Quantifying Feature Contributions to Overall Disparity Using Information Theory

Sanghamitra Dutta, Praveen Venkatesh, Pulkit Grover

Abstract

When a machine-learning algorithm makes biased decisions, it can be helpful to understand the "sources" of disparity to explain why the bias exists. Towards this, we examine the problem of quantifying the contribution of each individual feature to the observed disparity. If we have access to the decision-making model, one potential approach (inspired from intervention-based approaches in explainability literature) is to vary each individual feature (while keeping the others fixed), and use the resulting change in disparity to quantify its contribution. However, we may not have access to the model or be able to test/audit its outputs for individually varying features. Furthermore, the decision may not always be a deterministic function of the input features (e.g., with human-in-the-loop). For these situations, we might need to explain contributions using purely distributional (i.e., observational) techniques, rather than interventional. We ask the question: *what is the "potential" contribution of each individual feature to the observed disparity in the decisions when the exact decision-making mechanism is not accessible?* We first provide canonical examples (thought experiments) that help illustrate the difference between distributional and interventional approaches to explaining contributions, and when either is better suited. When unable to intervene on the inputs, we quantify the "redundant" statistical dependency about the protected attribute that is present in both the final decision and an individual feature, by leveraging a body of work in information theory called Partial Information Decomposition. We also perform a simple case study to show how this technique could be applied to quantify contributions.

1 INTRODUCTION

Machine learning algorithms have permeated almost every aspect of our lives, including high-stakes applications, e.g. hiring and admissions. With their growing use, it has become increasingly important to incorporate algorithmic fairness to avoid bias with respect to protected attributes, such as, gender, race, nationality, age, etc. Existing literature on fairness [1–20] provides several measures of bias and disparity at the final model output, as well as, several techniques to mitigate these biases during model design.

However, in many applications, e.g., college admissions, the decision-making mechanism is a complex combination of algorithms and human-in-the-loop. Thus, only identifying bias and disparity in the final decision may not be enough to audit and, subsequently, mitigate them. E.g., there is an ongoing debate in the US on whether GRE/TOEFL scores should be used for college admissions because they may cause disparity in the decisions with respect to protected attributes [21, 22]. It would help to understand *how* the disparity in the decisions arose, e.g., which features could be potentially responsible for the disparity, and then evaluate how critical those features are for the specific application. In fact, several existing discrimination laws (e.g. Title VII of the US Civil Rights Act [23, 24]) allow exemptions if the disparity can be justified by an occupational necessity, e.g., coding test for software engineers, or weightlifting ability for firefighters [2, 10–18, 25].

The problem of quantifying contribution of different features to the overall disparity bridges the fields of both fairness and explainability. Explanations often provide an understanding (e.g. through visualization) of the contribution of each individual feature to the final decision [26–28]. Here we are interested in quantifying the contribution of features to the observed disparity in the decision making.

One possible approach inspired from existing techniques in explainability (e.g., QII [26], Influence Functions [29], SHAP [27, 30]) could be to intervene on the input features and observe changes in the output. For instance, one can vary an input feature while keeping other features unchanged, and assign the resulting change in disparity in the final output to this feature to quantify its contribution to the disparity. However, these techniques are not a good fit when one does not have access to the model to be able to intervene on its inputs.

Furthermore, in many applications, e.g., admissions, the machine-learnt algorithm might be used in conjunction with human-in-the-loop. This introduces an additional challenge in quantifying contributions of individual features: the output is no longer a deterministic function of the input features. The final decision may not be the same for two candidates having the same values of input features as additional, non-quantified aspects are taken into consideration.

In this work, the question we ask is: what is the potential contribution of each feature to the observed disparity when the exact decision-making mechanism is not accessible? To answer this question, we introduce two perspectives to this problem: (i) an interventional perspective; and (ii) a purely distributional perspective (i.e., observational). We then demonstrate when one is better suited than the other. We call our first approach *interventional* because it requires one to intervene on specific input features and observe the final outputs, thereby changing their joint distribution. We call our second approach *distributional/observational* because it only depends on the joint distribution of the observed input features and the final output, and does not require access to the model to be able to intervene on specific input features.

Our work makes the following contributions:

1. Explaining Contribution of Individual Features to the Observed Disparity: We first quantify the observed disparity as the mutual information [31] $I(Z; \hat{Y})$ where Z is the protected attribute, e.g., gender, race, etc., and $\hat{Y}$ is the final decision. Then, we introduce an *interventional* approach (see Candidate Measure 1 in Section 3) to quantify the contribution of each individual feature to the observed disparity. This approach is inspired from existing works [26, 27] in explainability, and can be used when one has access to the model making the decisions, and is able to intervene on the inputs to observe the outputs (decisions). However, for situations when access to the model is either not available or the decisions are not deterministic given the input features, we also introduce a *distributional* approach (see Candidate Measure 2 in Section 3). We provide a measure to quantify the "potential" contribution of each individual feature to the decision. This technique leverages a body of work in information theory called Partial Information Decomposition (PID) that quantifies the "redundant" information about the protected attribute Z present both in $\hat{Y}$ and an input feature.

Remark 1. *We use the term "potential" because even if a feature has a non-zero potential contribution (by our definition), one is unable to check if intervening on this feature actually results in a change in disparity in the final decision. Being unable to intervene on the model inputs, we essentially capture the "redundant" statistical dependency (with the protected attribute) that is present in both an individual feature and the final decision. Thus, in spirit, this may seem similar to capturing statistical correlations with the input features and the output, as discussed in explainability literature [32]. However, note that, our quantification involves three random variables: we want to capture the statistical dependency about the protected attribute that is present in both the final decision and an input feature. Statistical correlation is only defined for two random variables. This leads us to examine "redundant" information that precisely quantifies this dependency.*

2. Canonical Examples to Illustrate the Difference Between Various Explainability Approaches: We also discuss several canonical examples in Section 3 that illustrate the differences between an *interventional* and a *distributional* approach to explaining the contribution of each individual feature. These examples also help us understand when one approach is better suited than the other.

3. Case Study on an artificial admissions dataset: We finally demonstrate a case study on an artificial graduate admissions dataset to demonstrate how these techniques could be used to explain contributions.

Related Works: Algorithmic fairness is an important field of research with several measures of fairness as well as approaches to incorporate them in model design. The most closely related works to our work are [2, 12–19, 33, 34]. However, [2, 12–19] focus on quantifying exempt and non-exempt discrimination (either using observational measures or causal modelling) given a choice of critical features, rather than quantifying the contribution of each individual feature. Alternatively, [33, 34] propose information-theoretic techniques to carefully select features for fair decision making. In this work, we focus on explaining the contributions of all the individual features to the overall disparity, even when we may not have access to the exact decision-making mechanism (including scenarios with human-in-the-loop). In [35], the focus is on leveraging the underlying causal graph to understand feature contributions to disparity. Another recent related work [36] focus on counterfactual explainability techniques for fairness.

This work is also connected with the literature on explainability [26, 27, 29, 30, 37–46] (see [32] for a survey). In particular, [27, 30] use Shapley values, but their goal is to quantify the contribution of individual features to the
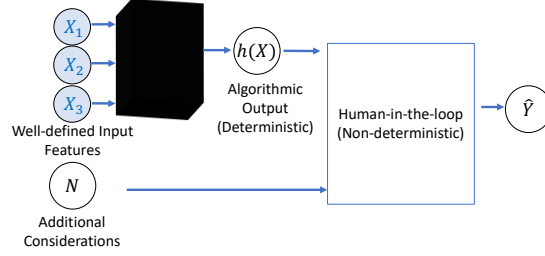
Figure 1: Illustration of our setup: Here $X = (X_1, X_2, \ldots, X_n)$ denotes a set of well-defined input features that are input to a (deterministic) model $h(X)$. The final decision $\hat{Y}$ combines $h(X)$ and additional considerations (which usually includes the protected attribute Z). Thus, $\hat{Y}$ is not necessarily a deterministic function of X.

decision (or, its accuracy), and propose several approximations for the ease of computation. Here, our goal is not to explain the contribution of individual features to the *decision* (or, its accuracy), but rather their contribution to the overall *disparity in the decision*. We note that [30] also extend their Shapley-value-based explainability technique for an application in fairness: quantifying contributions of neurons to the accuracy only on a protected group, e.g., people of a certain gender or race. In contrast, here we are interested in quantifying contribution of different features to the mutual information $\mathrm{I}(Z; \hat{Y})$ (statistical dependence between protected attribute and final output). In this context, our work brings out the contrast between interventional and distributional approaches to explainability through canonical examples, and illustrates when one is better suited than the other. In doing so, our work leverages tools from Partial Information Decomposition, that also can have broad applications in explainability.

2 PRELIMINARIES

2.1 Our Notations and System Model

We let $X = (X_1, X_2, \ldots, X_n)$ denote a set of well-defined input features that go into a model. The output of the model is denoted by $h(X)$ which is usually a deterministic function of the model inputs. We also let Z denote the protected attribute, e.g., gender, race, etc., which may or may not be an input feature explicitly fed into the model ($h(X)$). The final decision is denoted by $\hat{Y}$ which is a complex combination of the deterministic model output $h(X)$ and subjective evaluation by human-in-the-loop who may take additional factors (non-quantified aspects) into consideration. These additional factors almost always include the protected attribute Z, e.g., gender, race, age, etc. Therefore, the final decision $\hat{Y}$ may not be a deterministic function of the model inputs X, and could also depend on Z (see Fig. 1).

Next, we provide a brief background on Partial Information Decomposition (PID) in Section 2.2 and Shapley Values in Section 2.3 for completeness. Readers familiar with these concepts may directly proceed to our main results in Section 3, which is followed by a case study on an artificial admissions dataset in Section 4.

2.2 Background on Partial Information Decomposition (PID)

PID [47–49] is an emerging body of work in information theory that decomposes the mutual information $\mathrm{I}(Z; (A, B))$ about a random variable Z contained in the tuple (A, B) into four *non-negative* terms (also see Fig. 2):

$$\begin{aligned} \mathrm{I}(Z; (A, B)) = \mathrm{Uni}(Z : A|B) + \mathrm{Uni}(Z : B|A) \\ + \mathrm{Red}(Z : (A, B)) + \mathrm{Syn}(Z : (A, B)). \end{aligned} \tag{1}$$

Here, $\mathrm{Uni}(Z : A|B)$ denotes the unique information about Z that is present only in A and not in B. Similarly, $\mathrm{Uni}(Z : B|A)$ is the unique information about Z that is present only in B and not in A. The term $\mathrm{Red}(Z : (A, B))$ denotes the redundant information about Z that is present in both A and B, and $\mathrm{Syn}(Z : (A, B))$ denotes the synergistic information not present in either of A or B individually, but present jointly in (A, B). *All four of these terms are non-negative. Also notice that,* $\mathrm{Red}(Z : (A, B))$ *and* $\mathrm{Syn}(Z : (A, B))$ *are symmetric in A and B.* Before defining these PID terms formally, let us understand them through an intuitive example.
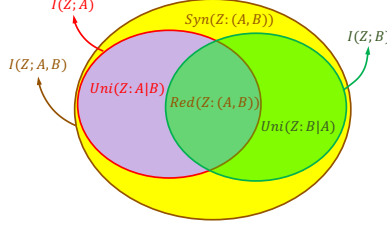
Figure 2: PID: Mutual information $\mathrm{I}(Z;(A,B))$ is decomposed into 4 non-negative terms, namely, $\mathrm{Uni}(Z:A|B)$, $\mathrm{Uni}(Z:B|A)$, $\mathrm{Red}(Z:(A,B))$ and $\mathrm{Syn}(Z:(A,B))$. Also note that, $\mathrm{I}(Z;(A,B)) = \mathrm{I}(Z;B) + \mathrm{I}(Z;A\mid B)$, each of which is in turn a sum of two PID terms. $\mathrm{Red}(Z:(A,B))$ is the sub-volume between $\mathrm{I}(Z;A)$ and $\mathrm{I}(Z;B)$, and $\mathrm{Uni}(Z:A|B)$ is the sub-volume between $\mathrm{I}(Z;A\mid B)$ and $\mathrm{I}(Z;A)$.

Example 1 (Understanding PID). *Let $Z = (Z_1, Z_2, Z_3)$ with $Z_1, Z_2, Z_3 \sim$ i.i.d. Bern(½). Let $A = (Z_1, Z_2, Z_3 \oplus N)$, $B = (Z_2, N)$, $N \sim$ Bern(½) is independent of Z. Here, $\mathrm{I}(Z;(A,B)) = 3$ bits.*

The unique information about Z that is contained only in A and not in B is effectively contained in Z_1 and is given by $\mathrm{Uni}(Z:A|B) = \mathrm{I}(Z;Z_1) = 1$ bit. The redundant information about Z that is contained in both A and B is effectively contained in Z_2 and is given by $\mathrm{Red}(Z:(A,B)) = \mathrm{I}(Z;Z_2) = 1$ bit. Lastly, the synergistic information about Z that is not contained in either A or B alone, but is contained in both of them together is effectively contained in the tuple $(Z_3 \oplus N, N)$, and is given by $\mathrm{Syn}(Z:(A,B)) = \mathrm{I}(Z;(Z_3 \oplus N, N)) = 1$ bit. This accounts for the 3 bits in $\mathrm{I}(Z;(A,B))$. Here, B does not have any unique information about Z that is not contained in A, i.e., $\mathrm{Uni}(Z:B|A) = 0$.

Irrespective of the formal definition of the individual PID terms, the following identities also hold:

$$\mathrm{I}(Z;A) = \mathrm{Uni}(Z:A|B) + \mathrm{Red}(Z:(A,B)). \tag{2}$$

$$\mathrm{I}(Z;A\mid B) = \mathrm{Uni}(Z:A|B) + \mathrm{Syn}(Z:(A,B)). \tag{3}$$

Note that, $\mathrm{Uni}(Z:A|B)$ can be viewed as the information-theoretic sub-volume of the intersection between $\mathrm{I}(Z;A)$ and $\mathrm{I}(Z;A\mid B)$. Similarly, $\mathrm{Red}(Z:(A,B))$ is the sub-volume between $\mathrm{I}(Z;A)$ and $\mathrm{I}(Z;B)$.

These equations also demonstrate that $\mathrm{Uni}(Z:A|B)$ and $\mathrm{Red}(Z:(A,B))$ are the information contents that exhibit themselves in $\mathrm{I}(Z;A)$ which is the statistically visible information content about Z present in A. Because both of these PID terms are non-negative, if any one of them is non-zero, we will have $\mathrm{I}(Z;A) > 0$. Similarly, $\mathrm{Uni}(Z:B|A)$ and $\mathrm{Red}(Z:(A,B))$ also exhibit themselves in $\mathrm{I}(Z;B)$. On the other hand, $\mathrm{Syn}(Z:(A,B))$ is the information content that does not exhibit itself in $\mathrm{I}(Z;A)$ or $\mathrm{I}(Z;B)$ individually, i.e., these terms can still be 0 even if $\mathrm{Syn}(Z:(A,B)) > 0$. But, $\mathrm{Syn}(Z:(A,B))$ exhibits itself in $\mathrm{I}(Z;(A,B))$. Also note that,

$$\begin{aligned}
\mathrm{I}(Z;(A,B)) &= \underbrace{\mathrm{Uni}(Z:A|B) + \mathrm{Red}(Z:(A,B))}_{\mathrm{I}(Z;A)} \\
&\quad + \underbrace{\mathrm{Uni}(Z:B|A) + \mathrm{Syn}(Z:(A,B))}_{\mathrm{I}(Z;B|A)} \\
&= \underbrace{\mathrm{Uni}(Z:B|A) + \mathrm{Red}(Z:(A,B))}_{\mathrm{I}(Z;B)} \\
&\quad + \underbrace{\mathrm{Uni}(Z:A|B) + \mathrm{Syn}(Z:(A,B))}_{\mathrm{I}(Z;A|B)}.
\end{aligned} \tag{4} \tag{5}$$

Given three independent equations (1), (2) and (3) in four unknowns (the four PID terms), defining any one of the terms (e.g., $\mathrm{Uni}(Z:A|B)$) is sufficient to obtain the other three. For completeness, we include the definition of unique information from [47] (that also allows for estimation via convex optimization [50]) with the specific properties used in the proofs in the Appendix. To follow the paper, only an intuitive understanding (from Fig. 2) may be sufficient.

Definition 1 (Unique Information [47]). *Let Δ be the set of all joint distributions on (Z,A,B) and Δ_p be the set of joint distributions with the same marginals on (Z,A) and (Z,B) as their true distribution, i.e.,*

$$\Delta_p = \{Q \in \Delta : q(z,a) = \mathrm{Pr}(Z{=}z, A{=}a) \text{ and } q(z,b) = \mathrm{Pr}(Z{=}z, B{=}b)\}.$$

$$\text{Uni}(Z:A|B) = \min_{Q \in \Delta_p} \text{I}_Q(Z;A \mid B), \tag{6}$$

where $\text{I}_Q(Z;A \mid B)$ is the conditional mutual information when (Z, A, B) have joint distribution Q.

The key intuition behind this definition is that the unique information should only depend on the marginal distribution of the pairs (Z, A) and (Z, B). This is motivated from an **operational** perspective that if A has unique information about Z (with respect to B), then there must be a situation where one can predict Z better using A than B (more details in [47, Section 2]). Therefore, all the joint distributions in the set Δ_p with the same marginals essentially have the same unique information, and the distribution Q^* that minimizes $\text{I}_Q(Z;A \mid B)$ is the joint distribution that has no synergistic information leading to $\text{I}_{Q^*}(Z;A \mid B) = \text{Uni}(Z:A|B)$. Definition 1 also defines $\text{Red}(Z:(A,B))$ and $\text{Syn}(Z:(A,B))$ using (2) and (3).

Definition 2 (Redundant Information [47])**.** *The redundant information about Z contained in both A and B is given by:*

$$\text{Red}(Z:(A,B)) = \text{I}(Z;A) - \text{Uni}(Z:A \mid B). \tag{7}$$

2.3 Background on Shapley Values

Here, we provide a background on Shapley values [51], a concept from cooperative game theory, that we use to arrive at our techniques of explainability. The Shapley value is a way of distributing a total reward generated by the coalition of all players. The Shapley values are a unique set of values satisfying certain assumptions [51].

Consider a set $\mathcal{N}$ (of $|\mathcal{N}| = n$ total players) and a function $v(S)$ that maps a subset of players $S \subseteq \mathcal{N}$ to a real number. Here, $v(\emptyset) = 0$, where $\emptyset$ denotes the empty set, and $v(\mathcal{N})$ is the total reward generated by the coalition of all players. In general, $v(S)$ is called the worth of coalition S, that describes the total expected reward that the members of S can obtain by cooperation. The function v is called a characteristic function.

According to the Shapley values, the reward-amount $\varphi_i(v)$ that player i is allocated in a coalition game specified by $(v, \mathcal{N})$ is: $\varphi_i(v) = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{|S|! \, (n-|S|-1)!}{n!} \big(v(S \cup \{i\}) - v(S) \big)$. Note that the sum extends over all subsets S of $\mathcal{N}$ not containing player i.

The Shapley values satisfy many desirable properties, of which we include the most relevant one here.

Property 1. *The sum of the Shapley values of all players equals the total reward, i.e., $\sum_{i \in \mathcal{N}} \varphi_i(v) = v(\mathcal{N})$.*

3 MAIN RESULTS

This section is organized as follows: we begin (Section 3.1) with a canonical example that illustrates the difference between *interventional* and *distributional* approaches to explaining contributions of individual features to the observed disparity. Next, in Section 3.2, we discuss some properties of our proposed *distributional* measure of contribution to disparity, focusing on the case when one does not have access to the underlying decision-making mechanism, e.g., with human-in-the-loop. In Section 3.3, we discuss some more canonical examples to further illustrate the differences between our two proposed approaches.

3.1 Interventional and Distributional Approaches to Explaining Contribution

We consider the following canonical example that represents scenarios where two or more features are highly correlated with each other.

Canonical Example 1 (Highly Dependent Input Features)**.** *Consider two highly dependent features[1] being used for deciding student admissions. These two features are the scores in two courses: $X_1 = Z + U$ and $X_2 = Z + U$. Here Z is the protected attribute, distributed as Bern(½), and U is another independent random variable with distribution Bern(½) that could potentially represent inner ability (inspired from existing works in causal fairness, e.g., [18]). Now, suppose that the admission decision happens to be based on a score given by: $\hat{Y} = X_1$.*

---

[1]Often, such features are regarded as co-linear features.

Let us now introduce our first approach to quantify the contribution of the features X_1 and X_2 when we have access to the model. Inspired from literature in explainability, we can vary an individual feature while keeping the other features unchanged, and assign the resulting change in disparity (mutual information between Z and the output) to this feature. So, in essence, the change due to X_1 in the observed disparity $\mathrm{I}(Z;\hat{Y})$ can be quantified as:

$$\mathrm{I}(Z;\hat{Y}(X)) - \mathrm{I}(Z;\hat{Y}(X\backslash X_1)),$$

where $\hat{Y}(X_S)$ denotes the output of the model/decision-making system that is based only on the features in the set $X_S \subseteq X$. In existing works on explainability [27], $\hat{Y}(X_S)$ is typically computed by setting the input features in the set $X\backslash X_S$ as constants for all data-points (e.g., equal to their respective means), and then evaluating the output of the model. Notice that, $\hat{Y}(X)$ is thus essentially $\hat{Y}$ which is the actual model output, and $\hat{Y}(\emptyset)$ is a constant. Thus, $\mathrm{I}(Z;\hat{Y}(X)) = \mathrm{I}(Z;\hat{Y})$ and $\mathrm{I}(Z;\hat{Y}(\emptyset)) = 0$.

A Shapley-value-based approach (building on this idea) would take an average of the change due to X_i across all possible subsets $X_S \subseteq X\backslash X_i$, leading to the following candidate measure:

Candidate Measure 1 (Interventional Contribution)**.**

$$\mathrm{Contri}(X_i) = \sum_{X_S \subseteq X\backslash X_i} \frac{|X_S|!\,(n - |X_S| - 1)!}{n!}$$
$$(\mathrm{I}(Z;\hat{Y}(X_S \cup X_i)) - \mathrm{I}(Z;\hat{Y}(X_S))). \tag{8}$$

Using this candidate measure of contribution for this canonical example, we get:

$$\mathrm{Contri}(X_1) = 1/2 \text{ bits.} \tag{9}$$
$$\mathrm{Contri}(X_2) = 0 \text{ bits.} \tag{10}$$
$$\mathrm{Contri}(X_1) + \mathrm{Contri}(X_2) = \mathrm{I}(Z;\hat{Y}) = 1/2 \text{ bits.} \tag{11}$$

This perspective is in alignment with the actual mechanism of how the decision is being made, which is based only on X_1. However, based on this explanation of contributions to disparity, the admissions committee could choose to drop the first feature, and retrain their model using the remaining features (here, X_2) in the hope that it would remove disparity. This would turn out to be misleading since, now, disparity can still arise from X_2. Thus, from a distributional perspective, we might want to capture this statistical redundancy between these two features X_1 and X_2, even if the second feature is not actually being used in the mechanism.

Furthermore, consider a scenario where an external observer does not have access to the actual mechanism of the model or how the decision is being made (also see Section 2.1). They do not know if the decision is a deterministic function of the input X, or if the decision is made with human-in-the-loop and their subjective evaluation as well (since the human has access to both of the features along with additional non-quantified aspects). The observer only gets to observe the inputs to the model, and the final output decisions for a dataset. In such a situation, it is not possible to intervene on the inputs to the decision-making system and compare the outputs for different values of the input. In fact, even if one retrains the original model with X_1 and X_2 as inputs, it may not converge to the same weights because of the *distributional redundancy* between the two features. Then, the question that for the observer is: *what is the "potential" contribution of each individual feature to the observed disparity* $\mathrm{I}(Z;\hat{Y})$ *when the exact decision-making mechanism is not accessible?*

For answering this question, in this work, we first examine an information-theoretic quantity called redundant information (recall Definition 2 in Section 2.2). Redundant information captures the redundant statistical dependency about Z present in the decision $\hat{Y}$ and a set of features X_S, and is written as $\mathrm{Red}(Z : (\hat{Y}, X_S))$.

A Shapley-value-based approach would take an average of the change in the redundant information due to X_i across all possible subsets $X_S \subseteq X\backslash X_i$, leading to the following:

Candidate Measure 2 (Potential Contribution)**.**

$$\mathrm{PotentContri}(X_i) = \sum_{X_S \subseteq X\backslash X_i} \frac{|X_S|!\,(n - |X_S| - 1)!}{n!}$$
$$(\mathrm{Red}(Z : (\hat{Y}, X_S \cup X_i)) - \mathrm{Red}(Z : (\hat{Y}, X_S))). \tag{12}$$

6

Now, let us see what $\text{PotentContri}(X_i)$ signifies for this canonical example. Notice that,

$$\text{Red}(Z : (\hat{Y}, X_1)) \overset{(a)}{=} \text{I}(Z; \hat{Y}) - \text{Uni}(Z : \hat{Y} \mid X_1)$$
$$\overset{(b)}{=} \text{I}(Z; \hat{Y}) = 1/2 \text{ bits.} \tag{13}$$

Here (a) holds from (2), and (b) holds because $\text{Uni}(Z : \hat{Y} \mid X_1) \leq \text{I}(Z; \hat{Y} \mid X_1) = 0$ (non-negativity of PID and (2); see Lemma 1 in Appendix). Because $\text{Red}(Z : (\hat{Y}, X_1))$ is a purely distributional quantity,

$$\text{Red}(Z : (\hat{Y}, X_2)) = \text{Red}(Z : (\hat{Y}, X_1)) = 1/2 \text{ bits.} \tag{14}$$

We also have $\text{Red}(Z : (\hat{Y}, X)) = 1/2$ bits, and $\text{Red}(Z : (\hat{Y}, \emptyset)) = 0$ bits. This leads to:

$$\text{PotentContri}(X_1) = 1/2(\text{Red}(Z : (\hat{Y}, X))$$
$$- \text{Red}(Z : (\hat{Y}, X_2))) + 1/2(\text{Red}(Z : (\hat{Y}, X_1))$$
$$- \text{Red}(Z : (\hat{Y}, \emptyset))) = 1/4 \text{ bits.} \tag{15}$$
$$\text{PotentContri}(X_2) = 1/2(\text{Red}(Z : (\hat{Y}, X))$$
$$- \text{Red}(Z : (\hat{Y}, X_1))) + 1/2(\text{Red}(Z : (\hat{Y}, X_2))$$
$$- \text{Red}(Z : (\hat{Y}, \emptyset))) = 1/4 \text{ bits.} \tag{16}$$
$$\text{PotentContri}(X_1) + \text{PotentContri}(X_2)$$
$$= \text{I}(Z; \hat{Y}) = 1/2 \text{ bits.} \tag{17}$$

3.2 Properties of our Distributional Measure of Contribution

Our distributional approach to quantifying contributions can identify if any additional disparity has been introduced in the decision-making process from non-quantified aspects, e.g., due to human-in-the-loop who almost always have access to the protected attributes.

Theorem 1 (Unexplained Disparity). *In general, we have* $\text{PotentContri}(X_i) \leq \text{I}(Z; \hat{Y})$. *An equality holds if and only if* $\text{Uni}(Z : \hat{Y} \mid X) = 0$.

Proof. In general, we have,

$$\sum_{X_i \in X} \text{PotentContri}(X_i) = \text{Red}(Z : (\hat{Y}, X))$$

(property of Shapley values; see Property 1 in Section 2.3). Now, observe that,

$$\text{Red}(Z : (\hat{Y}, X)) = \text{I}(Z; \hat{Y}) - \text{Uni}(Z : \hat{Y} \mid X) \overset{(a)}{\leq} \text{I}(Z; \hat{Y}),$$

where (a) holds from the non-negativity of PID terms. An equality holds if and only if $\text{Uni}(Z : \hat{Y} \mid X) = 0$. □

Remark 2 (Significance of Theorem 1). *Since the observer does not have access to the model, they do not know if the decision $\hat{Y}$ is an entirely deterministic function of the input X (or not, e.g., because of human-in-the-loop). When $\hat{Y}$ is entirely determined by input X, the Markov chain $Z - X - \hat{Y}$ holds, implying $\text{Uni}(Z : \hat{Y} \mid X) = 0$ (Lemma 1 in Appendix). The sum of the individual contributions of all the features also add up to the observed disparity $\text{I}(Z; \hat{Y})$. However, when $\hat{Y}$ is not an entirely deterministic function of the input X, $\text{Uni}(Z : \hat{Y} \mid X)$ may be non-zero. This signifies that there is some unique information about Z in the output $\hat{Y}$ that cannot be attributed to the input features X, implying additional unexplained disparity introduced in the decision-making process.*

Theorem 2 (Bounds on PotentContri). *For any feature X_i, we have $0 \leq \text{PotentContri}(X_i) \leq \text{I}(Z; \hat{Y})$.*

Proof. For the lower-bound, it suffices to show that

$$\text{Red}(Z : (\hat{Y}, X_S \cup X_i)) \geq \text{Red}(Z : (\hat{Y}, X_S))$$

7

for any set $X_S \subseteq X \setminus X_i$. Note that,

$$\text{Red}(Z : (\hat{Y}, X_S \cup X_i))$$

$$= \text{I}(Z; \hat{Y}) - \text{Uni}(Z : \hat{Y} \mid X_S \cup X_i) \tag{18}$$

$$\overset{(a)}{\geq} \text{I}(Z; \hat{Y}) - \text{Uni}(Z : \hat{Y} \mid X_S) \tag{19}$$

$$= \text{Red}(Z : (\hat{Y}, X_S)). \tag{20}$$

Here (a) holds from a property of unique information (see Lemma 2 in Appendix).

For the upper bound, notice that

$$\text{PotentContri}(X_i) \overset{(a)}{\leq} \sum_{X_i \in X} \text{PotentContri}(X_i)$$

$$\overset{(b)}{=} \text{Red}(Z : (\hat{Y}, X)) \overset{(c)}{\leq} \text{I}(Z; \hat{Y}). \tag{21}$$

Here (a) holds from non-negativity of $\text{PotentContri}(X_i)$ (proved just above), (b) holds from Theorem 1, and (c) holds since PID is a non-negative decomposition. $\qquad\square$

Next, we examine more canonical examples to better understand the *interventional* and *distributional* approaches.

3.3   More Canonical Examples

Canonical Example 2 (Disparity Amplification)**.** *Consider two features: score in two courses, $X_1 = Z + U$ and $X_2 = U$. Here Z is the protected attribute, distributed as Bern(½), and $U \sim Bern(½)$ is another independent random variable (e.g. representing inner ability). Now, suppose the decision is: $\hat{Y} = X_1 - X_2 = Z$.*

This is a case of disparity amplification. Here, though X_2 does not individually have any information about Z (indeed $\text{I}(Z; X_2) = 0$), it leads to an increase in the observed disparity in the final output $\hat{Y}$. We have $\text{I}(Z; \hat{Y}) = 1$ bit which would not have been possible from using the feature X_1 alone, since $\text{I}(Z; X_1)$ is much less. Thus, we would want X_2 to also be assigned a contribution to the overall disparity.

For this example, the interventional perspective to quantifying contributions would lead to the following explanations:

$$\text{Contri}(X_1) = 3/4 \text{ bits.} \tag{22}$$

$$\text{Contri}(X_2) = 1/4 \text{ bits.} \tag{23}$$

$$\text{Contri}(X_1) + \text{Contri}(X_2) = \text{I}(Z; \hat{Y}) = 1 \text{ bit.} \tag{24}$$

Interestingly, the distributional perspective to quantifying "potential" contributions would also lead to the same:

$$\text{PotentContri}(X_1) = 3/4 \text{ bits.} \tag{25}$$

$$\text{PotentContri}(X_2) = 1/4 \text{ bits.} \tag{26}$$

$$\text{PotentContri}(X_1) + \text{PotentContri}(X_2) = \text{I}(Z; \hat{Y})$$

$$= 1 \text{ bit.} \tag{27}$$

Thus, both the approaches are able to quantify the contribution of the individual features to the observed disparity, as desired.

Canonical Example 3 (Disparity Masking)**.** *Let $Z \sim Bern(½)$ be the protected attribute, and $U \sim Bern(½)$ be another independent random variable (e.g. representing inner ability). Consider two features: $X_1 = Z$ (the protected attribute) and $X_2 = U$ (score in a course). Now, suppose that the admission decision is: $\hat{Y} = X_1 \oplus X_2 = Z \oplus U$.*

This is a case of masked discrimination: high-scoring candidates of the non-protected group and low-scoring candidates of the protected group are admitted. From an observational perspective, there is no disparity in the overall
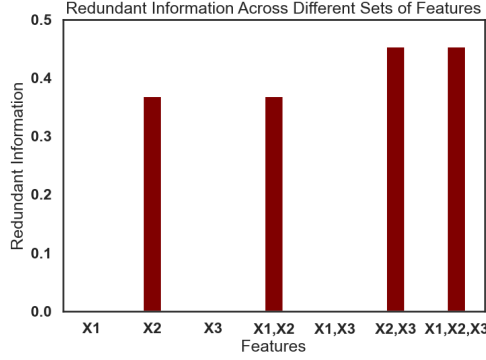
Figure 3: Understanding redundant information about gender Z in $\hat{Y}$ and subset $S \subseteq (X_1, X_2, X_3)$ across different subsets of features, i.e., $\mathrm{Red}(Z : (\hat{Y}, S))$: Notice that, the features (X_2, X_3) jointly have higher redundant information about Z with $\hat{Y}$ than either of the features alone.

decisions. This is also exhibited by the fact that $\mathrm{I}(Z; \hat{Y}) = 0$. However, a careful examination using interventions reveal that the decisions are discriminatory towards high-scoring candidates of the protected group who are not admitted.

For this example, the interventional perspective leads to the following contributions:

$$\mathrm{Contri}(X_1) = 1/2 \text{ bits and } \mathrm{Contri}(X_2) = -1/2 \text{ bits.}$$

$$\mathrm{Contri}(X_1) + \mathrm{Contri}(X_2) = \mathrm{I}(Z; \hat{Y}) = 0 \text{ bits.} \tag{28}$$

On the other hand, the distributional perspective for "potential" contributions leads to the following:

$$\mathrm{PotentContri}(X_1) = 0 \text{ bits.} \tag{29}$$

$$\mathrm{PotentContri}(X_2) = 0 \text{ bits.} \tag{30}$$

$$\mathrm{PotentContri}(X_1) + \mathrm{PotentContri}(X_2) = \mathrm{I}(Z; \hat{Y}) = 0 \text{ bits.} \tag{31}$$

Remark 3 (On Negative Contribution). *We note that while* $\mathrm{PotentContri}$ *is always non-negative,* Contri *can take negative values. Here, a negative value of* Contri *demonstrates that a feature (X_2), in fact, reduces (effectively masks) the overall disparity that was introduced by another feature (X_1).*

4   CASE STUDY

Here, we include a simple case study on an artificial dataset to demonstrate how to compute potential contributions of features on datasets.

Example 2. *We consider a scenario of college admissions with three features: discrete scores corresponding to the GPA, GRE, and Recommendation Letters. Let $Z \sim Bern(\frac{1}{2})$ denote gender, $X_1 = U_1 \sim Bern(\frac{9}{10})$ denote whether GPA is above a desirable threshold, $X_2 = Z + U_2$ (with $U_2 \sim Bern(\frac{1}{2})$) denote a thresholded GRE score, and $X_3 = U_2$ denote a score from reference letters. Here, U_1, and U_2 can be thought of as latent random variables denoting inner abilities of a candidate that are independent of Z. Suppose that the admission committee uses an automated figure of merit $F = Z + U_1 + 2U_2$ whose exact mechanism is not known to us (could be something like $F = X_1 + X_2 + X_3$ or $F = X_1 + 2X_2 - X_3$). We can only observe the final decisions based on $\hat{Y} = F + R = Z + U_1 + 2U_2 + N$ where $N \sim Bern(\frac{1}{10})$ denotes a manual evaluation score (which might be subjective and not a deterministic function of the inputs X_1, X_2, X_3).*

In Fig. 3 and Fig. 4, we employ our technique to quantify the potential contribution of each individual feature to the overall disparity $\mathrm{I}(Z; \hat{Y})$. We use the discrete information theory (`dit`) [52] package to compute all of the PID terms. The package solves a convex optimization problem for the computation of unique information using the empirical distribution of the data (10K samples).
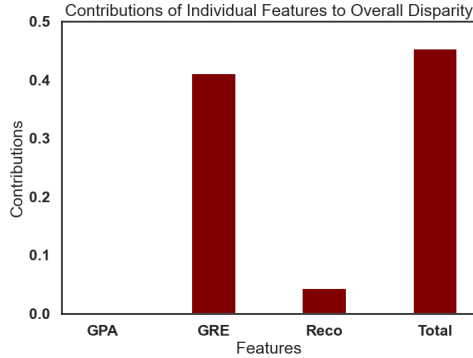
Figure 4: Quantifying Potential Contribution of Individual Features to the Overall Disparity I$(Z; \hat{Y})$ based on Redundant Information: Notice that, the feature Reco (X_3) is also assigned a potential contribution since (X_2, X_3) jointly have a higher redundant information about Z with $\hat{Y}$ than either of them alone.

Remark 4 (Early Truncation for Scalability). *We note that there are computational challenges as we scale this technique to larger number of features. To address this, we employ an approximation that is based on the following critical observation:* $\mathrm{Red}(Z : (\hat{Y}, X_S))$ *is non-decreasing as the number of features in the set X_S increases (Lemma 3 in Appendix), and is upper bounded by* I$(Z; \hat{Y})$. *Thus, if we arrive at a value of* $\mathrm{Red}(Z : (\hat{Y}, X_S))$ *such that* $|\mathrm{Red}(Z : (\hat{Y}, X_S)) - \mathrm{I}(Z; \hat{Y})| \leq \epsilon$ *for a small choice of ϵ, then we can approximate* $|\mathrm{Red}(Z : (\hat{Y}, X_S' \cup X_i)) - \mathrm{Red}(Z : (\hat{Y}, X_S'))|$ *with ϵ for all supersets $X_S' \supseteq X_S$. In practice, this can reduce the number of sets X_S on which we need to compute* $\mathrm{Red}(Z : (\hat{Y}, X_S))$.

5 CONCLUSION AND BROADER IMPACT

This work proposes two novel approaches to quantifying contribution of individual features to the overall disparity (I$(Z; \hat{Y})$) in a decision, and discusses when either is better suited. These techniques could be applicable for auditing decision-making mechanisms: when one is able to intervene or not intervene on the inputs, as well as in scenarios with human-in-the-loop. Future work will examine how these approaches can inform intervention and repair of the decision-making mechanism to reduce disparity.

We note that the computational complexity of computing Shapley values definitely becomes a challenge for complex neural networks with very large number of features/neurons that are used in computer vision or natural language processing. In future work, we will also explore techniques of approximating Shapley values for faster computation, e.g., in [30]. However, we envision our technique to still be applicable for smaller models used in more consequential applications, e.g., college admissions, or credit decision that use simple models with a few features and human-in-the-loop [44].

Another future work could be improving the estimation of these quantities for continuous random variables (either by discretizing them [18] or leveraging techniques in [33, 53–56] for the ease of computation).

References

[1] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 2012, pp. 214–226.

[2] F. Kamiran, I. Žliobaitė, and T. Calders, "Quantifying explainable discrimination and removing illegal discrimination in automated decision making," *Knowledge and information systems*, vol. 35, no. 3, pp. 613–644, 2013.

[3] A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A reductions approach to fair classification," in *International Conference on Machine Learning*. PMLR, 2018, pp. 60–69.

[4] M. Hardt, E. Price, N. Srebro *et al.*, "Equality of opportunity in supervised learning," in *Advances in Neural Information Processing Systems*, 2016, pp. 3315–3323.

[5] K. R. Varshney, "Trustworthy machine learning and artificial intelligence," *XRDS: Crossroads, The ACM Magazine for Students*, vol. 25, no. 3, pp. 26–29, 2019.

[6] F. P. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney, "Optimized pre-processing for discrimination prevention," in *Advances in Neural Information Processing Systems*, 2017, pp. 3992–4001.

[7] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. R. Varshney, "Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing," in *International Conference on Machine Learning (ICML)*, 2020.

[8] J. Cho, G. Hwang, and C. Suh, "A fair classifier using mutual information," 2020.

[9] H. Wang, H. Hsu, M. Diaz, and F. P. Calmon, "To split or not to split: The impact of disparate treatment in classification," *IEEE Transactions on Information Theory*, 2021.

[10] A. Datta, M. Fredrikson, G. Ko, P. Mardziel, and S. Sen, "Use privacy in data-driven systems: Theory and experiments with machine learnt programs," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2017, pp. 1193–1210.

[11] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Advances in Neural Information Processing Systems*, 2017, pp. 656–666.

[12] J. Zhang and E. Bareinboim, "Fairness in decision-making—the causal explanation formula," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[13] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '17. ACM, 2017, pp. 797–806.

[14] R. Nabi and I. Shpitser, "Fair inference on outcomes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[15] S. Chiappa, "Path-specific counterfactual fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 7801–7808.

[16] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu, "Interventional Fairness: Causal Database Repair for Algorithmic Fairness," in *Proceedings of the 2019 International Conference on Management of Data*, ser. SIGMOD '19. ACM, 2019, pp. 793–810.

[17] S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover, "An information-theoretic quantification of discrimination with exempt features," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 3825–3833.

[18] S. Dutta, P. Venkatesh, P. Mardziel, A. Datta, and P. Grover, "Fairness under feature exemptions: Counterfactual and observational measures," *IEEE Transactions on Information Theory*, 2021.

[19] R. Xu, P. Cui, K. Kuang, B. Li, L. Zhou, Z. Shen, and W. Cui, "Algorithmic decision making with conditional fairness," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 2125–2135.

[20] J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell, "Fairness under unawareness: Assessing disparity when protected class is unobserved," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 339–348.

[21] "It's Time for an Honest Conversation about Graduate Admissions," https://news.ets.org/stories/its-time-for-an-honest-conversation-about-graduate-admissions/, 2020.

[22] "The Problem With the GRE," https://www.theatlantic.com/education/archive/2016/03/the-problem-with-the-gre/471633/, 2016.

[23] EEOC Website, "Title VII of the Civil Rights Act of 1964," https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964, 1964.

[24] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[25] S. S. Grover, "The business necessity defense in disparate impact discrimination cases," *Ga. L. Rev.*, vol. 30, p. 387, 1995.

[26] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE Symposium on Security and Privacy (SP)*, 2016, pp. 598–617.

[27] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[28] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[29] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, pp. 1885–1894.

[30] A. Ghorbani and J. Zou, "Neuron shapley: Discovering the responsible neurons," *arXiv preprint arXiv:2002.09815*, 2020.

[31] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 2012.

[32] C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.

[33] S. Galhotra, K. Shanmugam, P. Sattigeri, and K. R. Varshney, "Fair data integration," *arXiv preprint arXiv:2006.06053*, 2020.

[34] S. Khodadadian, M. Nafea, A. Ghassami, and N. Kiyavash, "Information theoretic measures for fairness-aware feature selection," *arXiv preprint arXiv:2106.00772*, 2021.

[35] W. Pan, S. Cui, J. Bian, C. Zhang, and F. Wang, "Explaining algorithmic fairness through fairness-aware causal path decomposition," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 1287–1297.

[36] Y. Ge, J. Tan, Y. Zhu, Y. Xia, J. Luo, S. Liu, Z. Fu, S. Geng, Z. Li, and Y. Zhang, "Explainable fairness in recommendation," *arXiv preprint arXiv:2204.11159*, 2022.

[37] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The building blocks of interpretability," *Distill*, 2018. [Online]. Available: https://distill.pub/2018/building-blocks

[38] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 3319–3328.

[39] P. Dabkowski and Y. Gal, "Real time image saliency for black box classifiers," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017, pp. 6970–6979.

[40] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 3016–3022, main track.

[41] B. Kim, O. Koyejo, R. Khanna *et al.*, "Examples are not enough, learn to criticize! Criticism for interpretability." in *NIPS*, 2016, pp. 2280–2288.

[42] H. Harutyunyan, A. Achille, G. Paolini, O. Majumder, A. Ravichandran, R. Bhotika, and S. Soatto, "Estimating informativeness of samples with smooth unique information," *arXiv preprint arXiv:2101.06640*, 2021.

[43] P. Venkatesh, S. Dutta, and P. Grover, "Information flow in computational systems," *IEEE Transactions on Information Theory*, vol. 66, no. 9, pp. 5456–5491, 2020.

[44] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison, "A machine learning framework to identify students at risk of adverse academic outcomes," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '15. New York, NY, USA: ACM, 2015, p. 1909–1918.

[45] P. Adler, C. Falk, S. A. Friedler, T. Nix, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing black-box models for indirect influence," *Knowledge and Information Systems*, vol. 54, no. 1, pp. 95–122, 2018.

[46] I. E. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler, "Problems with shapley-value-based explanations as feature importance measures," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5491–5500.

[47] N. Bertschinger, J. Rauh, E. Olbrich, J. Jost, and N. Ay, "Quantifying unique information," *Entropy*, vol. 16, no. 4, pp. 2161–2183, 2014.

[48] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," *arXiv preprint arXiv:1004.2515*, 2010.

[49] V. Griffith and C. Koch, "Quantifying synergistic mutual information," in *Guided Self-Organization: Inception*. Springer, 2014, pp. 159–190.

[50] P. K. Banerjee, J. Rauh, and G. Montúfar, "Computing the unique information," in *2018 IEEE International Symposium on Information Theory (ISIT)*, 2018, pp. 141–145.

[51] L. S. Shapley, *A value for n-person games*. Princeton University Press, 1953.

[52] R. G. James, C. J. Ellison, and J. P. Crutchfield, "dit: a Python package for discrete information theory," *The Journal of Open Source Software*, vol. 3, no. 25, p. 738, 2018.

[53] S. Mukherjee, H. Asnani, and S. Kannan, "CCMI: Classifier based Conditional Mutual Information estimation," *arXiv preprint arXiv:1906.01824*, 2019.

[54] P. Venkatesh, S. Dutta, N. Mehta, and P. Grover, "Can information flows suggest targets for interventions in neural circuits?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 3149–3162, 2021.

[55] G. Schamberg and P. Venkatesh, "Partial information decomposition via deficiency for multivariate gaussians," *arXiv preprint arXiv:2105.00769*, 2021.

[56] A. Pakman, A. Nejatbakhsh, D. Gilboa, A. Makkeh, L. Mazzucato, M. Wibral, and E. Schneidman, "Estimating the unique information of continuous variables," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20 295–20 307, 2021.

[57] P. K. Banerjee, E. Olbrich, J. Jost, and J. Rauh, "Unique informations and deficiencies," in *2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2018, pp. 32–38.

Disclaimer

Appendix

Lemma 1 (Zero Unique Information). *When the Markov chain $Z - X - \hat{Y}$ holds, we have* $\mathrm{Uni}(Z : \hat{Y} \mid X) = 0$.

From the definition of PID, we have, $\mathrm{I}(Z; \hat{Y} \mid X) = \mathrm{Uni}(Z : \hat{Y}|X) + \mathrm{Syn}(Z : (\hat{Y}, X))$ where both the terms are non-negative. Thus, $\mathrm{Uni}(Z : \hat{Y}|X) \leq \mathrm{I}(Z; \hat{Y} \mid X)$. Notice that, when the Markov chain $Z - X - \hat{Y}$ holds, we have $\mathrm{I}(Z; \hat{Y} \mid X) = 0$, thus implying $\mathrm{Uni}(Z : \hat{Y}|X) = 0$ as well.

Lemma 2 (Monotonicity of Unique Information). *For all (Z, B, X_c, X'_c), we have:*

$$\mathrm{Uni}(Z : B|X_c \cup X'_c) \leq \mathrm{Uni}(Z : B|X_c).$$

This result is derived in [57, Lemma 32].

Lemma 3 (Monotonicity of Redundant Information). *For all (Z, B, X_c, X'_c), we have:*

$$\mathrm{Red}(Z : (B, X_c \cup X'_c)) \geq \mathrm{Red}(Z : (B, X_c)).$$

The proof holds from Lemma 2, since $\mathrm{Uni}(Z : B|X_c \cup X'_c) + \mathrm{Red}(Z : (B, X_c \cup X'_c)) = \mathrm{I}(Z; B) = \mathrm{Uni}(Z : B|X_c) + \mathrm{Red}(Z : (B, X_c))$.