the case of χ, and according to w in Theorem E.4); next, *fixing* the states of all other features, we measure the probability that i can change the outcome, by sampling a different state according to the distribution $w(\cdot \mid \mathbf{a}_{-i})$.

Before we prove Theorem E.4, let us prove the following lemma.

Lemma E.5.

$$\chi^p(\mathcal{G}, w) = 2 \sum_{\mathbf{a}_{-i} \in A_{-i}} w(\mathbf{a}_{-i}) w(W_{\mathbf{a}_{-i}}) w(L_{\mathbf{a}_{-i}})$$

Proof.

$$\begin{aligned}
\chi^p(\mathcal{G}) &= \sum_{\mathbf{a} \in A} w(\mathbf{a}) \sum_{b \in A_i} w(b \mid \mathbf{a}_{-i}) |v(\mathbf{a}_{-i}, b) - v(\mathbf{a})| \\
&= 2 \sum_{\mathbf{a}_{-i} \in A} w(\mathbf{a}_{-i}) \sum_{\substack{c \in A_i: \\ v(\mathbf{a}_{-i}, c) = 0}} \sum_{\substack{b \in A_i: \\ v(\mathbf{a}_{-i}, b) = 1}} w(c \mid \mathbf{a}_{-i}) w(b \mid \mathbf{a}_{-i}) \\
&= 2 \sum_{\mathbf{a}_{-i} \in A} w(\mathbf{a}_{-i}) \sum_{\substack{c \in A_i: \\ v(\mathbf{a}_{-i}, c) = 0}} w(c \mid \mathbf{a}_{-i}) w(W_{\mathbf{a}_{-i}}) \\
&= 2 \sum_{\mathbf{a}_{-i} \in A} w(\mathbf{a}_{-i}) w(L_{\mathbf{a}_{-i}}) w(W_{\mathbf{a}_{-i}})
\end{aligned}$$

□

Lemma E.6. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function that satisfies*

(i) $f(x, 0) = f(0, y) = 0$.

(ii) $f(x_1, y) + f(x_2, y) = f(x_1 + x_2, y)$.

(iii) $f(x, y_1) + f(x, y_2) = f(x, y_1 + y_2)$.

Then there is some constant c such that $f(x, y) = cxy$.

Proof. First, we show that $f(rx, y) = rf(x, y)$ for all $r \in \mathbb{R}$. Given any $n \in \mathbb{N}$, $f(nx, y) = nf(x, y)$ by property (2). Similarly, $f(\frac{x}{n}, y) = \frac{1}{n} f(x, y)$. Thus, for any rational number $q \in \mathbb{Q}$, we have $f(qx, y) = f(x, qy) = qf(x, y)$. Now, take

any real number $r \in \mathbb{R}$. There exists a sequence of rational numbers $(q_n)_{n=1}^{\infty}$ such that $\lim_{n\to\infty} q_n = r$. Thus, $f(rx, y) = \lim_{n\to\infty} f(q_n x, y) = \lim_{n\to\infty} q_n f(x, y) = rf(x, y)$ (and similarly $f(x, ry) = rf(x, y)$).

Let us observe the partial derivatives of f at $x \neq 0$:

$$\begin{aligned}\frac{\partial f}{\partial x}(x^*, y^*) &= \lim_{\varepsilon\to 0} \frac{f(x^* + \varepsilon, y^*) - f(x^*, y^*)}{\varepsilon} \\ &= \lim_{\varepsilon\to 0} \frac{(\frac{x^*+\varepsilon}{x^*} - 1)f(x^*, y^*)}{\varepsilon} = \frac{f(x^*, y^*)}{x^*}\end{aligned}$$

and similarly $\frac{\partial f}{\partial y}(x^*, y^*) = \frac{f(x^*, y^*)}{y^*}$. We obtain the following differential equation: $x\frac{\partial f}{\partial x} - f = 0$. Its only solution is $f(x, y) = g(y)x + h(y)$. However, since $f(0, y) = 0$ for all y, we get that $h(y) \equiv 0$. Similarly, $f(x, y) = k(x)y$. Putting it all together, we get that $f(x, y) = cxy$.

□

Lemma E.7. *If a value ϕ satisfies the (WDU) and (Sym-w) property, then it agrees with χ^p on any game $\mathcal{G} = \langle\{i\}, A_i, v\rangle$ with any weight function $w : A_i \to \mathbb{R}_+$*

Proof. Let us write W_i and L_i to be the winning and losing states in A_i. By state symmetry we know that ϕ is only a function of $(w(b))_{b\in W_i}$ and $(w(b))_{b\in L_i}$. By the weighted disjoint union property, we know that

$$\phi_i((w(b))_{b\in W_i}, (w(b))_{b\in L_i}) = \sum_{b\in W_i}\sum_{c\in L_i} \phi_i(w(b), w(c)).$$

Using the (WDU) property, we know that the following holds for single-feature games with only two states. Given $x_1, x_2, y \in \mathbb{R}_+$, the following holds:

$$\begin{aligned}\phi_i(x_1 + x_2, y) =& \phi_i(x_1, y) + \phi_i(x_2, y) \\ \phi_i(y, x_1 + x_2) =& \phi_i(y, x_1) + \phi_i(y, x_2)\end{aligned}$$

By Lemma E.7, we know that $\phi_i(x, y) = cxy = c\chi_i^p(x, y)$. In particular, this implies that $\phi_i(\mathcal{G}, w) = \chi_i^p(\mathcal{G}, w)$, and we are done. □

Proof of Theorem E.4. First, we note that χ^p satisfies (D), (Sym-w) and (WDU) (this is an easy exercise). We write W to be the winning state vectors in A and L to be the losing state vectors in A. Now, if either $w(W) = 0$ or $w(L) = 0$, any influence measure that satisfies (D) assigns a value of zero to all $i \in N$, and the claim trivially holds. Thus, we assume that $w(W), w(L) > 0$.

Next, according to the (DU) property, we can write

$$\phi_i(W, L, w) = \sum_{\mathbf{a}_{-i}\in A_{-i}} \phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}}, w).$$

The argument is the same as the one used for the decomposition of χ in Theorem 2.8. By the above lemmas, $\phi_i(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}}, w) = C\chi_i^p(W_{\mathbf{a}_{-i}}, L_{\mathbf{a}_{-i}}, w)$. Note that by feature symmetry, it must be the case that the constant C is independent of i. □

F Generalized Distance Measures

Suppose that we have a set of feature vectors $B \subseteq A$. In previous sections we had assumed that there was some function $v : A \to \{0, 1\}$ that classified a vector as either having a value of 0 or a value of 1. We then proceeded to provide an axiomatic characterization of influence measures in such settings. Influence was largely based on the following notion: a feature $i \in N$ can influence the vector $\mathbf{a} \in A$, if $|v(\mathbf{a}_{-i}, b) - v(\mathbf{a})| = 1$. Let us now consider a more general setting; instead of defining a classifier over data points, we have some semi-distance measure over the vectors. Recall that a pseudo-distance measure is a function $d : A \times A \to \mathbb{R}$ that satisfies all of the distance axioms, but $d(\mathbf{a}, \mathbf{b}) = 0$ does not necessarily imply that $\mathbf{a} = \mathbf{b}$. Given some pseudo-distance measure d over A, rather than measuring influence by the measure $|v(\mathbf{a}_{-i}, b) - v(\mathbf{a})|$, we measure influence by $d((\mathbf{a}_{-i}, b), \mathbf{a})$.

We observe that if $d(\mathbf{a}, \mathbf{b}) \in \{0, 1\}$ for all $\mathbf{a}, \mathbf{b} \in A$, then we revert to the original setting.

Given a pseudo-distance measure d over A and a dataset $B \subseteq A$, let us define $\mathcal{P}_d(B)$ to be the partition of B into the equivalence classes defined by $\mathbf{a} \sim \mathbf{b}$ iff $d(\mathbf{a}, \mathbf{b}) = 0$. In other words, $\mathcal{P}_d(B)$ is the clustering of B into points that are of equal distance to each other. Fixing a pseudo-distance d, we provide the following extensions of the axioms defined in Section 2.

We keep the notion of symmetry used in Section 2 (Definition 2.1): an influence measure satisfies symmetry if it is invariant under coordinate permutations, both for individual features (e.g. renaming males to females and vice versa should not change the influence of any feature), and between the features (e.g. renaming gender and age should not change feature influence). We do, however, adopt more general definitions of the dummy and disjoint union properties.

Definition F.1 (d-Dummy)**.** We say that an influence measure satisfies the d-Dummy property if $\phi_i(B) = 0$ whenever $d((\mathbf{a}_{-i}, b), \mathbf{a}) = 0$ for all $\mathbf{a} \in B$ and all $b \in A_i$ such that $(\mathbf{a}_{-i}, b) \in B$.

Definition F.2 (Feature Independence)**.** Let $B \subseteq A$ be a dataset, and let $B(\mathbf{a}_{-i}) = \{\mathbf{b} \in B \mid \mathbf{b}_{-i} = \mathbf{a}_{-i}\}$. An influence measure satisfies feature independence (FD) if

$$\phi_i(B) = \sum_{\mathbf{a}_{-i} \in A_{-i}} \phi_i(B(\mathbf{a}_{-i})).$$

Definition F.3 (d-Disjoint Union)**.** Let $B \subseteq A$ be a dataset, and let $\mathcal{B} = \{B_1, \ldots, B_m\}$ be the equivalence classes of B according to the pseudo-distance d. An influence measure ϕ satisfies the d-disjoint union, if for any $j \in \{1, \ldots, m\}$, any partition C, C' of B_j satisifies

$$\phi_i(B_1, \ldots, B_m) = \phi_i(\mathcal{B}_{-j}, C) + \phi_i(\mathcal{B}_{-j}, C') - \phi_i(\mathcal{B}_{-j}).$$

Finally, the following axiom requires that in very minimal settings, a feature's influence should agree with d.

Definition F.4 (Agreement with Distance)**.**

Given a dataset $B \subseteq A$, define

$$\chi_i^d(B) = \sum_{\mathbf{a} \in B} \sum_{b \in A_i : (\mathbf{a}_{-i}, b) \in B} d((\mathbf{a}_{-i}, b), \mathbf{a}) \tag{20}$$

Lemma F.5. *Let B be a dataset of single-feature points. Then if ϕ satisfies, d-(D), d-(DU), (Sym), and (AD), then $\phi(B) = \chi^d(B)$*

Proof Sketch. We partition B into its equivalence classes according to d, $\mathcal{B} = \{B_1, \ldots, B_m\}$. In an argument similar to Lemma 2.7, we can show that the symmetry axiom implies that ϕ is a function of $|B_1|, \ldots, |B_m|$. Let $w_j = |B_j|$; employing the d-disjoint union property and the dummy property, we obtain that there exists some $m \times m$ matrix D' such that $\phi(B) = \mathbf{w}^T D' \mathbf{w}$, and D' is 0 on the diagonal, non-negative, and symmetric (symmetry here is obtained via state symmetry).

To show that D' must identify with the pseudo-distance, we employ the agreement with distance axiom on inputs to ϕ that have only two non-zero coordinates, to obtain the desired result. □

Theorem F.6. *If an influence measure ϕ satisfies the d-dummy, d-disjoint union, symmetry and agreement with distance axioms, then*

$$\phi_i(B) = \alpha \sum_{\mathbf{a} \in B} \sum_{b \in A_i : (\mathbf{a}_{-i}, b) \in B} d((\mathbf{a}_{-i}, b), \mathbf{a}),$$

where α is a constant independent of i.

Proof Sketch. The proof mostly follows the proof technique of Theorem 2.8. Let us write the influence of i under d to be $\phi_i^d(A)$.

Using the (FI) property, we decompose ϕ_i^d into $|A_{-i}|$ different single-feature datasets. Next, we apply Lemma F.5 on each of the datasets to show that identity holds. □