Information Theoretic Measures for Fairness-aware Feature Selection

Sajad Khodadadian[*]   Mohamed Nafea[†]   AmirEmad Ghassami[‡]   Negar Kiyavash[§]

June 9, 2021

Abstract

Machine learning algorithms are increasingly used for consequential decision making regarding individuals based on their relevant features. Features that are relevant for accurate decisions may however lead to either explicit or implicit forms of discrimination against unprivileged groups, such as those of certain race or gender. This happens due to existing biases in the training data, which are often replicated or even exacerbated by the learning algorithm. Identifying and measuring these biases at the data level is a challenging problem due to the interdependence among the features, and the decision outcome. In this work, we develop a framework for fairness-aware feature selection which takes into account the correlation among the features and the decision outcome, and is based on information theoretic measures for the accuracy and discriminatory impacts of features. In particular, we first propose information theoretic measures which quantify the impact of different subsets of features on the accuracy and discrimination of the decision outcomes. We then deduce the marginal impact of each feature using Shapley value function; a solution concept in cooperative game theory used to estimate marginal contributions of players in a coalitional game. Finally, we design a fairness utility score for each feature (for feature selection) which quantifies how this feature influences accurate as well as nondiscriminatory decisions. Our framework depends on the joint statistics of the data rather than a particular classifier design. We examine our proposed framework on real and synthetic data to evaluate its performance.

1 Introduction

Machine learning algorithms are increasingly utilized in many domains of human life such as advertising, healthcare, loan assessment, job applications, and predictive policing [29, 33, 2]. While learning algorithms may help improve prediction accuracy in these applications, there is a growing concern about potential discriminatory practices against underrepresented groups. These practices often result due to existing biases in training data which are replicated (or even exacerbated) by the learning algorithm. Due to legal and/or ethical implications of these decisions[1], mere improvements in reducing risk or decision error does not suffice, and the system should also take into account the potential discriminatory consequences of a decision.

We consider the fairness problem that arises in supervised learning. In this setting, the goal is to assign an accurate label (decision) to each individual based on a set of features she/he possesses, while hindering certain features, referred to as *protected attributes*, from influencing the decisions, [6, 16, 20, 44, 22]. The law recognizes two doctrines of discrimination: (1) Disparate treatment (explicit discrimination) where the protected attribute (group membership) is deliberately used to treat underrepresented groups differently, and (2) Disparate impact (implicit discrimination) where the protected attribute is hidden from the decision maker but other features act as proxies for the group membership, unfavoring those of certain demographic [20, 44, 3, 13, 43]). For example, redlining, a systematic denial of some services such as banking, insurance, or healthcare to residents of certain neighborhoods, was adopted for decades in certain urban areas in the United States; here, the zip code acts as a proxy for the individual's race [21].

A candid approach to alleviate discriminatory decisions is to identify features with significant proxy behaviour and prevent these features from influencing the decision outcome, see for example [40]. However, a feature that acts as a

---

[*]H. Milton Stewart School of Industrial & Systems Engineering, Georgia Institute of Technology, skhodadadian3@gatech.edu

[†]Department of Electrical Engineering and Computer Science, University Of Detroit Mercy, nafeamo@udmercy.edu

[‡]Department of Computer Science, Johns Hopkins University, aghassa1@jhu.edu

[§]College of Management of Technology, École Polytechnique Fédérale de Lausanne (EPFL), negar.kiyavash@epfl.ch

[1]As a legal example, Title VII of the Civil Rights Act of 1964 prohibits employers from discriminating against employees on the basis of features such as race, gender, religion, etc.

proxy for the sensitive attributes(s) may contain information necessary for accurate classification. The question then is how to simultaneously quantify the marginal impact of each feature both on the accuracy and discrimination of the system? By answering this question, one can choose to select the subset of features that optimally satisfies certain accuracy/fairness requirements.

In this paper, we propose two information-theoretic measures that separately quantify the accuracy and discriminatory impact of subsets of features. Subsequently, we deduce the marginal impacts of each feature using Shapely-value analysis [37]. Introducing the measures for subsets of features, rather than for individual features, allows us to account for the interdependencies among the features, sensitive attributes, and decision outcome. Finally, we define a fairness-utility score for each feature which combines both impacts for each feature. Using these scores, we can choose to select the subset of features that optimally satisfy certain accuracy/fairness requirements.

Information-theoretic measures are advantageous because they (i) capture non-linear dependencies among the system variables [8] and (ii) allow the system variables, particularly the protected attribute(s) and the true outcome, to have arbitrary cardinalities (as opposed to [20, 44]). Further, our measures are defined with respect to the features, the sensitive attributes, and the true outcome. Thus, our method only requires knowledge of their joint statistics, and is not limited to a particular choice of the classifier.

1.1   Related work

Fairness. The problem of algorithmic fairness and various notions for nondiscriminative learning, both in the context of supervised learning [12, 13, 20], and otherwise [23, 24], have been the focus of recent work. Several methods have been proposed to address disparate impact (implicit discrimination), see [26, 12, 44, 13, 20, 28, 39]. These references adopted different notions of fairness, and proposed algorithms to mitigate unwanted biases. Notions of fairness can be coarsely divided into (i) *group fairness* which typically fixes some protected attributes (group memberships) and requires parity for a certain statistical measure across these groups [20, 30, 31], and (ii) *individual fairness*, which requires that similar individuals should be treated similarly, for certain similarity measures between individuals, as well as between classification outcomes [12, 44, 24]. Furthermore, approaches for mitigating discrimination bias can be generally categorized into (i) pre-processing methods, which modify the distribution of the training data, see [25, 19, 13, 6]; (ii) in-processing methods, which modify the cost function or the constraints of the learning algorithm, see [5, 14, 26, 42]; (iii) post-processing methods, which modify the prediction outcome, see [20, 36], and finally, (iv) causal reasoning, which introduces the concepts of counterfactual and interventional fairness, see [28, 31, 35]. This work adopts a group fairness approach, and classifies as a pre-processing method for bias mitigation. Specifically, we introduce a fairness-utility score for each (non-sensitive) feature, based on information theoretic measures, and use these scores for fairness-aware feature selection: Features with high discriminatory and low accuracy impact shall be removed. Further, we relate our proposed measure for the discriminatory impact of features to the existing notions of group fairness such as statistical parity and equalized odds (see Section 2.4). Finally, we provide causal reasoning for the desired properties of the measures we construct for the accuracy and discriminatory impacts of features.

Fairness-aware feature selection. Feature selection under fairness constraints has been previously studied by a few works. [18] proposed a feature selection criteria based on the populations' consensus on whether features are unfair to use, and investigated the impact of removing these features on prediction accuracy. In [27], the authors considered a submodular robust optimization problem, in which a set function is maximized under a set-size constraint, despite deletions in data points. The authors examined their proposed algorithms on a feature selection setting where the sensitive features (protected attributes) are deleted and the objective function is the mutual information between subsets of features and the true label. In [16], the authors proposed an optimization framework where the objective is to find the optimal compressed version of the features which maximizes the mutual information with the true label, while their mutual information with the protected attribute, given the true label, is upper bounded. In [15], the authors considered the problem of integrating new features to an existing training dataset, and proposed an algorithm to identify the new features that can be added to the original training data while still ensuring interventional fairness.

Transparency and Shapely value function. There is a growing interest in transparency and interpretability analysis of machine learning systems [9, 10, 17]. In [9], the authors proposed a framework for transparency analysis which is easy to interpret for humans. This is achieved by quantifying the individual impacts of features on various metrics of a decision maker, such as decision error, risk, or discrimination. Our measures for accuracy and discriminatory impacts can also be viewed as measures for transparency. Besides, there has been a surge in using Shapely value for transparent

machine learning, i.e., quantifying the importance of features for accurate and/or fair decisions of a given classifier [9, 32, 34]. In addition, a few work proposed using Shapely value for feature selection, without fairness considerations, see [7, 38]. In this work, we also use Shapely value function to deduce marginal impacts of features, while accounting for their interdependencies. However, our main contribution in this work lies in constructing *information theoretic* measures for the accuracy and discriminatory impacts of features, which do not depend on a certain choice of the classifier.

Information theoretic measures for fairness. Another related work is [11] which proposed using information theoretic tools to measure the non-exempt and exempt components of discrimination: The non-exempt discrimination component quantifies the part of discrimination that cannot be accounted for by the features critical for accuracy, and the exempt component quantifies the remaining part of discrimination. In [1], the authors proposed an information theoretic formulation for "model projection", where a reference probablistic classifier is projected to the set of classifiers that satisfy certain fairness criteria.

1.2 Challenges and our contributions

Our goal is to develop a framework for fairness-aware feature selection by computing a score for each feature which captures both its accuracy and discriminatory impacts. Thus, we need to define measures which (i) precisely quantify the accuracy and discriminatory impacts of a feature (or a subset of features), and (ii) capture the inter-dependencies among the features, the protected attribute, and the true outcome. These measures depend only on the joint statistic of the data and not on the particular classifier at hand.

We tackle these challenges by proposing information-theoretic accuracy and discrimination measures that depend only on the joint distribution of the data. In order to account for the *correlation* among the model variables, these measures are defined on subsets of features rather than a single feature. As our ultimate goal is to estimate the marginal impact of each feature, we propose extracting a score for each feature using the so-called Shapley value [37], a concept from cooperative game theory that allows assigning values to quantify an individual's contribution to the game. Our measures are based on a bivariate decomposition of mutual information [4], in order to precisely quantify the accuracy and discriminatory impacts.

The remainder of this paper is organized as follows. Section 2 introduces accuracy and discrimination measures, defined for subsets of features. In Section 3, we compute the aggregate discriminatory and accuracy impact of each feature. Section 4 discusses experimental results. All the proofs are deferred to the appendix.

2 Accuracy and discrimination measures

As mentioned earlier, we seek information-theoretic measures. But generic measures such as mutual information are not suitable for estimating the accuracy and discriminatory effects of a subset of features. We clarify this in Sections 2.3 and 2.4. Instead, we propose information-theoretic measures based on a bivariate decomposition of mutual information [4], which are indeed able to meet our requirements.

2.1 Bivariate decomposition of information

Consider an arbitrary triple of random variables R_1, R_2, and T, and let P_{T,R_1,R_2} denote their joint probability distribution. The amount of information R_1 and R_2 contain about T is measured by the mutual information $I(T; R_1, R_2)$. [4] proposed a non-negative decomposition of $I(T; R_1, R_2)$ into unique, shared, and synergistic information; see Figure 1. The unique information of R_1 with respect to T, denoted by $UI(T; R_1 \setminus R_2)$, represents the information content related to T that is only available in R_1. The shared information of R_1 and R_2, denoted by $SI(T; R_1, R_2)$, represents the information content related to T that both R_1 and R_2 possess. Finally, the synergistic information of R_1 and R_2, denoted by $CI(T; R_1, R_2)$, represents the information content that can be obtained only if both R_1 and R_2 are available. For instance, in a cryptographic system, both the secret key and the cypher text are needed to decode the encrypted message. According to Figure 1, this decomposition of mutual information satisfies the following:

$$I(T; R_1, R_2) = UI(T; R_1 \setminus R_2) + UI(T; R_2 \setminus R_1) + SI(T; R_1, R_2) + CI(T; R_1, R_2), \qquad (2.1)$$

$$I(T; R_i) = UI(T; R_i \setminus R_j) + SI(T; R_1, R_2), \ i \neq j, \ i, j \in \{1, 2\}. \qquad (2.2)$$
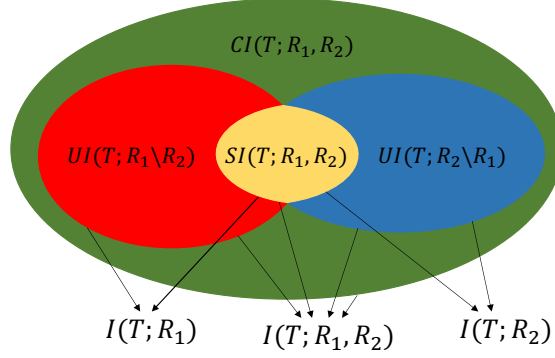
Figure 1: Decomposition of Information.

By defining a measure for either unique, shared, or synergistic information, and using the decomposition in (2.1), (2.2), the other two quantities are well defined. We use the following measure for unique information [4]:

$$UI(T; R_1 \backslash R_2) = \min_{Q \in \Delta_P} I_Q(T; R_1 | R_2), \qquad (2.3)$$

where $I_Q(T; R_1 | R_2) = \sum_{t,r_1,r_2} Q_{T,R_1,R_2}(t, r_1, r_2) \log \frac{Q_{T|R_1,R_2}(t|r_1,r_2)}{Q_{T|R_2}(t|r_2)}$ is the conditional mutual information between T and R_1 given R_2, calculated with respect to the joint probability distribution $Q \in \Delta_P$. $\Delta_P = \{Q \in \Delta : Q_{T,R_1} = P_{T,R_1}, Q_{T,R_2} = P_{T,R_2}\}$, where Δ is the simplex of joint probability distributions over T, R_1, R_2.

Using the unique information definition in (2.3), and the information decomposition in (2.1)–(2.2), we can uniquely define the shared and synergistic information. In Sections 2.3, 2.4, we use the information decomposition in (2.1)–(2.3) to construct information-theoretic measures for the accuracy and discriminatory impacts of subsets of features.

2.2 Problem formulation

We consider a supervised learning setting in which each individual in the dataset is associated with the protected attribute(s) $A \in \mathcal{A}$, and a set of n features $X^n = \{X_1, \cdots, X_n\}$, where $X_i \in \mathcal{X}$, $i \in [n] = \{1, 2, \cdots, n\}$. For $S \subseteq [n]$, X_S is the subset of features $\{X_i : i \in S\}$, and $X_{S^c} = X^n \setminus X_S$. For the classification task, let $Y \in \mathcal{Y}$ and $\hat{Y} \in \mathcal{Y}$ be the true label and the predicted label of an individual, respectively.

To demonstrate the desired properties for our measures, we use a causal graph to represent the relations among the system variables [35]. A causal graph is a directed acyclic graph whose nodes are random variables. A directed edge from X to Y indicates that X is a direct cause of Y; we call X a *parent* of Y, and Y a *child* of X. If there is a directed path from X to Y, Y is called a *descendant* of X, and X an *ancestor* of Y. For a joint probability distribution P on the variables in a causal graph G, we assume that P satisfies both Markov and faithfulness conditions with respect to G [35]. Thus, conditional independencies among the variables can be read from the graph [35]. We require faithfulness since the desired properties for our measures are defined on the corresponding causal graph.

For simplicity of demonstrating our measures, we assume that the protected attributes A influence the true label Y only through the features X^n. That is, A does not have a direct causal influence on Y. As in [31], we also assume ancestral closure of A, i.e., any parent of A should be in A. For example, if race is a protected attribute, and mother's race is a parent of race, then mother's race should be a protected attribute as well. The following graphical model governs the statistical relations among the variables:

$$\begin{array}{c} A \to X^n \to Y \\ \downarrow \\ \hat{Y} \end{array} \qquad (2.4)$$

where $\hat{Y} = f(X^n)$. $f : \mathcal{X}^n \mapsto \mathcal{Y}$ represents the classifier, where $\mathcal{X}^n$ is the nth Cartesian product of $\mathcal{X}$.

2.3 Quantifying accuracy effect

Suppose $X = \{X_1, X_2\}$. To measure the impact of $\{X_1\}$ in accurately predicting Y, one candidate measure is $I(X_1; Y)$. However, this quantity contains the shared information between X_1 and X_2, with respect to Y, i.e.,

$SI(Y; X_1, X_2)$, which should not be attributed to X_1 alone. For example, in the extreme case where X_1 and X_2 are copies of one another, i.e., $X_1 = X_2$, it follows that X_1 is not essential for prediction. However, $I(X_1; Y)$ is equal to the total content of information. This suggests omitting the shared information from $I(X_1; Y)$. A second candidate measure is the unique information of X_1, i.e., $UI(Y; X_1 \setminus X_2)$. This measure is not adequate either. For example, let $X_1, X_2 \in \{0, 1\}$ be such that $P(X_1 = 1) = P(X_2 = 1) = 1/2 + \epsilon$, and let $Y = X_1 \oplus X_2$. As $\epsilon \to 0$, $I(X_1; Y) \to 0$, which implies that $UI(Y; X_1 \setminus X_2) \to 0$, while X_1 clearly impacts the accuracy of prediction.

We postulate that a good accuracy measure for a subset of features $X_S \subseteq X^n$, denoted by $v^{Acc}(X_S)$, should satisfy the following properties:

- Non-negativity: $v^{Acc}(X_S) \geq 0, \ \forall X_S \subseteq X^n$.

- Monotinicity: $v^{Acc}(X_{S_1}) \leq v^{Acc}(X_{S_2}), \ \forall S_1 \subseteq S_2$.

- Blocking: $Y \perp\!\!\!\perp X_S | \{A, X_{S^c}\} \iff v^{Acc}(X_S) = 0$.

Including an additional feature should not decrease prediction accuracy, hence $v^{Acc}(X_S)$ should satisfy monotinicity. By assigning a zero measure to the empty set ($v^{Acc}(\emptyset) = 0$), $v^{Acc}(X_S)$ is non-negative. The blocking property implies that, in the causal graph representing the relations among the variables, $v^{Acc}(X_S)$ is non-zero for features that form a *Markov blanket* of Y, and is zero for the remaining features. The two measures mentioned earlier (mutual and unique information) do not satisfy these properties. Mutual information does not satisfy blocking, and unique information is not monotone in X_S [4].

We propose to use the sum of unique information (that is only available in X_S) and synergistic information of X_S and all other variables, as a measure of the accuracy impact of X_S. This is equivalent to the conditional mutual information $I(Y; X_S | X_{S^c})$, which is the content of information that would be lost if we eliminate $\{X_S\}$ from the set X^n.

Definition 1 (Accuracy coefficient). For a subset of features $X_S \subseteq X^n$, the accuracy coefficient of X_S is given by

$$v^{Acc}(X_S) = I(Y; X_S | \{A, X_{S^c}\}) = UI(Y; X_S \setminus \{A, X_{S^c}\}) + CI(Y; X_S, \{A, X_{S^c}\}). \tag{2.5}$$

$v^{Acc}(X_S)$ satisfies all the desired properties. Non-negativity and monotinicity are straightforward. Theorem 1 below states that $v^{Acc}(X_S)$ satisfies the blocking property.

Theorem 1. For any subset of features $X_S \subseteq X^n$, the accuracy coefficient $v^{Acc}(X_S)$ is zero if and only if X_S is not a direct cause of Y in the corresponding causal graph, i.e., $Y \perp\!\!\!\perp X_S | X_{S^c}$.

Proof. The proof of Theorem 1 appears in Appendix A. $\qquad\square$

Remark 1. In Definition 1, we included the protected attribute A along with the other features X_{S^c} in the conditioning (and in the unique and synergistic information). This follows due to the assumption that the protected attribute A is not a direct cause of Y, and hence $A \not\subset X^n$. If A were to be a subset of X^n, then $\{A, X_{S^c}\}$ should be replaced by X_{S^c} in (2.5).

2.4 Quantifying discriminatory effect

As we mentioned earlier, discrimination can be explicit when the protected attribute A is used as an input to the classifier, or implicit when the protected attribute influences the outcome through proxy features. Our goal in this section is to measure the discriminatory impact of a subset of features $X_S \subseteq X^n$. A measure for the discriminatory impact of X_S, denoted by $v^D(X_S)$, should satisfy the following properties:

- Non-negativity: $v^D(X_S) \geq 0$

- Monotinicity: $v^D(X_{S_1}) \leq v^D(X_{S_2})$ for $S_1 \subseteq S_2$

- Y-independence: $Y \perp\!\!\!\perp X_S \implies v^D(X_S) = 0$.

- A-independence: $A \perp\!\!\!\perp X_S \implies v^D(X_S) = 0$.

- AY-independence: $A \perp\!\!\!\perp X_S | Y \implies v^D(X_S) = 0$.

Non-negativity and monotonicity have the same reasoning as for the accuracy measure in Section 2.3. We justify Y-independence and A-independence as follows. A subset X_S that is "irrelevant" to the classification task ($Y \perp\!\!\!\perp X_S$), or does not act as a proxy for the protected attributes ($A \perp\!\!\!\perp X_S$), should not be considered as discriminatory, i.e., $v^D(X_S) = 0$. Similarly, AY-independence implies that, conditioned on the class label $\{Y = y\}$, if X_S is independent of A, $v^D(X_S)$ should be zero. Note that neither $I(X_S; A)$ nor $I(X_S; A|Y)$ satisfies our desired properties. $I(X_S; A)$ does not satisfy Y-independence and AY-independence, and $I(X_S; A|Y)$ does not satisfy Y-independence and A-independence. Instead, we propose the following discrimination measure:

Definition 2 (Discrimination coefficient). For a subset of features $X_S \subseteq X^n$, the discrimination coefficient is

$$v^D(X_S) \triangleq SI(Y; X_S, A) \times I(X_S; A) \times I(X_S; A|Y). \tag{2.6}$$

$SI(Y; X_S, A)$ measures the information content in Y provided by X_S, and is "discriminatory" in the sense that it is shared with the protected attribute A. $I(X_S; A)$ measures the dependence between X_S and A. $I(X_S; A|Y)$ measures the conditional dependence between X_S and A, given Y. Our discrimination measure satisfies all the aforementioned properties:

Proposition 1. The discrimination coefficient $v^D(X_S)$ is monotone and non-negative. That is, (i) For any $S_1 \subseteq S_2$, we have $v^D(X_{S_1}) \leq v^D(X_{S_2})$, and (ii) $v^D(X_S) \geq 0$, for all $X_S \subseteq X^n$.

Proof. The proof of Proposition 1 is given in Appendix B. $\qquad \square$

Proposition 2. If a subset of features $X_S \subseteq X^n$ satisfies $X_S \perp\!\!\!\perp Y$, $X_S \perp\!\!\!\perp A$, or $X_S \perp\!\!\!\perp A|Y$, then $v^D(X_S) = 0$.

Proof. The proof of Proposition 2 appears in Appendix C. $\qquad \square$

Besides satisfying the desired properties, our discrimination measure $v^D(X_S)$ is in line with the notions of fairness in the literature [12, 44, 20].

Definition 3 (Notions of Fairness). A classifier $\hat{Y} : \mathcal{X}^n \to \mathcal{Y}$ satisfies demographic parity if it is statistically independent of the protected attribute ($\hat{Y} \perp\!\!\!\perp A$). Further, $\hat{Y}$ satisfies equalized odds if it is statistically independent of the protected attribute conditioned on the true label ($\hat{Y} \perp\!\!\!\perp A|Y$).

To restrict our measures in (2.5) and (2.6) to a certain classifier $f(X^n)$, we can replace X_S by $\hat{Y} = f|_{X_S}(X^n)$, where $f|_{X_S}(X^n) = f(X^n)$ such that X_{S^c} is held constant. For the discrimination measure $v^D(X_S)$, by replacing X_S with $\hat{Y}$, A-independence and AY-independence in Proposition 2 correspond to demographic parity and equalized odds.

We further investigate the properties of $v^D(X_S)$. In the following definition, a direct path refers to a chain of variables connected with directed arrows in the graphical model.

Definition 4. A subset of features $X_S \subseteq X^n$ is called path-discriminatory if it blocks all the direct paths from the protected attribute A to the true label Y, i.e., $A \perp\!\!\!\perp Y|X_S$.

Theorem 2. For every subset of features X_S that is path-discriminatory, $v^D(X_S) = I(Y; A)I(X_S; A)I(X_S; A|Y)$. Further, for a path-discriminatory subset of features K that is directly connected to the protected attribute A, i.e., $A \perp\!\!\!\perp \{Y \cup \{X^n \setminus K\}\}|K$, we have $v^D(K) \geq v^D(X_S)$, for all $X_S \subseteq X^n$.

Proof. The proof of Theorem 2 is given in Appendix D. $\qquad \square$

3 Aggregation of Effects

In Section 2, we defined accuracy and discrimination measures for subsets of features, $v^{Acc}(X_S)$ and $v^D(X_S)$. However, we shall not use $v^{Acc}(X_i)$ and $v^D(X_i)$ to measure the *marginal* accuracy and discrimination impacts of a single feature X_i; these do not take into account the correlation among features. For instance, from the definition of shared information, it is evident that $SI(Y; X_i, A)$ is a function of only the marginals P_{YX_i} and P_{YA}. In order to account for this correlation, we need to factor in the aggregate effect of all subsets of features that include a certain feature. For example, to measure the contribution of a single State's electoral votes on winning the election, we need to first measure the impact of that state when included with all possible combinations of other states, and aggregate these to deduce its marginal impact. We can calculate the effect of adding X_i to an arbitrary subset X_S by $v(X_S \cup \{X_i\}) - v(X_S)$. An appropriately weighted sum of these effects can provide an aggregation measure. This leads us to the *Shapley value* function [37]:

Definition 5. Let $\mathcal{P}$ denote the power set. Given a *characteristic function* $v(\cdot) : \mathcal{P}([n]) \to \mathbb{R}$, the Shapley value function $\phi_{(\cdot)} : [n] \to \mathbb{R}$ is defined as:

$$\phi_i = \sum_{T \subseteq [n] \setminus i} \frac{|T|!(n - |T| - 1)!}{n!} (v(T \cup \{i\}) - v(T)), \; \forall i \in [n].$$

Given the characteristic functions $v^{Acc}(\cdot)$ and $v^D(\cdot)$, the corresponding Shapley value functions are denoted by $\phi_{(\cdot)}^{Acc}$ and $\phi_{(\cdot)}^D$. We refer to these as *marginal accuracy coefficient* and *marginal discrimination coefficient*, respectively.

The weights $\frac{|T|!(n-|T|-1)!}{n!}$ in the definition of the Shapley value function are chosen so that the following lemma holds:

Lemma 3. [41] Shapley value is the unique aggregation function satisfying the following properties:

- Symmetry: If $v(T \cup \{i\}) = v(T \cup \{j\})$, for all $T \subseteq [n] \setminus \{i, j\}$, $\implies \phi_i = \phi_j$.

- Efficiency: $\sum_{i \in [n]} \phi_i = v([n])$.

- Monotonicity: Given two characteristic functions $v^{(1)}(\cdot)$, $v^{(2)}(\cdot)$, and the corresponding Shapley value functions $\phi_{(\cdot)}^{(1)}, \phi_{(\cdot)}^{(2)}$, if $v^{(1)}(T \cup \{i\}) - v^{(1)}(T) \geq v^{(2)}(T \cup \{i\}) - v^{(2)}(T), \forall T \subseteq [n] \implies \phi_i^{(1)} \geq \phi_i^{(2)}$.

Besides these properties, we have the following result:

Corollary 1. Marginal accuracy and discrimination coefficients are non-negative, i.e., $\phi_i^{Acc}, \phi_i^D \geq 0$, for all $i \in [n]$.

Proof. The proof of Corollary 1 appears in Appendix E. $\qquad\square$

The following corollary captures the special case when one of the features, say X_a, where $a \in [n]$, is a sufficient statistics for the true label Y. In this case, X_a should attain the highest marginal accuracy coefficient among all features.

Corollary 2. In the special case where Y has a single parent X_a, i.e., $Y \perp \{A \cup \{X^n \setminus X_a\}\}|X_a$, we have $\phi_a^{Acc} \geq \phi_j^{Acc}$, for all $j \in [n] \setminus a$.

Proof. The proof for Corollary 2 appears in Appendix F. $\qquad\square$

Corollary 3 below states, if a certain feature is a sufficient statistics for the protected attribute, its marginal discrimination coefficient should be the highest among all features.

Corollary 3. In the special case where A has a single child X_d, and $A \perp (Y \cup \{X^n \setminus X_d\})|X_d$; $\phi_d^D \geq \phi_j^D, \forall j \in [n] \setminus d$.

Proof. The proof of Corollary 3 is provided in Appendix G. $\qquad\square$

A feature that is a single child of A is the "worst" feature in terms of discrimination. This feature is the most correlated with A, and all the information that pertains to A passes through this feature to the other nodes of the graph. In other words, removing this feature results in A and Y being independent.

We can use the marginal coefficients ϕ_i^{Acc} and ϕ_i^D to define a score for each feature (for feature selection), termed as the fairness-utility score. Specifically, given ϕ_i^{Acc} and ϕ_i^D, the fairness-utility score for a feature X_i is defined as $\mathcal{F}_i = \phi_i^{Acc} - \alpha \phi_i^D$, where α is a positive hyperparametter which trades off between accuracy and discrimination.

4 Experimental Results

We evaluate our accuracy and discrimination measures both on real and synthetic datasets.

Features	$10^2 \times \phi^{Acc}$	$10^7 \times \phi^D$
1) X_1	0.027	2.7927
2) X_2	0.052	0.0923
3) X_3	4.467	0.0959
4) X_4	3.122	0.0899
5) X_5	0.141	1.4301

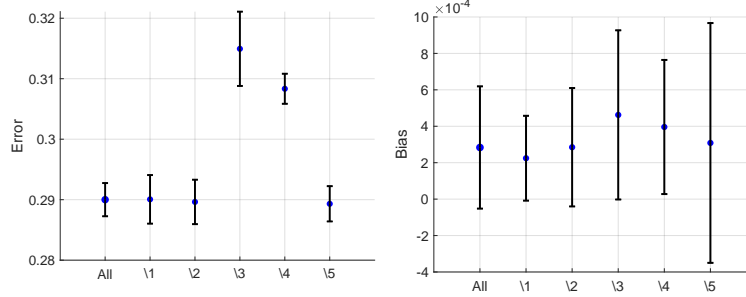Table 1: Discrimination and accuracy coefficients For synthetic dataset.

Figure 2: a) Error and b) Bias of the classifier over synthetic dataset when all the features are used, and when one of the features is removed.

4.1 Synthetic Dataset

We evaluate our measures on a randomly generated dataset, following the graphical model in Figure 3. We randomly assigned a conditional distribution for each variable given their parents in the graph. We calculated the joint distribution P_{XAY} from these conditional probabilities. We calculated the marginal accuracy and discrimination coefficients for this joint distribution. The results are shown in Table 1.

Furthermore, we randomly split the whole dataset into training and test subsets. We trained a neural network classifier with the features as the input, and prediction of Y as the output. We removed one feature at a time and repeated the same procedure with the remaining features. Similar to the real dataset, the prediction error is calculated using cross entropy loss, and the bias is calculated by the KL divergence between $P_{\hat{Y}|A=0}$ and $P_{\hat{Y}|A=1}$. We ran this procedure 100 times, and calculated the average and the confidence intervals. The results are shown in Figure 2. It is clear that removal of X_3 and X_4, which are the most important features according to the marginal accuracy coefficient, results in a higher increase in the prediction error. In addition, as suggested by the marginal discrimination coefficient, X_1 is the most discriminatory feature. We observe that removing X_1 results in the lowest bias in the prediction.
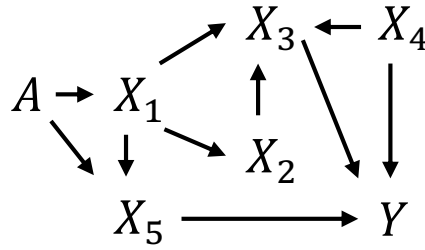

Figure 3: The graphical model of the data generating process of the synthetic data.

Features	$10^2 \times \phi^{Acc}$	$10^6 \times \phi^{D}$
1) `Age`	1.7892	6.0904
2) `Charge Degree`	0.1723	1.3916
3) `Gender`	0.4186	1.6849
4) `Prior Counts`	3.6116	5.0466
5) `Length of Stay`	1.8290	2.3126

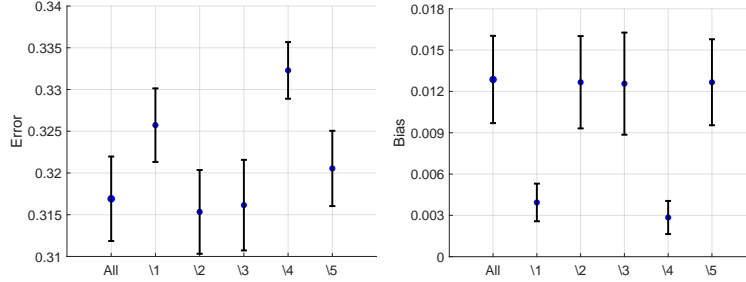Table 2: Discrimination and accuracy coefficients for COMPAS dataset.

Figure 4: a) Error and b) Bias of the classifier over COMPAS dataset when all the features are used, and when one of the features is removed. $\setminus i$ denotes the case where i'th feature (as numerated in Table 2) is removed.

4.2 ProPublica COMPAS Dataset

ProPublica COMPAS dataset [2] contains the criminal history and demographic makeup of defendants in Broward County, Florida from 2013-2014. We processed the raw dataset by dropping records with missing information and converted categorical variables to numerical values. The race of each individual is provided in the COMPAS dataset and constitutes our protected attribute. We restrict our analysis to individuals who are African American ($A = 0$) or Caucasian ($A = 1$). Our processed dataset contains 5334 records (3247 African Americans and 2087 Caucasians). Each individual in our dataset has a feature vector (`Age, Charge Degree, Gender, Prior Counts, Length Of Stay`), and a binary true label which indicates whether the individual was arrested for a crime within 2 years of release. `Age` variable takes three levels of `Age` < 25, $25 <$ `Age` < 45, or `Age` > 45. `Charge Degree` has two values Misdemeanor or Felony, `Gender` is either Male or Female, `Prior Counts` can be 0, $1 - 3$, or larger than 3, and `Length of Stay` can be ≤ 1 week, ≤ 3 months, or > 3 months.

We applied our measures of accuracy and discrimination to this dataset (see Table 2). The result shows that `Prior Count` and `Age` exhibit strongest proxies for discrimination. This is in line with the findings of [40]. In addition, the result shows that `Charge Degree` and `Gender` are the least informative features for the prediction task.

Furthermore, same as the synthetic dataset, we designed a classifier with all the features included, and with one of the features removed. The results are shown in Figure 4. As it is predicted by the marginal discrimination coefficient, removal of `Age` or `Prior Counts` results in the lowest bias in the classifier output. In addition, as it is anticipated by the accuracy coefficient, removal of `Charge Degree` and `Gender` has very small effect on the error of the predictor.

References

[1] Wael Alghamdi, Shahab Asoodeh, Hao Wang, Flavio P Calmon, Dennis Wei, and Karthikeyan Natesan Ramamurthy. Model projection: Theory and applications to fair machine learning. In *2020 IEEE International Symposium on Information Theory (ISIT)*, pages 2711–2716, 2020.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *Pro Publica*, 2016.

[3] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

[4] Nils Bertschinger, Johannes Rauh, Eckehard Olbrich, Jürgen Jost, and Nihat Ay. Quantifying unique information. *Entropy*, 16(4):2161–2183, 2014.

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

[6] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pages 3992–4001, 2017.

[7] Shay Cohen, Eytan Ruppin, and Gideon Dror. Feature selection based on the shapley value. In *19th international joint conference on Artificial intelligence*, pages 665–670. Morgan Kaufmann Publishers, 2005.

[8] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

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

[10] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[11] Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. An information-theoretic quantification of discrimination with exempt features. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3825–3833, 2020.

[12] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[13] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

[14] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

[15] Sainyam Galhotra, Karthikeyan Shanmugam, Prasanna Sattigeri, and Kush R Varshney. Fair data integration. *arXiv preprint arXiv:2006.06053*, 2020.

[16] AmirEmad Ghassami, Sajad Khodadadian, and Negar Kiyavash. Fairness in supervised learning: An information theoretic approach. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 176–180. IEEE, 2018.

[17] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[18] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016.

[19] Sara Hajian and Josep Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering*, 2013.

[20] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[21] D Bradford Hunt. Redlining. *Encyclopedia of Chicago*, 2005.

[22] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

[23] Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 1(2), 2016.

[24] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, pages 325–333, 2016.

[25] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[26] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650. IEEE, 2011.

[27] Ehsan Kazemi, Morteza Zadimoghaddam, and Amin Karbasi. Scalable deletion-robust submodular maximization: Data summarization with privacy and fairness constraints. In *International Conference on Machine Learning*, pages 2549–2558, 2018.

[28] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[29] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Estimating skin tone and effects on classification performance in dermatology datasets. *arXiv preprint arXiv:1910.13268*, 2019.

[30] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[31] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

[32] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

[33] John F Mahoney and James M Mohen. Method and system for loan origination and underwriting, October 23 2007. US Patent 7,287,008.

[34] Masayoshi Mase, Art B Owen, and Benjamin B Seiler. Cohort shapley value for algorithmic fairness. *arXiv preprint arXiv:2105.07168*, 2021.

[35] Judea Pearl. *Causality*. Cambridge university press, 2009.

[36] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the SIAM International Conference on Data Mining*. SIAM, 2009.

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

[38] Xin Sun, Yanheng Liu, Jin Li, Jianqi Zhu, Xuejie Liu, and Huiling Chen. Using cooperative game theory to optimize the feature selection problem. *Neurocomputing*, 97:86–93, 2012.

[39] Hao Wang, Berk Ustun, and Flavio Calmon. Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pages 6618–6627. PMLR, 2019.

[40] Hao Wang, Berk Ustun, and Flavio P Calmon. On the direction of discrimination: An information-theoretic analysis of disparate impact in machine learning. *arXiv preprint arXiv:1801.05398*, 2018.

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

[42] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017.

[43] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.

[44] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 325–333, 2013.

Supplementary Material

A   Proof of Theorem 1

If part: From (2.4), Y is a descendant of all the features in the graphical model. Assuming $X_S \subseteq X^n$ is not a direct parent of Y in the graphical model, given X_{S^c}, Y and X_S are independent, i.e., $P_{Y|A,X_{S^c},X_S} = P_{Y|A,X_{S^c}}$. We have

$$
\begin{aligned}
I(Y;A,X^n) &= I(Y;A,X_{S^c},X_S) \\
&= \sum_{y,a,x^n} P_{Y,A,X_{S^c},X_S}(y,a,x_{S^c},x_S) \; \log \frac{P_{Y|A,X_{S^c},X_S}(y|a,x_{S^c},x_S)}{P_Y(y)} \\
&= \sum_{y,a,x^n} P_{Y,A,X_{S^c},X_S}(y,a,x_{S^c},x_S) \; \log \frac{P_{Y|A,X_{S^c}}(y|a,x_{S^c})}{P_Y(y)} \\
&= \sum_{y,a,x_{S^c}} P_{Y,A,X_{S^c}}(y,a,x_{S^c}) \log \frac{P_{Y|A,X_{S^c}}(y|a,x_{S^c})}{P_Y(y)} \\
&= I(Y;A,X_{S^c}) \\
&\implies v^{Acc}(X_S) = I(Y;A,X_S|X_{S^c}) = 0.
\end{aligned}
$$

Only if part: Now assume $v^{Acc}(X_S) = 0$. We want to prove that X_S is not a direct parent of Y.

$$
\begin{aligned}
v^{Acc}&(X_S) = 0 \\
&\implies I(Y;X^n,A) = I(Y;X_{S^c},A) \\
&\implies \sum_{y,x^n,a} P_{Y,X^n,A}(y,x^n,a) \log \frac{P_{Y|X^n,A}(y|x^n,a)}{P_Y(y)} \\
&= \sum_{y,x_{S^c},a} P_{Y,X_{S^c},A}(y,x_{S^c},a) \log \frac{P_{Y|X_S^c,A}(y|x_{S^c},a)}{P_Y(y)} \\
&= \sum_{y,x^n,a} P_{Y,X^n,A}(y,x^n,a) \log \frac{P_{Y|X_{S^c},A}(y|x_{S^c},a)}{P_Y(y)} \\
&\implies \sum_{y,x^n,a} P_{Y,X^n,A}(y,x^n,a) \log \frac{P_{Y|X^n,A}(y|x^n,a)}{P_{Y|X_{S^c},A}(y|x_{S^c},a)} = 0 \\
&\implies \sum_{a,x^n \in \mathcal{X}^n} P_{A,X^n}(a,x^n) D_{KL}(P_{Y|A=a,X^n=x^n} || P_{Y|A=a,X_{S^c}=x_{S^c}}) = 0
\end{aligned}
$$

Thus, $D_{KL}(P_{Y|A=a,X^n=x^n} || P_{Y|A=a,X_{S^c}=x_{S^c}}) = 0$, for all a,x^n, since $D_{KL}(.||.) \geq 0$. Furthermore, according to the Gibbs' inequality, $D_{KL}(P||Q) = 0$ if and only if $P = Q$ almost every where. As a result, $P_{Y|A=a,X^n=x^n} = P_{Y|A=a,X_{S^c}=x_{S^c}}$, for all a, and $x^n \in \mathcal{X}^n \implies Y \perp X_S | \{A, X_{S^c}\}$. Since we assume that the joint distribution is faithful with respect to the graphical model, we conclude that X_S is not a direct parent of Y.

B   Proof of Proposition 1

According to [4], unique information is monotonic with respect to its third argument, i.e.,

$$
\begin{aligned}
S_1 \subseteq S_2 &\implies X_{S_1} \subseteq X_{S_2} \\
&\implies UI(Y;A \setminus X_{S_1}) \geq UI(Y;A \setminus X_{S_2}).
\end{aligned}
$$

In addition, we have:

$$
SI(Y;A,X_S) = I(Y;A) - UI(Y;A \setminus X_S)
$$

As a result:
$$S_1 \subseteq S_2 \implies SI(Y; A, X_{S_1}) \leq SI(Y; A, X_{S_2}).$$

The proof of non-negativity of shared information can be found in [4]. Further, both mutual information $I(A; X_S)$, and conditional mutual information $I(A; X_S|Y)$ are non-negative and monotonic with respect to X_S [8]. Since all three terms $SI(Y; X_S, A)$, $I(A; X_S)$, and $I(A; X_S|Y)$ are non-negative and monotonic in X_S, their multiplication is also non-negative and monotonic in X_S.

C   Proof of Proposition 2

$X_S \perp\!\!\!\perp A$ and $X_S \perp\!\!\!\perp A|Y$ imply that $I(X_S; A) = 0$ and $I(X_S; A|Y) = 0$, respectively, see [8]. Furthermore, as shown in [4, Lemma 12]:
$$SI(Y; X_S, A) \leq I(Y; X_S).$$

As a result,

$$Y \perp\!\!\!\perp X_S \implies SI(Y; X_S, A) = 0$$
$$\implies v^D(X_S) = 0.$$

D   Proof of Theorem 2

According to [4, Lemma 13], for a path-discriminatory subset of features X_S, we have $SI(Y; X_S, A) = I(Y; A)$.

According to [4, Lemma 12], for any $X_S \subseteq X^n$, $SI(Y; X_S, A) \leq I(Y; A)$. Thus, the discrimination coefficient of a path-discriminatory subset of features is an upper bound of the discrimination coefficient of all other subsets of features, and hence Theorem 2 follows.

E   Proof of Corollary 1

The accuracy and discrimination coefficients $v^{Acc}(.)$ and $v^D(.)$ both are non-negative and monotonic, see Proposition 1. The marginal accuracy and discrimination coefficients ϕ_i^{Acc} and ϕ_i^D, $i \in [n]$, are the summation of terms on the form $v^{Acc/D}(X_S \cup \{X_i\}) - v^{Acc/D}(X_S)$, which are non-negative due to the monotonicity and non-negativity of v^{Acc} and v^D.

F   Proof of Corollary 2

For the marginal accuracy coefficient ϕ_i^{Acc}, $i \in [n]$, we have

$$\phi_i^{Acc} = \sum_{X_S \subseteq X_{-i}^n} w(X_S) \left[v^{Acc}(X_S \cup \{X_i\}) - v^{Acc}(X_S) \right],$$

where $w(X_S) = \frac{|X_S|!(n-|X_S|-1)!}{n!}$ and $X_{-i}^n = X^n \setminus \{X_i\}$. Recall, $v^{Acc}(\cdot)$ is defined in (2.5). We thus have

$$\phi_i^{Acc} = \sum_{X_S \subseteq X_{-i}^n} w(X_S) \Big[I\left(Y; X_S \cup \{X_i\} \big| A, (X_{S^c} \setminus X_i)\right) - (I(Y; X_S|A, X_{S^c})\Big]$$

$$= \sum_{X_S \subseteq X_{-i}^n} w(X_S) \Big[I(Y; A, X^n) - I(Y; A, X_{S^c} \setminus \{X_i\}) - I(Y; A, X^n) + I(Y; A, X_{S^c})\Big]$$

$$= \sum_{X_S \subseteq X_{-i}^n} w(X_S) \Big[I(Y; A, X_{S^c}) - I(Y; A, X_{S^c} \setminus \{X_i\})\Big]$$

$$= \sum_{X_S \subseteq X_{-i}^n} w(X_S) \Big[I(Y; A, X_S \cup \{X_i\}) - I(Y; A, X_S)\Big].$$

14

The last equality follows by applying a change of variables, where we replace $(X_{S^c} \setminus \{X_i\})$ by X_S, and thus X_{S^c} is replaced by $X_S \cup \{X_i\}$.

Suppose $i \in [n] \setminus a$. For $X_S \subseteq X^n \setminus \{X_i, X_a\}$, we have:

$$I(Y; A, X_S \cup \{X_a\}) = I(Y; X_a) + I(Y; A, X_S | X_a)$$
$$= I(Y; X_a),$$

where the last equality follows from the fact that X_a is the only parent of Y, and hence $I(Y; A, X_S | X_a) = 0$. Therefore, for any X_S such that $X_a \in X_S$, we have

$$I(Y; A, X_S \cup \{X_i\}) = I(Y; A, X_S) = I(Y; X_a)$$
$$\implies I(Y; A, X_S \cup \{X_i\}) - I(Y; A, X_S) = 0.$$

As a result, $\phi_a^{Acc} \geq \phi_i^{Acc}, \forall i \in [n] \setminus a$.

G   Proof of Corollary 3

Suppose that X_d satisfies the condition of Corollary 3: X_d is a single child of A, and $A \perp\!\!\!\perp (Y \cup \{X^n \setminus X_d\}) | X_d$. Also, suppose that $i \in [n] \setminus d$.

For $X_S \subseteq X^n \setminus \{X_i\}$, if $X_d \in X_S$, we have $v^D(X_S) = v^D(X_S \cup \{X_i\}) = I(Y; A)I(A; X_d)I(A; X_d | Y) \implies v^D(S \cup \{X_i\}) - v^D(X_S) = 0$. In addition, if $X_d \notin X_S$, due to the Theorem 2 and data processing inequality [8], we have $v^D(X_S \cup \{X_i\}) - v^D(X_S) \leq v^D(X_S \cup \{X_d\}) - v^D(X_S)$.

As in Appendix F, let $w(X_S) = \frac{|X_S|!(n-|X_S|-1)!}{n!}$. For all $i \in [n] \setminus d$, we have

$$
\begin{aligned}
\phi_i^D &= \sum_{X_S \subseteq X^n \setminus \{X_i\}} w(X_S) \left(v^D(X_S \cup \{X_i\}) - v^D(X_S) \right) \\
&\leq \sum_{X_S \subseteq X^n \setminus \{X_d\}} w(X_S) \left(v^D(X_S \cup \{X_k\}) - v^D(X_S) \right) \\
&= \phi_d^D.
\end{aligned}
$$