Manifold Restricted Interventional Shapley Values

Muhammad Faaiz Taufiq[1]
University of Oxford

Patrick Blöbaum
Amazon Research

Lenon Minorics
Amazon Research

Abstract

Shapley values are model-agnostic methods for explaining model predictions. Many commonly used methods of computing Shapley values, known as *off-manifold methods*, rely on model evaluations on out-of-distribution input samples. Consequently, explanations obtained are sensitive to model behaviour outside the data distribution, which may be irrelevant for all practical purposes. While *on-manifold methods* have been proposed which do not suffer from this problem, we show that such methods are overly dependent on the input data distribution, and therefore result in unintuitive and misleading explanations. To circumvent these problems, we propose *ManifoldShap*, which respects the model's domain of validity by restricting model evaluations to the data manifold. We show, theoretically and empirically, that ManifoldShap is robust to off-manifold perturbations of the model and leads to more accurate and intuitive explanations than existing state-of-the-art Shapley methods.

1 INTRODUCTION

Explaining model predictions is highly desirable for reliable applications of machine learning. This is especially important in risk-sensitive settings like medicine and credit scoring [Hakkoum et al., 2022, Lee et al., 2019, Ahmad et al., 2018, Kvamme et al., 2018] where an incorrect model prediction could prove very costly. Explainability is becoming increasingly relevant because of regulations like the General Data Protection Regulation [Regulation, 2016], which may require being able to explain model predictions before deploying a model in the real world. This is less of a challenge in models like linear models and decision trees, which tend to be easier to interpret. However, the same is not true for more complex models like Neural Networks, where explaining predictions may not be straightforward [Ribeiro et al., 2016].

Explainable AI is an area of machine learning which aims to provide methodologies for interpreting model predictions. Various different techniques of explaining models have been proposed, with each approach satisfying different properties [Linardatos et al., 2021]. In this paper, we focus on Shapley values [Strumbelj and Kononenko, 2010, 2014, Lundberg and Lee, 2017], a popular approach for quantifying feature relevance, which is model-agnostic, i.e., is independent of model implementation. Additionally, this is a local explanation method, i.e., it can be used to explain individual model predictions. Shapley values are based on ideas from cooperative game theory [Bilbao, 2000] and come with various desirable theoretical properties [Sundararajan and Najmi, 2020] which make it a very attractive method in practice.

At a high-level, Shapley values treat features as 'players' in a game, where the total payout is the model prediction at a given point. To quantify the feature importance, this method distributes the total payout among each player in a 'fair' manner using a *value* function. Different types of Shapley value functions have been proposed which differ in the way they distribute payout among players [Sundararajan and Najmi, 2020, Frye et al., 2021]. These can be broadly divided into two categories: (i) *on-manifold* value functions, which only depend on the model behaviour on the input data distribution, and (ii) *off-manifold* value functions which also depend on the model behaviour outside the input data distribution.

Off-manifold Shapley values are not robust to changes in model behaviour outside the data distribution. This means that the explanations obtained using these methods may be highly influenced if the model behaviour outside the data distribution changes, even if it remains fixed on the data distribution [Frye et al., 2021, Slack et al., 2020, Yeh et al., 2022]. Such changes to the model can change the Shapley values drastically, resulting in misleading explanations, and can even be used to hide model biases. On the other hand, while the on-manifold Shapley values are robust to

[1]Work done during internship at Amazon Research. Correspondence to: muhammad.taufiq@stats.ox.ac.uk.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

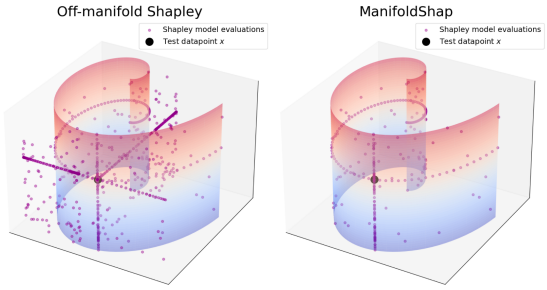


Figure 1: The datapoints at which model is evaluated when computing Shapley values for test point $\mathbf{x}$, along with the data manifold. Off-manifold methods evaluate the model outside the data manifold whereas our proposal, ManifoldShap, restricts model evaluations to the data manifold.

such model perturbations, the explanations obtained are highly sensitive to changes in the feature distribution. Additionally, these methods do not capture the *causal* contribution of features as they attribute importance based on feature correlations. For example, we show that on-manifold Shapley values can be 'fooled' into attributing similar importance to two positively correlated features, even if the model depends on only one of them.

In this paper, we bridge this gap between *on-manifold* and *off-manifold* Shapley values by proposing ManifoldShap (illustrated in Figure 1), a Shapley value function, which remains robust to changes in model behaviour outside the data distribution, while estimating the *causal* contribution of features. We show that ManifoldShap is significantly less sensitive to changes in the feature distribution than other on-manifold value functions. We extend the formal notion of robustness in Yeh et al. [2022] by providing an alternative definition which may be more desirable in many cases. We additionally show that our proposed method satisfies both notions of robustness, while other methods do not. Moreover, ManifoldShap satisfies a number of other desirable properties which we verify theoretically and empirically on real-world datasets.

2 SHAPLEY VALUES

In this section, we will introduce Shapley values for model explainability. For any given model $f : \mathcal{X} \to \mathcal{Y}$, our goal is to obtain localised model explanations at a given point $\mathbf{x} \in \mathcal{X}$. We assume that $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$.

Shapley values [Strumbelj and Kononenko, 2010, 2014, Lundberg and Lee, 2017] provide a natural tool for obtaining such explanations. For a specific input $\mathbf{x}$, Shapley values define a way of distributing the difference between $f(\mathbf{x})$ and a baseline, which we denote as b_0, among the d input features. This can naturally be interpreted as the contribution of each feature towards the difference $f(\mathbf{x}) - b_0$, and is commonly referred to as feature attributions. One possible choice of baseline explored in the literature is the model evaluated at an auxiliary input $\mathbf{x}'$, i.e., $b_0 = f(\mathbf{x}')$. Alternatively, many methods use the average model output $\mathbb{E}[f(\mathbf{X})]$ as the baseline, i.e., $b_0 = \mathbb{E}[f(\mathbf{X})]$. This can be used to explain *why* the output at a point $\mathbf{x}$ deviates from the average output. The average output provides a more intuitive and interpretable baseline compared to the choice of an auxiliary input $\mathbf{x}'$, which can be arbitrary. In this work, we therefore restrict our attention to the latter category.

As an example, consider a model which predicts an individual's salary, with input features corresponding to individual's information. If feature $i \in [d]$ represents the age of the individual, the attribution for feature i, which we will denote as ϕ_i, tells us the contribution of individual's age to the salary prediction for $\mathbf{x}$, relative to the average salary prediction, i.e., $f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})]$. To compute the contribution for feature i at $\mathbf{x}$, Shapley values consider a value function $v : 2^{[d]} \to \mathbb{R}$ where v may implicitly depend on $\mathbf{x}$. Given a subset $S \subseteq [d] \setminus \{i\}$, we can intuitively interpret the difference $v(S \cup \{i\}) - v(S)$ as the contribution of feature i w.r.t. the set S. Next, the Shapley values for feature i is defined as a weighted sum over all possible subsets S:

$$\phi_i := \sum_{S \subseteq [d] \setminus \{i\}} \frac{|S|!(d - |S| - 1)!}{d!} (v(S \cup \{i\}) - v(S)).$$

The quantity ϕ_i can be intuitively considered as the average contribution of feature i to the prediction at $\mathbf{x}$. In order for the explanations obtained to be interpretable and intuitive, the value function v must be chosen such that it satisfies a number of desirable properties. We present some of the most important such properties here:

1. *Sensitivity:* If f does not depend on x_i, then $v(S \cup \{i\}) = v(S)$, and hence $\phi_i = 0$.
2. *Symmetry:* If f is symmetric in components i and j and $x_i = x_j$, then $v(S \cup \{i\}) = v(S \cup \{j\})$ and hence $\phi_i = \phi_j$.
3. *Efficiency:* If ϕ_i denotes the attribution of feature i to $f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})]$, then $v([d]) - v(\emptyset) = f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})]$ and hence, $\sum_i \phi_i = f(\mathbf{x}) - \mathbb{E}[f(\mathbf{X})]$.

Next, we present various commonly used value functions, which can be classified into *off-manifold* and *on-manifold* value functions.

2.1 Off-Manifold Value Functions

This class of value functions does not restrict function evaluations to the data distribution, and consequently, computing Shapley values involves evaluating the model on out-of-distribution inputs, where the model has not been trained (see Figure 1). The most commonly used off-manifold value function is Marginal Shapley (MS) (also called RBShap [Sundararajan and Najmi, 2020]):

Marginal Shapley (MS).

$$v^{\text{MS}}_{\mathbf{x},f}(S) \coloneqq \mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})].$$

Specifically, Marginal Shapley takes the expectation of $f(\mathbf{x}_s, \mathbf{X}_{\bar{S}})$ over the marginal density of $\mathbf{X}_{\bar{S}}$.

In addition to this, there has been some recent work proposing a causal perspective when computing Shapley values [Janzing et al., 2020, Heskes et al., 2020, Jung et al., 2022]. Specifically, these works observe that manually fixing the values of features $\mathbf{X}_S$ to $\mathbf{x}_S$ when computing Shapley values, corresponds to *intervening* on the feature values. In Pearl's do calculus [Pearl, 2000, 2012], this is expressed as $do(\mathbf{X}_S = \mathbf{x}_S)$. This leads to the definition of Interventional Shapley (IS) value functions:

Interventional Shapley (IS).

$$v^{\text{IS}}_{\mathbf{x},f}(S) \coloneqq \mathbb{E}[f(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S)]. \tag{1}$$

A detailed discussion of how Interventional Shapley differs from other *non-causal* value functions has been deferred to Section 2.4. How to compute $v^{\text{IS}}_{\mathbf{x},f}(S)$ depends on the causal structure of the features. Janzing et al. [2020] only consider the causal relations between the function inputs and outputs, rather than between the real-world features and the true output Y. This corresponds to the set-up in Figure 2, where the true feature values $\tilde{X}_i$ are formally distinguished from the features X_i input into the function, f, with X_i being a direct causal descendant of $\tilde{X}_i$ and no interactions between X_i. In this set-up, intervening on $\mathbf{X}_S$ yields the following interventional distribution:

$$p(\mathbf{X}_{\bar{S}} \mid do(\mathbf{X}_S = \mathbf{x}_S)) = p(\mathbf{X}_{\bar{S}}).$$

In this case, the value function, $v^{\text{IS}}_{\mathbf{x},f}(S)$ can straightforwardly be computed as

$$v^{\text{IS}}_{\mathbf{x},f}(S) = \mathbb{E}[f(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S)] = \mathbb{E}_{\mathbf{X}_{\bar{S}} \sim p(\mathbf{X}_{\bar{S}})}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})].$$

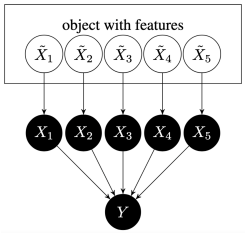


Figure 2: Causal structure considered in Janzing et al. [2020]. The true features are $\tilde{X}_i$ while features input into the model are X_i.

This is equivalent to Marginal Shapley. Therefore, Marginal Shapley can be considered a special case of Interventional Shapley. In contrast, Heskes et al. [2020] seeks to estimate the causal contributions of the real-world features towards the true output Y, and therefore, does not distinguish between the true features and the features input into the model. The resulting IS value function also takes into account the causal relations among the true features themselves.

2.2 On-Manifold Value Functions

These value functions only rely on function values in data distribution when computing Shapley values. As a result, any changes in the function outside data distribution does not change the explanations obtained. One of the first on-manifold value functions proposed was Conditional Expectation Shapley (CES) [Sundararajan and Najmi, 2020]:

Conditional Expectation Shapley (CES).

$$v^{\text{CES}}_{\mathbf{x},f}(S) \coloneqq \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S].$$

Unlike Marginal Shapley, CES takes the expectation of $f(\mathbf{x}_s, \mathbf{X}_{\bar{S}})$ over the conditional density of $\mathbf{X}_{\bar{S}}$ given $\mathbf{X}_S = \mathbf{x}_S$ (and not the marginal density of $\mathbf{X}_{\bar{S}}$). This has undesired implications for the obtained Shapley values, which we discuss in detail in Section 2.4.

Apart from this, recently Yeh et al. [2022] proposed Joint Baseline Shapley (JBShap), a value function which aims to make Shapley values robust to model changes in regions of low data-density. This value function explicitly takes the density $p(\mathbf{x})$ into consideration when calculating explanations:

Joint Baseline Shapley (JBShap).

$$v^{\text{J}}_{\mathbf{x},f,p}(S) \coloneqq f(\mathbf{x}_S, \mathbf{x}'_{\bar{S}}) p(\mathbf{x}_S, \mathbf{x}'_{\bar{S}}),$$

where $\mathbf{x}'$ is an auxiliary baseline. The authors also propose an extension of JBShap, called *Random Joint Baseline Shapley* (RJBShap) where the value function averages over all possible baseline values:

Random Joint Baseline Shapley (RJBShap).

$$v^{\text{RJ}}_{\mathbf{x},f,p}(S) \coloneqq \mathbb{E}_{p_b(\mathbf{X}_{\bar{S}})}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}}) p(\mathbf{x}_S, \mathbf{X}_{\bar{S}})].$$

Here, $p_b(\mathbf{X}_{\bar{S}})$ is some prior distribution over features $\mathbf{x}'_{\bar{S}}$. A natural choice of prior is the marginal density $p(\mathbf{X}_{\bar{S}})$, which we use to compute RJBShap later.

Having listed the most relevant on and off manifold value functions, we discuss their limitations in the following sections. This will motivate our proposal of an alternative value function, which aims to circumvent these limitations.

2.3 Limitations of off-manifold value functions

As Slack et al. [2020], Frye et al. [2021] point out, dependence of Shapley explanations on off-manifold behaviour

of the model can be problematic. For example, computing Interventional Shapley at $\mathbf{x}$ requires evaluating the model at points $(\mathbf{x}_S, \mathbf{X}_{\bar{S}})$ for $S \subseteq [d]$ where $\mathbf{X}_{\bar{S}} \sim p(\mathbf{X}_{\bar{S}} \mid do(\mathbf{X}_S = \mathbf{x}_S))$. Such points may lie outside the distribution of training data, where the model was not trained. Consider a model which is identical to the ground truth function on the data distribution. The train/test errors of the model will be 0, suggesting that it captures the ground truth function perfectly. However, if the model differs from the ground truth outside the data distribution, the model's Shapley values may be drastically different from the ground truth Shapley values, resulting in highly misleading explanations.

This limitation of off-manifold Shapley values can be exploited to 'fool' Shapley values into hiding model biases. In Slack et al. [2020], the authors consider models which are highly biased on the data manifold (i.e., solely rely on sensitive features, like racial background, for predictions). They show that these models can be perturbed outside the data manifold in such a way that the resulting Shapley values give no attribution to the sensitive features, despite the models relying solely on these sensitive features on the data manifold. Therefore, off-manifold Shapley values are highly vulnerable to off-manifold manipulations.

2.4 Limitations of on-manifold value functions

While the on-manifold value functions do not consider model behaviour outside data distribution, the existing methods can lead to unintuitive or misleading Shapley explanations as they do not consider the *causal* contributions of features, and are highly sensitive to feature correlations. Specifically, as Janzing et al. [2020] point out, when computing feature contributions at $\mathbf{x}$, the value function for a subset S, $v(S)$, must capture the effect of fixing the feature values $\mathbf{X}_S$ to $\mathbf{x}_S$. This is *not* given by $\mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$ as in CES, because observing $\mathbf{X}_S = \mathbf{x}_S$ also changes the distribution of $\mathbf{X}_{\bar{S}}$. Instead, the impact of setting $\mathbf{X}_S$ to $\mathbf{x}_S$ is captured by $\mathbb{E}[f(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S)]$, which in general is different from conditional expectation. Therefore, Interventional Shapley is inherently proposed to capture the *causal* effect of fixing feature values.

Since CES considers the conditional expectation $\mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$ when computing Shapley values, the resulting Shapley values are highly influenced by feature correlations. As a result, two highly correlated features may receive similar feature attributions even if the model under consideration depends on only one of them. We make this concrete with an example in Appendix D. We also demonstrate empirically in Section 5 and Appendix G that CES can be highly sensitive to the feature correlations, and consequently can lead to wrong explanations. Additionally, computing CES is computationally challenging when the feature-space is continuous. While Frye et al. [2021] propose training a surrogate model g with masked inputs to estimate the conditional expectation (see Appendix E), training g is even more difficult than training the model f.

Aside from this, the JBShap and RJBShap value functions proposed by Yeh et al. [2022], explain the feature contributions for the function $\tilde{f}_p(\mathbf{x}) \coloneqq f(\mathbf{x})p(\mathbf{x})$, rather than $f(\mathbf{x})$ itself. Specifically, RJBShap explain the contribution of individual features towards the difference $\tilde{f}_p(\mathbf{x}) - \mathbb{E}_{p_b(\mathbf{X})}[\tilde{f}_p(\mathbf{X})]$. This means that the resulting Shapley values therefore do not explain the underlying function f itself. We make this more concrete with an example with $\mathcal{X} \subseteq \mathbb{R}^2$:

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}, I_2), \quad f(\mathbf{x}) = \exp\left(x_1^2/2\right). \tag{2}$$

For this example, $\tilde{f}_p(\mathbf{x})$ only depends on x_2 and consequently, the RJBShap values for feature 1, $\phi_1 = 0$, for all $\mathbf{x} \in \mathcal{X}$, even though the function $f(\mathbf{x})$ *only* depends on x_1. RJBShap can therefore lead to *highly* misleading explanations. We confirm this empirically in Appendix G.2.4. Additionally, the notion of off-manifold robustness satisfied by JBShap and RJBShap value functions can be restrictive. We expand upon this in Section 3.1, where we propose an alternative definition of robustness which is less restrictive, and is not satisfied by JBShap and RJBShap.

3 MANIFOLD RESTRICTED SHAPLEY VALUES

In this paper, we argue that a model must be mainly characterised by it's behaviour on the data manifold. While *intervening* on features provides the correct notion of fixing features, we must restrict our attention to the data manifold when estimating Shapley values. This allows us to avoid the issues of non-identifiability outside the data manifold, thereby making the Shapley estimates robust against adversarial attacks as in Slack et al. [2020]. In order to estimate Shapley values which are robust to off-manifold manipulations, we must restrict the function evaluation to the data manifold. Before we proceed, we introduce our value function in terms of general sets $\mathcal{Z} \subseteq \mathcal{X}$.

Definition 1 (ManifoldShap)**.** *Let $\mathcal{Z} \subseteq \mathcal{X}$ be an open set with $\boldsymbol{x} \in \mathcal{Z}$, and $\mathbb{P}(\boldsymbol{X} \in \mathcal{Z} \mid do(\boldsymbol{X}_S = \boldsymbol{x}_S)) > 0$ for $S \subseteq [d]$. Then, we define the ManifoldShap on $\mathcal{Z}$ as:*

$$v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}(S) \coloneqq \mathbb{E}[f(\boldsymbol{X}) \mid do(\boldsymbol{X}_S = \boldsymbol{x}_S), \boldsymbol{X} \in \mathcal{Z}]. \tag{3}$$

Remark. The notation $\mathbb{E}[\cdot \mid do(\mathbf{X}_S = \mathbf{x}_S), \mathbf{X} \in \mathcal{Z}]$ denotes the expectation w.r.t. the density $p_{\mathcal{Z},\mathbf{x}_S}(\cdot)$ where

$$p_{\mathcal{Z},\mathbf{x}_S}(\mathbf{y}) \coloneqq \frac{p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S))\mathbb{1}(\mathbf{y} \in \mathcal{Z})}{\mathbb{P}(\mathbf{X} \in \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S))}. \tag{4}$$

The condition $\mathbb{P}(\mathbf{X} \in \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S)) > 0$ ensures that $p_{\mathcal{Z},\mathbf{x}_S}(\mathbf{x})$ (and hence $v^{\text{MAN}}_{\mathbf{x},f,\mathcal{Z}}(S)$) is well-defined. By conditioning on the event $\mathbf{X} \in \mathcal{Z}$, the ManifoldShap value function restricts the function evaluations to the set $\mathcal{Z}$. In

practice, $\mathcal{Z}$ can be chosen to be the data manifold, or any other region of interest, where model behaviour is relevant to explanations sought. In this way, ManifoldShap will disregard the model behaviour outside the region of interest when computing Shapley values. A detailed discussion of how to choose the sets $\mathcal{Z}$ is deferred to the next section.

Our formulation of *ManifoldShap* is general as it is does not assume a specific causal structure on the features. In our methodology, we assume that the expectation $\mathbb{E}[f(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S)]$ can be computed using observational data. This is a standard assumption needed to compute Interventional Shapley, and holds true under the causal structure in Figure 2. Under this assumption, we can compute the value function using the following result.

Lemma 1. *The value function $v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}$ can be written as,*

$$v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}(S) = \frac{\mathbb{E}[f(\boldsymbol{X})\mathbb{1}(\boldsymbol{X} \in \mathcal{Z}) \mid do(\boldsymbol{X}_S = \boldsymbol{x}_S)]}{\mathbb{P}(\boldsymbol{X} \in \mathcal{Z} \mid do(\boldsymbol{X}_S = \boldsymbol{x}_S))}$$

In practice, all we need is a manifold classifier, trained to estimate the value of the indicator, i.e. $\hat{g}(\mathbf{x}) \approx \mathbb{1}(\mathbf{x} \in \mathcal{Z})$. The value function (3) can then be estimated using:

$$v^{\text{MAN}}_{\mathbf{x},f,\mathcal{Z}}(S) \approx \frac{\mathbb{E}[f(\mathbf{X})\hat{g}(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S)]}{\mathbb{E}[\hat{g}(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S)]}. \tag{5}$$

We also provide alternative methodologies of estimating ManifoldShap using rejection sampling and regression techniques in Appendix C.

Choosing the sets $\mathcal{Z}$. Next, we discuss general purpose methodologies of choosing sets $\mathcal{Z}$ which can serve as practical estimation of the data manifold in most cases. One can obtain $\mathcal{Z}$ by training an out-of-distribution classifier directly. Slack et al. [2020] do so by perturbing each datapoint on randomly chosen features, and subsequently using these to train the classifier. In general, users may wish to choose different regions of interest $\mathcal{Z}$ on an ad hoc basis when computing Shapley values. In what follows, we outline a few specific choices of $\mathcal{Z}$, each of which satisfy different notions of robustness to off-manifold manipulations. We discuss this in greater length in Section 3.1.

Definition 2 (Density manifold)**.** *Given an $\epsilon > 0$, we define the* ϵ-density manifold *(ϵ-DM) of the data distribution, denoted as $\mathcal{D}_\epsilon$, as: $\mathcal{D}_\epsilon \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : p(\boldsymbol{x}) > \epsilon\}$. Here, $p(\boldsymbol{x})$ denotes the joint density of the data.*

The ϵ-DM includes all regions of high density in the set. Using $\mathcal{Z} = \mathcal{D}_\epsilon$ in our value function therefore restricts function evaluations to regions of high density. An alternative way to choose $\mathcal{Z}$ is via the probability mass captured by $\mathcal{Z}$, i.e., for a given level α, we may pick sets $\mathcal{Z} = \mathcal{P}_\alpha$ such that $\mathbb{P}(\mathbf{X} \in \mathcal{P}_\alpha) \geq \alpha$. One such set can be defined as:

Definition 3 (Mass manifold)**.** *Given an $\alpha > 0$, we define the* α-mass manifold *(α-MM) of the data distribution, denoted as $\mathcal{P}_\alpha$, as $\mathcal{P}_\alpha \coloneqq \mathcal{D}_{\epsilon^{(\alpha)}}$, where $\epsilon^{(\alpha)} \coloneqq \sup\{\epsilon \geq 0 : \mathbb{P}(\boldsymbol{X} \in \mathcal{D}_\epsilon) \geq \alpha\}$.*

We show in Proposition 9 (Appendix B) that the Lebesgue measure of $\mathcal{P}_\alpha$ is smallest among the sets $\mathcal{Z}$ with $\mathbb{P}(\mathbf{X} \in \mathcal{Z}) \geq \alpha$. It should be noted that $\mathcal{P}_\alpha$ is not necessarily the unique such set. One can use techniques like kernel density estimation and VAEs to approximate the manifolds described in this section (more details in Appendix F).

3.1 Robustness to off-manifold manipulation

We say that a Shapley value function is robust to off-manifold manipulation, if changing the model f outside the data manifold does not lead to 'large' changes in its Shapley values. In this section, we formalise this idea of robustness and show that ManifoldShap satisfies this notion, while the existing value functions do not. First, we present the definition of robustness as used in Yeh et al. [2022], to formalise the notion of off-manifold manipulations.

Definition 4 (T-robustness [Yeh et al., 2022])**.** *Given two models $f_1(\boldsymbol{x}), f_2(\boldsymbol{x})$ and any probability density $p(\boldsymbol{x})$, we say that a value function, $v_{\boldsymbol{x},f}$, is strong T-robust if it satisfies the following condition: if $\max_{\boldsymbol{x}} |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})|p(\boldsymbol{x}) \leq \delta$, then, $|v_{\boldsymbol{x},f_1}(S) - v_{\boldsymbol{x},f_2}(S)| \leq T\delta$ for any $S \subseteq [d]$.*

As per Yeh et al. [2022],"The premise $\max_{\mathbf{x}} |f_1(\mathbf{x}) - f_2(\mathbf{x})|p(\mathbf{x}) \leq \delta$ bounds the maximum perturbation on low density regions." Additionally, Yeh et al. [2022] show that JBShap and RJBShap value functions satisfy strong T-robustness to off-manifold manipulation, while other value functions like MS and CES do not. Likewise, since MS is a special case of IS, the latter also does not satisfy strong T-robustness. On the other hand, ManifoldShap restricted to ϵ-density manifold, $\mathcal{D}_\epsilon$, satisfies this notion of robustness.

Proposition 1. *The value function $v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{D}_\epsilon}(S) = \mathbb{E}[f(\boldsymbol{X}) \mid do(\boldsymbol{X}_S = \boldsymbol{x}_S), \boldsymbol{X} \in \mathcal{D}_\epsilon]$ is strong T-robust for $T = 1/\epsilon$.*

Proposition 1 shows that with decreasing ϵ, the robustness parameter T increases and ManifoldShap gets less robust.

Alternative definition of Robustness. Definition 4 considers a very specific notion of model perturbation. In particular, the perturbation in model $f(\mathbf{x})$ must not exceed $\delta/p(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^d$ and some $\delta > 0$. This does not encapsulate the case where the function perturbation remains bounded on a region of interest $\mathcal{Z}$, but may increase arbitrarily outside $\mathcal{Z}$. For example, we may have the case that the function $f(\mathbf{x})$ remains fixed on a set $\mathcal{Z}$ with $\mathbb{P}(\mathbf{X} \in \mathcal{Z}) > 0.99$. Robustness of Shapley values should dictate that changing the function outside $\mathcal{Z}$ should not lead to arbitrarily different Shapley values. We later show that Def. 4 does not lead to such robustness guarantees. To encapsulate this, we provide an alternative definition of robustness, which allows us to take into account model manipulation on sets with small probability mass. First, we define the notion of robustness on a general feature subspace $\mathcal{Z}' \subseteq \mathcal{X}$:

Definition 5 (Subspace T-robustness)**.** *Let $\mathcal{Z}' \subseteq \mathcal{X}$ be such*

that $\mathbb{P}(\boldsymbol{X} \in \mathcal{Z}') > 0$. *We say that a value function* $v_{\boldsymbol{x},f}$ *is strong T-robust on subspace* $\mathcal{Z}'$ *if it satisfies the following condition: if* $\sup_{\boldsymbol{x}\in\mathcal{Z}'} |f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})| \leq \delta$, *then,* $|v_{\boldsymbol{x},f_1}(S) - v_{\boldsymbol{x},f_2}(S)| \leq T\delta$ *for any* $S \subseteq [d]$.

A value function satisfying strong T-robustness on $\mathcal{Z}$ would not result in drastically different Shapley values when the model perturbation is bounded on the set $\mathcal{Z}$, by some value $\delta > 0$. The above definition allows us to directly consider robustness of value functions on sets based on probability mass, $\mathcal{P}_\alpha$. Moreover, by restricting the function evaluations to a set $\mathcal{Z}$, ManifoldShap is naturally set up to provide subspace T-robustness guarantee. We formalise this as follows:

Proposition 2. *The value function* $v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}$ *is strong T-robust on any set* $\mathcal{Z}'$ *satisfying* $\mathcal{Z} \subseteq \mathcal{Z}'$ *with* $T = 1$.

In contrast, we show that all other value functions under consideration do not satisfy this notion of robustness:

Proposition 3. *For any set* $\mathcal{Z}'$ *with* $\mathbb{P}(\boldsymbol{X} \in \mathcal{Z}') < 1$, *the IS value function* $v^{\text{IS}}_{\boldsymbol{x},f}(S)$, *the CES value function* $v^{\text{CES}}_{\boldsymbol{x},f}(S)$, *and the MS value function* $v^{\text{MS}}_{\boldsymbol{x},f}(S)$, *the JBShap value function* $v^{\text{J}}_{\boldsymbol{x},f}(S)$ *and the RJBShap value function* $v^{\text{RJ}}_{\boldsymbol{x},f}(S)$ *are all* not *strong T-robust on subspace* $\mathcal{Z}'$ *for* $|T| < \infty$.

Consider the family of value functions which *drop* features in $\bar{S}$ through randomisation, i.e., $v_{f,p_S}(S) = \mathbb{E}_{\mathbf{X}\sim p_S}[f(\mathbf{X})]$. We note that IS, MS, CES and ManifoldShap all fall into this family. For example, when $p_S = p(\mathbf{X} \mid do(\mathbf{X}_S = \mathbf{x}_S))$ we obtain IS, and when $p_S = p(\mathbf{X} \mid \mathbf{X}_S = \mathbf{x}_S)$ we obtain CES. We show in Appendix A.1 that the choice of p_S in ManifoldShap (i.e. $p_{\mathcal{Z},\mathbf{x}_S}$ in Eq. (4)) minimises the Total Variation distance with interventional distribution $p(\mathbf{X} \mid do(\mathbf{X}_S = \mathbf{x}_S))$ subject to the condition that $v_{f,p_S}(S)$ is strong T-robust on $\mathcal{Z}$. This ensures that ManifoldShap values provide reasonable estimation of *causal* contribution of features.

3.2 Comparison with existing methods

Causal Accuracy. Recall that, CES attributes feature importance based on feature correlations. Consequently, two highly correlated features may be attributed similar feature importance even if the model only depends on one of them, i.e., the sensitivity property is violated. However, ManifoldShap on the other-hand, seeks to estimate the *causal* contribution of features towards the prediction $f(\mathbf{x})$, as it uses the *interventional* measure restricted to the manifold $\mathcal{Z}$ to drop features. The experiments in Appendix G confirm this, as the ManifoldShap results are significantly less sensitive to feature correlations than CES.

Our example in Eq. (2) shows how the explicit dependence of RJBShap on the density can lead to extremely inaccurate Shapley explanations. In Appendix G.2.4, we show that because of its causal nature, ManifoldShap provides significantly more accurate and intuitive explanations. Additionally, unlike RJBShap, ManifoldShap only depends on the density estimation via the indicator $\mathbb{1}(p(\mathbf{x}) \geq \epsilon)$. Therefore, as we show in Appendix G.2.6, ManifoldShap is significantly more robust to density estimation errors than RJBShap.

Aside from this, Ghalebikesabi et al. [2021] propose Neighbourhood SHAP, a value function aimed to provide explanations for the localised behaviour of the model near the datapoint $\mathbf{x}$ where explanations are sought. While the authors empirically show the robustness of the methodology against off-manifold perturbations, they do not consider the causal perspective and therefore the main object of interest is not the causal contribution of features.

Robustness. As outlined in Section 3.1, ManifoldShap is robust to model changes outside the manifold and therefore is not vulnerable to adversarial attacks as in Slack et al. [2020]. In light of this, we argue that ManifoldShap provides a compromise between conditional and interventional Shapley values. It attempts to estimate causal contributions of features, while providing robustness guarantees.

Trade-off between Accuracy and Robustness. Restricting function evaluations to the manifold $\mathcal{Z}$, as in ManifoldShap, means that the resulting Shapley values are dependant on the manifold itself, and may not purely reflect the causal contribution of features. This is because these are no longer pure Interventional Shapley values. This results in a trade-off between robustness to off-manifold manipulation and the 'causal accuracy' of the Shapley values. ManifoldShap provides us flexibility over this trade-off, through the size of the manifold $\mathcal{Z}$. When $\mathcal{Z} = \mathcal{D}_\epsilon$, the size of the manifold is modulated through the ϵ parameter. As $\epsilon \to 0$, the size of manifold increases and ManifoldShap values tend towards IS values. However, as mentioned above, it comes at the cost of reduced robustness, as the Shapley evaluations include increasing number of datapoints 'far' from the training data. On the other hand, increasing ϵ increases the robustness of Shapley values, while reducing their causal accuracy, as the resulting Shapley values discard a significant number of datapoints which lie outside $\mathcal{D}_\epsilon$.

Computational Considerations. Computing CES may be computationally expensive and may require different supervised or unsupervised learning techniques [Frye et al., 2021, Sundararajan and Najmi, 2020, Yeh et al., 2022]. In contrast, while ManifoldShap requires estimating a manifold classifier, estimating $v^{\text{MAN}}_{\mathbf{x},f,\mathcal{Z}}(S)$ does not incur any computational cost over and above computing the interventional expectations. Proposition 1 illustrates this by expressing the ManifoldShap value function as a ratio of interventional expectations. This is even more straightforward when the causal structure is as in Figure 2, and the interventional expectation is equivalent to marginal expectation. Additionally, to avoid the exponential time complexity of computing the value function for all $S \subseteq [d]$, we propose a sampling based estimation in Appendix C.2 which

makes computation of ManifoldShap feasible for high dimensional feature spaces (see Appendix G.2.5).

4 ROBUSTNESS IN OTHER EXPLANATION METHODS

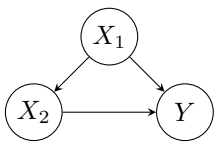
Shapley value is not the only *off-manifold* explanation method. This problem has also been explored in other explanation methods like LIME [Slack et al., 2020, Saito et al., 2020, Qiu et al., 2021] and gradient-based methods [Heo et al., 2019, Anders et al., 2020]. For example, Heo et al. [2019] illustrates this problem in gradient-based interpretability methods for Neural Networks. The paper shows that these explanations are not stable when model is manipulated without hurting the accuracy of the model. Numerous solutions have also been proposed such as Qiu et al. [2021], which addresses this problem for explanation methods like RISE, OCCLUSION and LIME by quantifying a similarity metric for perturbed data. This metric is then integrated into the explanation methods. Likewise Saito et al. [2020] proposes to make LIME robust to off-manifold manipulation, by using a GAN to sample more realistic synthetic data which are then used to generate LIME explanations. Aside from this, Anders et al. [2020] proposes an alternative robust gradient-based explanation method. However, unlike Shapley values, gradient-based methods rely on model properties (e.g., differentiability), and are not model agnostic.

5 EXPERIMENTAL RESULTS

In this section, we conduct experiments on synthetic and real world datasets to demonstrate the utility of ManifoldShap and compare it with existing methods. Instead of training the models, we compute Shapley values for the underlying true functions directly. Additional experiments investigating the sensitivity of the different Shapley methods to changing feature correlations, manifold size and feature dimensions have been included in Appendix G. The code to reproduce our experiments can be found at github.com/amazon-science/manifold-restricted-shapley.

5.1 Synthetic data experiments

Here we investigate the effect of model perturbation in low density regions on Shapley values.



Data generating mechanism. In this experiment, $\mathcal{Y} \subseteq \mathbb{R}$ and $\mathcal{X} \subseteq \mathbb{R}^2$ follow the Causal DAG shown on the left. In specific, the Structural Causal Model (SCM) [Pearl, 2000] for the ground truth data generating mechanism is:

$$X_1 = \epsilon_1, \quad X_2 = \rho X_1 + \sqrt{1-\rho^2}\epsilon_2, \quad Y = X_1.$$

Here, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ and $\rho = 0.85$ is the correlation between X_1, X_2. Next, we define the perturbed models.

Perturbed models. We define the following family of perturbed models $g_\delta : \mathcal{X} \to \mathbb{R}$, parameterised by $\delta \in \mathbb{R}$.

$$g_\delta(\mathbf{X}) \coloneqq Y + \delta X_2 \mathbb{1}(\mathbf{X} \notin \mathcal{P}_\alpha).$$

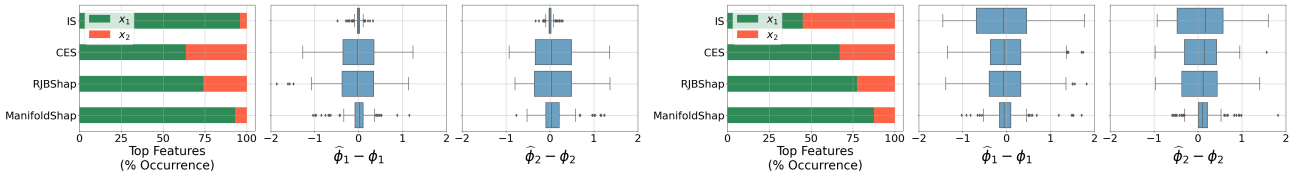
Here, we use VAEs to estimate $\mathcal{P}_\alpha$ (see Appendix F) and choose $\alpha = 1 - 10^{-3}$. By construction, the models g_δ should agree with the ground truth on the α-manifold, i.e. $g_\delta(\mathbf{X}) = Y$ when $\mathbf{X} \in \mathcal{P}_\alpha$, but these models differ from the ground truth for $\mathbf{X} \notin \mathcal{P}_\alpha$. Figure 4 shows the model heatmaps for $\delta = 0, 5$ along with the original data. It is impossible to distinguish between these models on the data manifold, as both have test mean squared error of 0.

Results. Recall that the ground truth model does not depend on X_2, so the ground truth Shapley value for feature 2 is $\phi_2 = 0$. As a result, for any prediction, feature 1 has greater absolute Shapley value than feature 2, i.e. $|\phi_1| \geq |\phi_2|$. We compute Shapley values for g_δ using different value functions on 500 datapoints $\{\mathbf{x}^{(i)}\}_{i=1}^{500}$, sampled from the SCM defined above. We compute CES using the ground truth conditional distributions of $X_i \mid X_j$ for $i \neq j$, which can be obtained analytically in this setting. Figure 3 shows the results, with the bar plots on the left of Figures 3a and 3b, showing the most important features as per different value functions for $\delta = 0, 5$.

For $\delta = 0$, Figure 3a confirms that the IS values of the ground truth model attribute greatest feature importance to feature 1 for all datapoints. This is expected as the ground truth model does not depend on x_2. For ManifoldShap, we observe that for 4% of the datapoints, feature 2 is attributed greater importance. This highlights that robustness of ManifoldShap comes at the cost of reduced causal accuracy of Shapley values. Furthermore, it can be seen that CES value function attributes greatest importance to feature 2 for more than 30% of the datapoints. This is because CES provides similar Shapley values for positively correlated features. We observe similar behaviour for RJBShap, which attributes greatest importance to feature 2 for about 20% of datapoints. This happens because RJBShap provides feature contributions for $\tilde{f}_p(\mathbf{x}) = f(\mathbf{x})p(\mathbf{x})$ rather than $f(\mathbf{x})$, and can therefore be misleading.

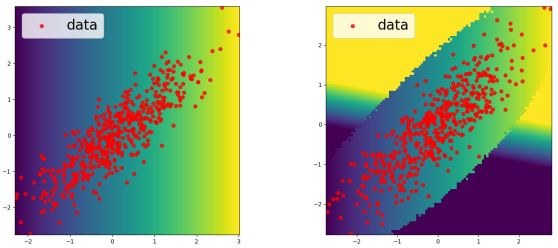
When $\delta = 5$, Figure 3b shows that, for more than 50% of datapoints IS attributes greater importance to feature 2 than feature 1 in the perturbed model. This shows that IS is sensitive to off-manifold perturbation. For ManifoldShap on the other hand, feature 2 is attributed greater importance for only about 10% of the datapoints, less than all other baselines.

We have also plotted the difference between estimated Shapley values and the ground truth IS values, for each value function. For a fair comparison between differ-



(a) $\delta = 0$ (b) $\delta = 5$

Figure 3: Synthetic data experiments for $\delta = 0, 5$. The barplots on the left of each subfigure shows the most important features for different Shapley value functions. The boxplots show the approximation errors of the Shapley values for different value functions.



(a) Heatmap of g_δ for $\delta = 0$. (b) Heatmap of g_δ for $\delta = 5$.

Figure 4: Heatmaps for ground truth and perturbed models g_δ. Each model has test mean squared error of 0.

ent value functions, we scale the Shapley values so that $\sum_{i\in\{1,2\}} |\phi_i| = 1$. As δ increases from 0 to 5, we can see that the errors in Shapley values increase for IS, while the errors in ManifoldShap are more concentrated around 0 than any other baseline.

The results show that ManifoldShap values, unlike IS, remain robust to off-manifold manipulations, while providing explanations which remain closer to ground truth IS values overall. CES and RJBShap, on the other hand can result in misleading explanations.

5.2 Real world datasets

In this subsection, we evaluate the effect of adversarial off-manifold manipulation of models on Shapley values using real-world datasets. Specifically, using the same setup as in Slack et al. [2020], we show that existing methodologies may fail to identify highly problematic model biases, whereas ManifoldShap can mitigate this problem due to its robustness properties. We consider the causal structure in Figure 2 where the true features $\tilde{X}_i$ are distinguished from input features X_i, and therefore IS is equivalent to MS here.

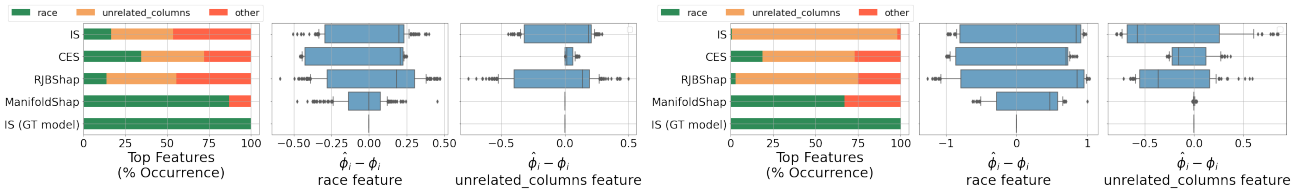
Datasets. The COMPAS dataset, collected by ProPublica [Angwin et al., 2016], includes information for 6172 defendants from Broward County, Florida. This information comprises 52 features including defendants' criminal history and demographic attributes. The sensitive attribute in this dataset is defendants' race. The second dataset, Communities and Crime (CC), is a UCI dataset [Dua and Graff, 2017] which includes crime data in communities across the US, where each community constitutes a datapoint comprising 128 features. The sensitive attribute in CC is the percentage of Caucasian population. From here onwards, we use 'race' to refer to the sensitive attribute for both datasets.

Biased classifier. Following the strategy of Slack et al. [2020], we construct the binary classifier f to be only dependant on the sensitive feature for both datasets. Additional details are given in Appendix G.1.

Manifold estimation. Just like in Slack et al. [2020], we determine the manifold $\mathcal{Z}$ by training an OOD classifier. In particular, we follow the strategy in Slack et al. [2020] by perturbing each datapoint on randomly chosen features, and subsequently using these newly generated perturbations to train an OOD classifier.

Out of manifold perturbation. To perturb the model outside the manifold $\mathcal{Z}$, we construct 2 synthetic features (referred to as 'unrelated columns') like Slack et al. [2020]. For datapoints that lie outside $\mathcal{Z}$, only the 'unrelated columns' are used to classify the datapoints. However, unlike Slack et al. [2020], these 'unrelated columns' are positively correlated with race. This is done to highlight a shortcoming of CES: even though CES is an on-manifold value function, the positive correlation between unrelated columns and race 'fools' CES into attributing non-zero credit to the synthetic features.

Results. We compute the Shapley values for the perturbed models on 500 datapoints from a randomly chosen held-out dataset. We use the supervised approach to estimate CES as outlined in Appendix E. The barplots in Figures 5a and 5b show the percentage of data points in COMPAS and CC datasets respectively, for which each feature shows up as the top feature as per different value functions. For RJBShap, CES, and IS, there are more data points in both datasets with top feature among 'unrelated columns' than data points with top feature of race. For IS, this happens as a result of OOD perturbation of the model, and shows that when using IS, we can hide biases in the model



(a) COMPAS dataset results

(b) Communities and Crime dataset results

Figure 5: Experiments on COMPAS and CC datasets. The barplots on the left of each subfigure shows the most important features for different Shapley values functions. The boxplots show the approximation errors of the Shapley values for different value functions.

by perturbing the model out of manifold. For RJBShap, this could be explained by the fact that it explicitly depends on the joint density $p(\mathbf{x})$ of the data. Since, 'unrelated columns' are positively correlated with race, the dependence of the density $p(\mathbf{x})$ on these features and race is similar. As a result, 'unrelated columns' get non-zero attributions in RJBShap.

This positive correlation between race and 'unrelated columns' also causes CES to attribute similar importance for features 'unrelated columns' as for race. This can be especially misleading when the data contains multiple correlated features which are not used by the model.

On the other hand, for ManifoldShap, majority of the datapoints have top feature race, whereas none of them have top feature among 'unrelated columns'. Figure 5 also shows the difference between estimated Shapley values and the ground truth IS values of the biased model. We have again rescaled the Shapley values so that $\sum_{i\in[d]} |\phi_i| = 1$ for fair comparison between different value functions. We can see that for the feature race, the errors of ManifoldShap are more concentrated around 0 than any other baseline considered. For 'unrelated columns', ManifoldShap values are $\hat{\phi}_i = \phi_i = 0$, i.e., ManifoldShap satisfies sensitivity property in this case. This shows that ManifoldShap is significantly more robust to adversarial manipulation of the function outside the manifold, as well as robust to the attribution of credit based on correlations among features.

6 DISCUSSION AND LIMITATIONS

In this paper, we propose ManifoldShap, a Shapley value function which provides a compromise between existing on and off manifold value functions, by providing explanations which are robust to off-manifold perturbations of the model while estimating the causal contribution of features. However, ManifoldShap also has its limitations.

While our work does not make any assumptions on the set $\mathcal{Z}$, the properties of ManifoldShap are inherently linked to the choice of $\mathcal{Z}$. ManifoldShap is only robust to perturbation of model outside $\mathcal{Z}$ and perturbations inside $\mathcal{Z}$ could lead to significant changes in the computed Shapley values. It is therefore important to choose $\mathcal{Z}$ that is a good representative of the true data manifold, as otherwise, the Shapley values may not be robust to off-manifold perturbations. Additionally, as pointed out in Section 3.2, restricting model evaluations to the set $\mathcal{Z}$ can reduce the causal accuracy of Shapley values. This becomes especially evident when the data manifold $\mathcal{Z}$ is *sparse* or low-dimensional relative to the space $\mathcal{X}$. We highlight this empirically in Appendix G.2.2. Likewise, as we show in Appendix A, the sensitivity and symmetry properties of ManifoldShap are also dependent on the properties of $\mathcal{Z}$. It is therefore worth exploring methodologies of choosing $\mathcal{Z}$ which provide the ideal trade-off between desirable properties like causal accuracy and robustness of explanations. We believe these limitations suggest interesting research questions that we leave for future work.

Acknowledgements

We would like to thank Dominik Janzing for his valuable suggestions and insightful discussions. We are also grateful to Kailash Budhathoki and Philipp Faller for providing feedback on an earlier version of the manuscript.

References

Muhammad Aurangzeb Ahmad, Ankur Teredesai, and Carly Eckert. Interpretable machine learning in healthcare. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 447–447, 2018. doi: 10.1109/ICHI.2018.00095.

Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 314–323. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/anders20a.html`.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of Data and Analytics*, pages 254–264. Auerbach Publications, 2016.

Jesús Bilbao. *Cooperative Games on Combinatorial Structures*. 01 2000. ISBN 978-1-4613-6976-9. doi: 10.1007/978-1-4615-4393-0.

Benoît Cadre, Bruno Pelletier, and Pierre Pudlo. Estimation of density level sets with a given probability content. *Journal of Nonparametric Statistics*, 25(1): 261–272, 2013. doi: 10.1080/10485252.2012.750319. URL https://doi.org/10.1080/10485252.2012.750319.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. Shapley explainability on the data manifold. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OPyWRrcjVQw.

Edward Gan and Peter Bailis. Scalable kernel density classification via threshold-based pruning. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 945–959, 2017.

Sahra Ghalebikesabi, Lucile Ter-Minassian, Karla DiazOrdaz, and Christopher C. Holmes. On locality of local explanation models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=6OkPFFMgBt.

Hajar Hakkoum, Ibtissam Abnane, and Ali Idri. Interpretability in the medical field: A systematic mapping and review study. *Applied Soft Computing*, 117:108391, 2022. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2021.108391. URL https://www.sciencedirect.com/science/article/pii/S1568494621011522.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/7fea637fd6d02b8f0adf6f7dc36aed93-Paper.pdf.

Tom Heskes, Ioan Gabriel Bucur, Evi Sijben, and Tom Claassen. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Dominik Janzing, Lenon Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pages 2907–2916. PMLR, 2020.

Yonghan Jung, Shiva Kasiviswanathan, Jin Tian, Dominik Janzing, Patrick Bloebaum, and Elias Bareinboim. On measuring causal contributions via do-interventions. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10476–10501. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jung22a.html.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2013.

Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.

Håvard Kvamme, Nikolai Sellereite, Kjersti Aas, and Steffen Sjursen. Predicting mortgage default using convolutional neural networks. *Expert Systems with Applications*, 102:207–217, 2018. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2018.02.029. URL https://www.sciencedirect.com/science/article/pii/S0957417418301179.

Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein Tajmir, Claude Guerrier, Sarah Ebert, Stuart Pomerantz, Javier Romero, Shahmir Kamalian, Ramon Gonzalez, Michael Lev, and Synho Do. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature Biomedical Engineering*, 3, 03 2019. doi: 10.1038/s41551-018-0324-9.

Pantelis Linardatos, Vasilis Papastefanopoulos, and Sotiris Kotsiantis. Explainable ai: A review of machine learning interpretability methods. *Entropy*, 23(1), 2021. ISSN 1099-4300. doi: 10.3390/e23010018. URL https://www.mdpi.com/1099-4300/23/1/18.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, USA, 2000. ISBN 0521773628.

Judea Pearl. The do-calculus revisited. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Arti-*

ficial Intelligence, UAI'12, page 3–11, Arlington, Virginia, USA, 2012. AUAI Press. ISBN 9780974903989.

Luyu Qiu, Yi Yang, Caleb Chen Cao, Jing Liu, Yueyuan Zheng, Hilary Hei Ting Ngai, Janet H. Hsiao, and Lei Chen. Resisting out-of-distribution data problem in perturbation of XAI. *CoRR*, abs/2107.14000, 2021. URL https://arxiv.org/abs/2107.14000.

Protection Regulation. Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, 679:2016, 2016.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

Sean Saito, Eugene Chua, Nicholas Capel, and Rocco Hu. Improving LIME robustness with smarter locality sampling. *CoRR*, abs/2006.12302, 2020. URL https://arxiv.org/abs/2006.12302.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 180–186, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371100. doi: 10.1145/3375627.3375830. URL https://doi.org/10.1145/3375627.3375830.

Erik Strumbelj and Igor Kononenko. An efficient explanation of individual classifications using game theory. *Journal of Machine Learning Research*, 11:1–18, 01 2010. doi: 10.1145/1756006.1756007.

Erik Strumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, 41(3):647–665, dec 2014. ISSN 0219-1377. doi: 10.1007/s10115-013-0679-x. URL https://doi.org/10.1007/s10115-013-0679-x.

Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/sundararajan20b.html.

Chih-Kuan Yeh, Kuan-Yun Lee, Frederick Liu, and Pradeep Ravikumar. Threading the needle of on and off-manifold value functions for shapley explanations. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 1485–1502. PMLR, 28–30 Mar 2022. URL https://proceedings.mlr.press/v151/yeh22a.html.

Appendices

A PROPERTIES OF MANIFOLDSHAP

In this section, we consider the theoretical properties of ManifoldShap. The proofs for results in this section are provided in Section B.

A.0.1 Sensitivity Property

The following result holds in the setting of Janzing et al. [2020], i.e., when the real feature values are formally distinguished from the feature values input into the function (see Figure 2). In this case, the interventional distribution of $p(\mathbf{X}_{\bar{S}} \mid do(\mathbf{X}_S = \mathbf{x}_S))$ is the same as the marginal distribution $p(\mathbf{X}_{\bar{S}})$.

Proposition 4 (Sensitivity)**.** *Let $i \in [d]$ be such that*

1. *the function $f(\boldsymbol{x})$ does not depend on x_i for all values of $\boldsymbol{x}$,*
2. *the set $\mathcal{Z}$ is of the form, $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_d$, where $\mathcal{Z}_j \subseteq \mathbb{R}$ and $X_i \in \mathcal{Z}_i$ almost surely.*

Then, if the causal graph of features is as shown in Figure 2, we have that $v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}(S) = v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}(S \cup \{i\})$, and therefore $\phi_i = 0$.

Remark. As mentioned previously, in this paper we argue that a function should mainly be characterised by it's behaviour on manifold. Note that in this case, unlike the classical formulation of Sensitivity axiom [Sundararajan and Najmi, 2020], we also need the condition 2 above, which implies that $\mathbb{1}(\mathbf{X} \in \mathcal{Z})$ is independent of X_i. This can be justified as follows: Define a function $h(\mathbf{x}) \coloneqq \mathbb{1}(\mathbf{x} \in \mathcal{Z}) f(\mathbf{x})$. If condition 2 does not hold, i.e., $\mathbb{1}(\mathbf{X} \in \mathcal{Z})$ depends on X_i, then $h(\mathbf{X})$ must depend on X_i. Moreover, by definition, $h(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{Z}$, i.e. h and f agree on $\mathcal{Z}$. Therefore, since f and h are indistinguishable on the $\mathcal{Z}$, and $h(\mathbf{X})$ depends on X_i, it would be misleading to have zero attribution for the i'th feature.

A.0.2 Symmetry Property

Like the previous result, the following result holds in the setting of Janzing et al. [2020], where Interventional Shapley is equivalent to Marginal Shapley.

Proposition 5 (Symmetry)**.** *Let $i, j \in [d]$ be such that*

1. *the function $f(\boldsymbol{x})$ is symmetric in components i and j on $\mathcal{Z}$,*
2. *the density $p(\boldsymbol{x})$ is symmetric in components i and j,*
3. *the function $\mathbb{1}(\boldsymbol{x} \in \mathcal{Z})$ is symmetric in components i and j.*

Then, if the causal graph of features is as shown in Figure 2, we have that for any $S \subseteq [d] \setminus \{i, j\}$ and $\boldsymbol{x}$ such that $x_i = x_j$, we have that $v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}(S \cup \{i\}) = v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}(S \cup \{j\})$, and therefore $\phi_i = \phi_j$.

Remark. The condition 3 above states that $\mathcal{Z}$ should be symmetric in components i and j. This condition will be satisfied if, for example, $\mathcal{Z}$ is a ball centred at origin. Moreover, if condition 2 is satisfied, i.e., the density $p(\mathbf{x})$ is symmetric in x_i and x_j, it is straightforward to show that the ϵ-density manifold will also satisfy condition 3, for any $\epsilon > 0$. Additionally, we emphasise that condition 2 is not specific to our value function, and is also needed for symmetry property to hold for CES and IS. Janzing et al. [2020] illustrates this with an example where symmetry fails to hold for both CES and IS without condition 2.

A.0.3 Efficiency Property

All results presented in this section from here onwards do not assume a specific causal structure on the features and hold for any general causal graph on features.

Proposition 6 (Efficiency)**.** *The value function $v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}(S) = \mathbb{E}[f(\boldsymbol{X}) \mid do(\boldsymbol{X}_S = \boldsymbol{x}_S), \boldsymbol{X} \in \mathcal{Z}]$ satisfies*

$$v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}([d]) - v^{\text{MAN}}_{\boldsymbol{x},f,\mathcal{Z}}(\emptyset) = f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{Z}],$$

for any $\boldsymbol{x} \in \mathcal{Z}$. Therefore, $\sum_i \phi_i = f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{X}) \mid \boldsymbol{X} \in \mathcal{Z}]$.

Proof. Follows straightforwardly from the definition of $v^{\text{MAN}}_{\mathbf{x},f,\mathcal{Z}}(S)$. □

A.0.4 Linearity Property

Proposition 7 (Linearity). *For any functions f_1, f_2 and $\alpha_1, \alpha_2 \in \mathbb{R}$,*

$$v^{\text{MAN}}_{\boldsymbol{x},\alpha_1 f_1+\alpha_2 f_2,\mathcal{Z}}(S) = \alpha_1 v^{\text{MAN}}_{\boldsymbol{x},f_1,\mathcal{Z}}(S) + \alpha_2 v^{\text{MAN}}_{\boldsymbol{x},f_2,\mathcal{Z}}(S)$$

Therefore, $\phi_i^{\alpha_1 f_1+\alpha_2 f_2} = \alpha_1 \phi_i^{f_1} + \alpha_2 \phi_i^{f_2}$, where ϕ_i^f denotes the Shapley value for feature i and function f.

Proof. Follows straightforwardly from the definition of $v^{\text{MAN}}_{\mathbf{x},f,\mathcal{Z}}(S)$ and the linearity of the expectation $\mathbb{E}[f(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S), \mathbf{X} \in \mathcal{Z}]$. □

A.1 Robustness and Causal Accuracy of ManifoldShap

We consider the family of value functions of the form $v_{f,p_S}(S) = \mathbb{E}_{\mathbf{X}\sim p_S}[f(\mathbf{X})]$ for some measure p_S. Both, Interventional Shapley and ManifoldShap are part of this family with corresponding densities $p^{\text{do}}_{\mathbf{x}_S}(\mathbf{y}) \coloneqq p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S))$ and $p_{\mathcal{Z},\mathbf{x}_S}(\mathbf{y}) \coloneqq \frac{p(\mathbf{y}|do(\mathbf{X}_S=\mathbf{x}_S))\mathbb{1}(\mathbf{y}\in\mathcal{Z})}{\mathbb{P}(\mathbf{X}\in\mathcal{Z}|do(\mathbf{X}_S=\mathbf{x}_S))}$ respectively. Next, we show that $p_{\mathcal{Z},\mathbf{x}_S}$ minimises the total variation distance with the interventional distribution $p^{\text{do}}_{\mathbf{x}_S}$ while satisfying subspace robustness in definition 5.

Proposition 8. *The measure $p_{\mathcal{Z},\boldsymbol{x}_S}$ satisfies*

$$p_{\mathcal{Z},\boldsymbol{x}_S} \in \arg\min_{p_S}\{\text{TV}(p_S, p^{\text{do}}_{\boldsymbol{x}_S}) : v_{f,p_S} \text{ is strong T-robust on subspace } \mathcal{Z}\}$$

Proof. If v_{f,p_S} is strong T-robust on subspace $\mathcal{Z}$, then consider functions f_1, f_2 satisfying

$$f_2(\mathbf{x}) \coloneqq f_1(\mathbf{x}) + \delta\mathbb{1}(\mathbf{x} \in \mathcal{Z}) + K\mathbb{1}(\mathbf{x} \notin \mathcal{Z})$$

for some $\delta, K > 0$. Then, $\max_{\mathbf{x}\in\mathcal{Z}} |f_1(\mathbf{x}) - f_2(\mathbf{x})| = \delta$. Moreover,

$$\begin{aligned}|v_{f_1,p_S}(S) - v_{f_2,p_S}(S)| =& |\mathbb{E}_{\mathbf{X}\sim p_S}[f_1(\mathbf{X})] - \mathbb{E}_{\mathbf{X}\sim p_S}[f_2(\mathbf{X})]| \\ =& \delta p_S(\mathbf{X} \in Z) + K p_S(\mathbf{X} \notin Z)\end{aligned}$$

Since we can pick K to be arbitrarily large, v_{f,p_S} satisfies strong T-robustness on subspace $\mathcal{Z}$ only if $p_S(\mathbf{X} \notin Z) = 0$.

Next, note that if $p_S(\mathbf{X} \in \mathcal{Z}) = 1$,

$$\begin{aligned}&\text{TV}(p_S, p^{\text{do}}_{\mathbf{x}_S}) \\ =&1/2 \int_{\mathbf{y}} |p_S(\mathbf{y}) - p^{\text{do}}_{\mathbf{x}_S}(\mathbf{y})| \mathrm{d}\mathbf{y} \\ =&1/2 \int_{\mathbf{y}} |p_S(\mathbf{y}) - p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S))| \mathrm{d}\mathbf{y} \\ =&1/2 \int_{\mathbf{y}\in\mathcal{Z}} |p_S(\mathbf{y}) - p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S))| \mathrm{d}\mathbf{y} + 1/2 \int_{\mathbf{y}\notin\mathcal{Z}} |p_S(\mathbf{y}) - p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S))| \mathrm{d}\mathbf{y} \\ \geq&1/2 \left| \int_{\mathbf{y}\in\mathcal{Z}} p_S(\mathbf{y}) - p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S)) \mathrm{d}\mathbf{y} \right| + 1/2 \left| \int_{\mathbf{y}\notin\mathcal{Z}} p_S(\mathbf{y}) - p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S)) \mathrm{d}\mathbf{y} \right| \\ =&1/2 \, |p_S(\mathbf{X} \in \mathcal{Z}) - \mathbb{P}(\mathbf{X} \in \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S))| + 1/2 |p_S(\mathbf{X} \notin \mathcal{Z}) - \mathbb{P}(\mathbf{X} \notin \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S))| \\ =&1/2 \, (1 - \mathbb{P}(\mathbf{X} \in \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S))) + 1/2 \mathbb{P}(\mathbf{X} \notin \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S)) \\ =&1/2 \int_{\mathbf{y}\in\mathcal{Z}} \left| \frac{p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S))\mathbb{1}(\mathbf{y} \in \mathcal{Z})}{\mathbb{P}(\mathbf{X} \in \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S))} - p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S)) \right| \mathrm{d}\mathbf{y} + 1/2 \mathbb{P}(\mathbf{X} \notin \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S)) \\ =&1/2 \int_{\mathbf{y}} \left| \frac{p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S))\mathbb{1}(\mathbf{y} \in \mathcal{Z})}{\mathbb{P}(\mathbf{X} \in \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S))} - p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S)) \right| \mathrm{d}\mathbf{y} \\ =&\text{TV}(p_{\mathcal{Z},\mathbf{x}_S}, p^{\text{do}}_{\mathbf{x}_S})\end{aligned}$$

□

Proposition 8 shows that among all the value functions of the form v_{f,p_S} which are strong T-robust on subspace $\mathcal{Z}$, ManifoldShap provides the *best* approximation to Interventional Shapley values. This further highlights that ManifoldShap provides a compromise between on and off manifold value functions – it satisfies subspace robustness while also approximating causal contribution of features.

B PROOFS

Proof of Lemma 1.

Proof. Using the definition of ManifoldShap, we get that

$$\begin{aligned}
v^{\text{MAN}}_{\mathbf{x},f,\mathcal{Z}}(S) =& \int_{\mathbf{y}} f(\mathbf{y}) p_{\mathcal{Z},\mathbf{x}_S}(\mathbf{y}) \mathrm{d}\mathbf{y} \\
=& \int_{\mathbf{y}} f(\mathbf{y}) \frac{p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S)) \mathbb{1}(\mathbf{y} \in \mathcal{Z})}{\mathbb{P}(\mathbf{X} \in \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S))} \mathrm{d}\mathbf{y} \\
=& \frac{1}{\mathbb{P}(\mathbf{X} \in \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S))} \int_{\mathbf{y}} f(\mathbf{y}) \mathbb{1}(\mathbf{y} \in \mathcal{Z}) p(\mathbf{y} \mid do(\mathbf{X}_S = \mathbf{x}_S)) \mathrm{d}\mathbf{y} \\
=& \frac{\mathbb{E}[f(\mathbf{X}) \mathbb{1}(\mathbf{X} \in \mathcal{Z}) \mid do(\mathbf{X}_S = \mathbf{x}_S)]}{\mathbb{P}(\mathbf{X} \in \mathcal{Z} \mid do(\mathbf{X}_S = \mathbf{x}_S))}.
\end{aligned}$$

□

Proof of Proposition 1.

Proof.

$$\begin{aligned}
&\max_{\mathbf{x}} |f_1(\mathbf{x}) - f_2(\mathbf{x})| p(\mathbf{x}) \leq \delta \\
\implies &\sup_{\mathbf{x} \in \mathcal{D}_\epsilon} |f_1(\mathbf{x}) - f_2(\mathbf{x})| p(\mathbf{x}) \leq \delta \\
\implies &\sup_{\mathbf{x} \in \mathcal{D}_\epsilon} |f_1(\mathbf{x}) - f_2(\mathbf{x})| \epsilon \leq \sup_{\mathbf{x} \in \mathcal{D}_\epsilon} |f_1(\mathbf{x}) - f_2(\mathbf{x})| p(\mathbf{x}) \leq \delta \\
\implies &\sup_{\mathbf{x} \in \mathcal{D}_\epsilon} |f_1(\mathbf{x}) - f_2(\mathbf{x})| \leq \delta/\epsilon
\end{aligned}$$

Using the above,

$$\begin{aligned}
|v^{\text{MAN}}_{\mathbf{x},f_1,\mathcal{D}_\epsilon}(S) - v^{\text{MAN}}_{\mathbf{x},f_2,\mathcal{D}_\epsilon}(S)| =& |\mathbb{E}[f_1(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S), \mathbf{X} \in \mathcal{D}_\epsilon] - \mathbb{E}[f_2(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S), \mathbf{X} \in \mathcal{D}_\epsilon]| \\
\leq& \mathbb{E}[|f_1(\mathbf{X}) - f_2(\mathbf{X})| \mid do(\mathbf{X}_S = \mathbf{x}_S), \mathbf{X} \in \mathcal{D}_\epsilon] \\
\leq& \sup_{\mathbf{x} \in \mathcal{D}_\epsilon} |f_1(\mathbf{x}) - f_2(\mathbf{x})| \leq \delta/\epsilon
\end{aligned}$$

□

Proof of Proposition 2.

Proof. Let $\sup_{\mathbf{x} \in \mathcal{Z}'} |f_1(\mathbf{x}) - f_2(\mathbf{x})| \leq \delta$. Then, for any $S \subseteq [d]$,

$$\begin{aligned}
|v^{\text{MAN}}_{\mathbf{x},f_1,\mathcal{Z}}(S) - v^{\text{MAN}}_{\mathbf{x},f_2,\mathcal{Z}}(S)| =& |\mathbb{E}[f_1(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S), \mathbf{X} \in \mathcal{Z}] - \mathbb{E}[f_2(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S), \mathbf{X} \in \mathcal{Z}]| \\
\leq& \mathbb{E}[|f_1(\mathbf{X}) - f_2(\mathbf{X})| \mid do(\mathbf{X}_S = \mathbf{x}_S), \mathbf{X} \in \mathcal{Z}] \\
\leq& \sup_{\mathbf{x} \in \mathcal{Z}} |f_1(\mathbf{x}) - f_2(\mathbf{x})| \\
\leq& \sup_{\mathbf{x} \in \mathcal{Z}'} |f_1(\mathbf{x}) - f_2(\mathbf{x})| \leq \delta.
\end{aligned}$$

□

Proof of Proposition 3.

Proof. Let $S = \emptyset$, then $v^{\text{IS}}_{\mathbf{x},f_1}(S) = v^{\text{CES}}_{\mathbf{x},f_1}(S) = v^{\text{MS}}_{\mathbf{x},f_1}(S) = \mathbb{E}[f(\mathbf{X})]$. Let $f_2(\mathbf{x}) \coloneqq f_1(\mathbf{x}) + K\mathbb{1}(\mathbf{x} \notin \mathcal{Z}')$ for some $K > 0$. Then, we have that $\sup_{\mathbf{x} \in \mathcal{Z}'} |f_1(\mathbf{x}) - f_2(\mathbf{x})| = 0$. Moreover,

$$|\mathbb{E}[f_1(\mathbf{X})] - \mathbb{E}[f_2(\mathbf{X})]| = |K\mathbb{E}[\mathbb{1}(\mathbf{X} \notin \mathcal{Z}')]| = K\mathbb{P}(\mathbf{X} \notin \mathcal{Z}') > 0.$$

Since we can choose K to be arbitrarily big, it follows that $|\mathbb{E}[f_1(\mathbf{X})] - \mathbb{E}[f_2(\mathbf{X})]|$ is not bounded for general functions f_1, f_2 satisfying $\sup_{\mathbf{x}\in\mathcal{Z}'} |f_1(\mathbf{x}) - f_2(\mathbf{x})| \leq \delta$.

Now, for JBShap and RJBShap, define f_1, f_2 such that $f_2(\mathbf{x}) \coloneqq f_1(\mathbf{x}) + K\mathbb{1}(\mathbf{x} \notin \mathcal{Z}', p(\mathbf{x}) > 0)/p(\mathbf{x})$. Then,

$$\sup_{\mathbf{x}\in\mathcal{Z}'} |f_1(\mathbf{x}) - f_2(\mathbf{x})| = 0.$$

Let $\mathbf{x} \in \mathbb{R}^d$ be such that $\mathbf{x} \notin \mathcal{Z}'$ and $p(\mathbf{x}) > 0$. Since $\mathbb{P}(\mathbf{X} \in \mathcal{Z}') < 1$, there must exist an $\mathbf{x} \in \mathbb{R}^d$ which satisfies this condition. Then for $S = \emptyset$,

$$\begin{aligned}|v^{\text{J}}_{\mathbf{x},f_1,p}(S) - v^{\text{J}}_{\mathbf{x},f_2,p}(S)| =& |f_1(\mathbf{x})p(\mathbf{x}) - f_2(\mathbf{x})p(\mathbf{x})| \\ =& K|\mathbb{1}(\mathbf{x} \notin \mathcal{Z}', p(\mathbf{x}) > 0)| = K.\end{aligned}$$

Since we can choose K to be arbitrarily big, it follows that $|v^{\text{J}}_{\mathbf{x},f_1,p}(\emptyset) - v^{\text{J}}_{\mathbf{x},f_2,p}(\emptyset)|$ is not bounded for general functions f_1, f_2 satisfying $\sup_{\mathbf{x}\in\mathcal{Z}'} |f_1(\mathbf{x}) - f_2(\mathbf{x})| \leq \delta$.

Moreover, we have that for $S = \emptyset$,

$$\begin{aligned}|v^{\text{RJ}}_{\mathbf{x},f_1,p}(S) - v^{\text{RJ}}_{\mathbf{x},f_2,p}(S)| =& |\mathbb{E}[f_1(\mathbf{X})p(\mathbf{X})] - \mathbb{E}[f_2(\mathbf{X})p(\mathbf{X})]| \\ =& K|\mathbb{E}[\mathbb{1}(\mathbf{X} \notin \mathcal{Z}', p(\mathbf{X}) > 0)]| \\ =& K|\mathbb{E}[\mathbb{1}(\mathbf{X} \notin \mathcal{Z}')]| \\ =& K\mathbb{P}(\mathbf{X} \notin \mathcal{Z}') > 0.\end{aligned}$$

Since we can choose K to be arbitrarily big, it follows that $|v^{\text{RJ}}_{\mathbf{x},f_1,p}(\emptyset) - v^{\text{RJ}}_{\mathbf{x},f_2,p}(\emptyset)|$ is not bounded for general functions f_1, f_2 satisfying $\sup_{\mathbf{x}\in\mathcal{Z}'} |f_1(\mathbf{x}) - f_2(\mathbf{x})| \leq \delta$. □

Proof of Proposition 4.

Proof. Recall that in the setting we are considering, the interventional distribution $p(\mathbf{X}_{\bar{S}} \mid do(\mathbf{X}_S = \mathbf{x}_S))$ is equal to the marginal distribution $p(\mathbf{X}_{\bar{S}})$.

Let S be such that $i \notin S$, and let $\mathbf{x} \in \mathcal{Z}$ be any point. Then, $\mathbb{1}((\mathbf{x}_S, \mathbf{X}_{\bar{S}}) \in \mathcal{Z}) = \prod_{j\in\bar{S}} \mathbb{1}(X_j \in \mathcal{Z}_j)$. Using the fact that $\mathbb{1}(X_i \in \mathcal{Z}_i) \overset{\text{a.s.}}{=} 1$, we get that $\mathbb{1}(\mathbf{X}_{\bar{S}} \in \mathcal{Z}) \overset{\text{a.s.}}{=} \prod_{j\in\bar{S}\setminus\{i\}} \mathbb{1}(X_j \in \mathcal{Z}_j)$.

$$\begin{aligned}v^{\text{MAN}}_{\mathbf{x},f,\mathcal{Z}}(S) &= \frac{\mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}})\mathbb{1}((\mathbf{x}_S, \mathbf{X}_{\bar{S}}) \in \mathcal{Z})]}{\mathbb{E}[\mathbb{1}((\mathbf{x}_S, \mathbf{X}_{\bar{S}}) \in \mathcal{Z})]} \\ &= \frac{\mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}}) \prod_{j\in\bar{S}} \mathbb{1}(X_j \in \mathcal{Z}_j)]}{\mathbb{E}[\prod_{j\in\bar{S}} \mathbb{1}(X_j \in \mathcal{Z}_j)]} \\ &= \frac{\mathbb{E}[f(\mathbf{x}_S, \mathbf{X}_{\bar{S}}) \prod_{j\in\bar{S}\setminus\{i\}} \mathbb{1}(X_j \in \mathcal{Z}_j)]}{\mathbb{E}[\prod_{j\in\bar{S}\setminus\{i\}} \mathbb{1}(X_j \in \mathcal{Z}_j)]} \\ &= \frac{\mathbb{E}[f(\mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}}) \prod_{j\in\bar{S}\setminus\{i\}} \mathbb{1}(X_j \in \mathcal{Z}_j)]}{\mathbb{E}[\prod_{j\in\bar{S}\setminus\{i\}} \mathbb{1}(X_j \in \mathcal{Z}_j)]} \\ &= \frac{\mathbb{E}[f(\mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}})\mathbb{1}((\mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}}) \in \mathcal{Z})]}{\mathbb{E}[\mathbb{1}((\mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}}) \in \mathcal{Z})]} = v^{\text{MAN}}_{\mathbf{x},f,\mathcal{Z}}(S \cup \{i\})\end{aligned}$$

where, in the second last step above we use the fact that, $f(\mathbf{x})$ is independent of x_i.

□

Proof of Proposition 5.

Proof. **Notation:** Let $m(\mathbf{x})$ be any function. We use the notation $m(\mathbf{X}_S = \mathbf{x}_S, \mathbf{X}_{\bar{S}} = \mathbf{x}'_{\bar{S}})$ to explicitly denote $m(\mathbf{x}_S, \mathbf{x}'_{\bar{S}})$.

Suppose $m(\mathbf{x})$ is a function symmetric in components i and j. Then, if $S \subseteq [d] \setminus \{i, j\}$. Then,

$$\begin{aligned}
&\mathbb{E}[m(\mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}})] \\
=&\int_{\mathbf{X}'_{\bar{S}\setminus\{i\}}} m(\mathbf{X}_{S\cup\{i\}} = \mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}} = \mathbf{X}'_{\bar{S}\setminus\{i\}}) \int_{\mathbf{Y}_{S\cup\{i\}}} p(\mathbf{X}_{S\cup\{i\}} = \mathbf{Y}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}} = \mathbf{X}'_{\bar{S}\setminus\{i\}}) \mathrm{d}\mathbf{Y}_{S\cup\{i\}} \mathrm{d}\mathbf{X}'_{\bar{S}\setminus\{i\}} \\
=&\int_{\mathbf{X}'_{\bar{S}\setminus\{i\}}} m(\mathbf{X}_{S\cup\{j\}} = \mathbf{x}_{S\cup\{j\}}, \mathbf{X}_i = \mathbf{X}'_j, \mathbf{X}_{\bar{S}\setminus\{i,j\}} = \mathbf{X}'_{\bar{S}\setminus\{i,j\}}) \\
&\times \int_{\mathbf{Y}_{S\cup\{i\}}} p(\mathbf{X}_S = \mathbf{Y}_S, \mathbf{X}_j = \mathbf{Y}_i, \mathbf{X}_i = \mathbf{X}'_j, \mathbf{X}_{\bar{S}\setminus\{i,j\}} = \mathbf{X}'_{\bar{S}\setminus\{i,j\}}) \mathrm{d}\mathbf{Y}_{S\cup\{i\}} \mathrm{d}\mathbf{X}'_{\bar{S}\setminus\{i\}}
\end{aligned}$$

where, in the last step above we use the fact that both, $m(\mathbf{x})$ and $p(\mathbf{x})$ are symmetric in components i and j and $x_i = x_j$. Next, relabelling the dummy variables $\mathbf{X}'_j$ as $\mathbf{X}'_i$ and $\mathbf{Y}_i$ as $\mathbf{Y}_j$, the above becomes

$$\begin{aligned}
&\int_{\mathbf{X}'_{\bar{S}\setminus\{j\}}} m(\mathbf{X}_{S\cup\{j\}} = \mathbf{x}_{S\cup\{j\}}, \mathbf{X}_{\bar{S}\setminus\{j\}} = \mathbf{X}'_{\bar{S}\setminus\{j\}}) \int_{\mathbf{Y}_{S\cup\{j\}}} p(\mathbf{X}_{S\cup\{j\}} = \mathbf{Y}_{S\cup\{j\}}, \mathbf{X}_{\bar{S}\setminus\{j\}} = \mathbf{X}'_{\bar{S}\setminus\{j\}}) \mathrm{d}\mathbf{Y}_{S\cup\{j\}} \mathrm{d}\mathbf{X}'_{\bar{S}\setminus\{j\}} \\
&= \mathbb{E}[m(\mathbf{x}_{S\cup\{j\}}, \mathbf{X}_{\bar{S}\setminus\{j\}})]
\end{aligned}$$

Next, we use the fact that the functions $m_1(\mathbf{x}) \coloneqq \mathbb{1}(\mathbf{x} \in \mathcal{Z})$ and $m_2(\mathbf{x}) \coloneqq f(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathcal{Z})$ are symmetric in components i and j. Therefore, using the result above, we get that,

$$\begin{aligned}
v^{\text{MAN}}_{\mathbf{x},f,\mathcal{Z}}(S \cup \{i\}) =& \frac{\mathbb{E}[f(\mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}})\mathbb{1}((\mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}}) \in \mathcal{Z})]}{\mathbb{E}[\mathbb{1}((\mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}}) \in \mathcal{Z})]} \\
=& \frac{\mathbb{E}[m(\mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}})]}{\mathbb{E}[\mathbb{1}((\mathbf{x}_{S\cup\{i\}}, \mathbf{X}_{\bar{S}\setminus\{i\}}) \in \mathcal{Z})]} \\
=& \frac{\mathbb{E}[m(\mathbf{x}_{S\cup\{j\}}, \mathbf{X}_{\bar{S}\setminus\{j\}})]}{\mathbb{E}[\mathbb{1}((\mathbf{x}_{S\cup\{j\}}, \mathbf{X}_{\bar{S}\setminus\{j\}}) \in \mathcal{Z})]} \\
=& v^{\text{MAN}}_{\mathbf{x},f,\mathcal{Z}}(S \cup \{j\})
\end{aligned}$$

□

Proposition 9. *Let $\mathcal{P}_\alpha$ be as defined in Def 3 and let $\mathcal{Z}$ be any set with $\mathbb{P}(\boldsymbol{X} \in \mathcal{Z}) \geq \mathbb{P}(\boldsymbol{X} \in \mathcal{P}_\alpha)$. Then, if $\epsilon^{(\alpha)} > 0$, we have that $|\mathcal{Z}| \geq |\mathcal{P}_\alpha|$, where $|\mathcal{S}| \coloneqq \int_{\mathcal{S}} \mathrm{d}\boldsymbol{x}$.*

Proof of Proposition 9.

Proof.

$$\begin{aligned}
|\mathcal{Z}| - |\mathcal{P}_\alpha| =& (|\mathcal{Z} \setminus \mathcal{P}_\alpha| + |\mathcal{Z} \cap \mathcal{P}_\alpha|) - (|\mathcal{P}_\alpha \setminus \mathcal{Z}| + |\mathcal{Z} \cap \mathcal{P}_\alpha|) \\
=& |\mathcal{Z} \setminus \mathcal{P}_\alpha| - |\mathcal{P}_\alpha \setminus \mathcal{Z}|
\end{aligned}$$

Similarly,

$$\begin{aligned}
0 \leq& \mathbb{P}(\mathbf{X} \in \mathcal{Z}) - \mathbb{P}(\mathbf{X} \in \mathcal{P}_\alpha) \\
=& (\mathbb{P}(\mathbf{X} \in \mathcal{Z} \setminus \mathcal{P}_\alpha) + \mathbb{P}(\mathbf{X} \in \mathcal{Z} \cap \mathcal{P}_\alpha)) - (\mathbb{P}(\mathbf{X} \in \mathcal{P}_\alpha \setminus \mathcal{Z}) + \mathbb{P}(\mathbf{X} \in \mathcal{Z} \cap \mathcal{P}_\alpha)) \\
=& \mathbb{P}(\mathbf{X} \in \mathcal{Z} \setminus \mathcal{P}_\alpha) - \mathbb{P}(\mathbf{X} \in \mathcal{P}_\alpha \setminus \mathcal{Z}) \\
=& \int_{\mathcal{Z}\setminus\mathcal{P}_\alpha} p(\mathbf{x}) \mathrm{d}\mathbf{x} - \int_{\mathcal{P}_\alpha\setminus\mathcal{Z}} p(\mathbf{x}) \mathrm{d}\mathbf{x} \\
\leq& \int_{\mathcal{Z}\setminus\mathcal{P}_\alpha} \epsilon^{(\alpha)} \mathrm{d}\mathbf{x} - \int_{\mathcal{P}_\alpha\setminus\mathcal{Z}} \epsilon^{(\alpha)} \mathrm{d}\mathbf{x} \\
=& \epsilon^{(\alpha)} \left(|\mathcal{Z} \setminus \mathcal{P}_\alpha| - |\mathcal{P}_\alpha \setminus \mathcal{Z}|\right)
\end{aligned}$$

In the second last step above, we use the fact that $p(\mathbf{x}) \geq \epsilon^{(\alpha)} \iff \mathbf{x} \in \mathcal{P}_\alpha$. Using the condition $\epsilon^{(\alpha)} > 0$, we get that

$$|\mathcal{Z}| - |\mathcal{P}_\alpha| = |\mathcal{Z} \setminus \mathcal{P}_\alpha| - |\mathcal{P}_\alpha \setminus \mathcal{Z}| \geq 0$$

□

C ALTERNATIVE METHODOLOGIES OF COMPUTING MANIFOLDSHAP

In this section, we outline alternative methodologies of computing ManifoldShap value function. As before, we assume that we can sample from the interventional distribution $p(\mathbf{X}_{\bar{S}} \mid do(\mathbf{X}_S = \mathbf{x}_S))$ for any $S \subseteq [d]$. This is a standard assumption needed to estimate Interventional Shapley.

C.1 Supervised approach

Here, we use the fact that the expectation $g(\mathbf{x}_S) \coloneqq \mathbb{E}[f(\mathbf{X}) \mid do(\mathbf{X}_S = \mathbf{x}_S), \mathbf{X} \in \mathcal{Z}]$ minimises the mean squared error $\mathcal{L}_S(h) = \mathbb{E}_{\tilde{\mathbf{X}}_S \sim p(\mathbf{X}_S), \tilde{\mathbf{X}} \sim p(\mathbf{X} \mid do(\mathbf{X}_S = \tilde{\mathbf{X}}_S), \mathbf{X} \in \mathcal{Z})}[f(\tilde{\mathbf{X}}) - h(\tilde{\mathbf{X}}_S)]^2$. Using this, we can define a surrogate model $g_\theta(\mathbf{x}_S)$ that takes as input coalition of features $\mathbf{x}_S$ (e.g., by masking features in $\bar{S}$) and that is trained to minimise the loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{\tilde{\mathbf{X}}_S \sim p(\mathbf{X}_S), \tilde{\mathbf{X}} \sim p(\mathbf{X} \mid do(\mathbf{X}_S = \tilde{\mathbf{X}}_S), \mathbf{X} \in \mathcal{Z})} \mathbb{E}_{S \sim \text{Shapley}}[f(\tilde{\mathbf{X}}) - g_\theta(\tilde{\mathbf{X}}_S)]^2$$

Here, $S \sim$ Shapley corresponds to sampling coalitions from the distribution where the probability assigned to each coalition is the combinatorial factor $|S|!(n - |S| - 1)!/n!$. Additionally, rejection sampling can be used to sample $\tilde{\mathbf{X}} \sim p(\mathbf{X} \mid do(\mathbf{X}_S = \tilde{\mathbf{X}}_S), \mathbf{X} \in \mathcal{Z})$. To be specific, we repeated sample $\tilde{\mathbf{X}} \sim p(\mathbf{X} \mid do(\mathbf{X}_S = \tilde{\mathbf{X}}_S))$ until the sampled value $\tilde{\mathbf{X}}$ lies in $\mathcal{Z}$.

C.2 Rejection sampling

In this subsection, we extend the approach in Strumbelj and Kononenko [2010] to propose an efficient sampling-based approximation. This approximation uses the following alternative formulation of Shapley values:

$$\phi_i = \sum_{\pi \in \Pi} \frac{1}{n!} \left[v(\{j : \pi(j) \leq \pi(i)\}) - v(\{j : \pi(j) < \pi(i)\}) \right]$$

where Π denotes the set of permutations of N, and $\pi(j) < \pi(i)$ means that j precedes i under ordering π. To derive the sampling procedure, we observe that the Shapley value for feature i, π_i, can be written as an average over the set of permutations, i.e.,

$$\phi_i = \mathbb{E}_\pi[v(\{j : \pi(j) \leq \pi(i)\}) - v(\{j : \pi(j) < \pi(i)\})]$$

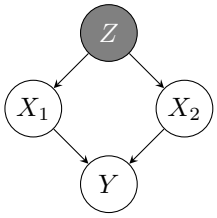
where the permutations π are drawn from a uniform distribution over Π. Using this, we derive the following procedure for obtaining an unbiased and consistent estimation of ManifoldShap.

Algorithm 1 Approximating ManifoldShap value ϕ_i for instance $\mathbf{x} \in \mathcal{Z}$.

Input: Instance $\mathbf{x}$; the desired number of samples m; feature i to compute Shapley value for;
$\phi_i \leftarrow 0$
for $j = 1$ to m **do**
  choose a random permutation of features $\pi \in \Pi$
  $S \leftarrow \{j : \pi(j) < \pi(i)\}$
  sample $\mathbf{y}_{\bar{S} \setminus \{i\}} \sim p(\mathbf{X}_{\bar{S} \setminus \{i\}} \mid do(\mathbf{X}_{S \cup \{i\}} = \mathbf{x}_{S \cup \{i\}}))$
  while $(\mathbf{x}_{S \cup \{i\}}, \mathbf{y}_{\bar{S} \setminus \{i\}}) \notin \mathcal{Z}$ **do**
    sample $\mathbf{y}_{\bar{S} \setminus \{i\}} \sim p(\mathbf{X}_{\bar{S} \setminus \{i\}} \mid do(\mathbf{X}_{S \cup \{i\}} = \mathbf{x}_{S \cup \{i\}}))$
  end while
  sample $\mathbf{z}_{\bar{S}} \sim p(\mathbf{X}_{\bar{S}} \mid do(\mathbf{X}_S = \mathbf{x}_S))$
  while $(\mathbf{x}_S, \mathbf{z}_{\bar{S}}) \notin \mathcal{Z}$ **do**
    sample $\mathbf{z}_{\bar{S}} \sim p(\mathbf{X}_{\bar{S}} \mid do(\mathbf{X}_S = \mathbf{x}_S))$
  end while
  $\phi_i \leftarrow \phi_i + (f(\mathbf{x}_{S \cup \{i\}}, \mathbf{y}_{\bar{S} \setminus \{i\}}) - f(\mathbf{x}_S, \mathbf{z}_{\bar{S}}))$
end for
$\phi_i \leftarrow \phi_i / m$
Return: ϕ_i

D INTERVENTIONAL SHAPLEY VS CONDITIONAL EXPECTATION SHAPLEY

Example. Assume that $\mathcal{X} = \{0,1\}^2$, and that the features X_1, X_2 follow the causal structure shown below. In this setting, interventional distributions are equivalent to marginal distributions, i.e., $p(\mathbf{X}_{\bar{S}} \mid do(\mathbf{X}_S = \mathbf{x}_S)) = p(\mathbf{X}_{\bar{S}})$.



Consider the case where $f(x_1, x_2) = x_1$ and Z, X_1, X_2 are binary variables, with

$$\begin{aligned} Z &= \begin{cases} 0 & \text{w.p. } 0.5 \\ 1 & \text{otherwise} \end{cases} \\ X_1 &= Z \\ X_2 &= \begin{cases} Z & \text{w.p. } p \text{ (for some } p > 0), \\ 1 - Z & \text{otherwise.} \end{cases} \end{aligned}$$

In this case, $\mathbb{E}[f(X_1, X_2) \mid do(X_2 = x_2)] = \mathbb{E}[f(X_1, X_2)] = \mathbb{E}[f(X_1, x_2)] = 1/2$ and

$$\mathbb{E}[f(X_1, X_2) \mid do(X_1 = x_1, X_2 = x_2)] = \mathbb{E}[f(X_1, X_2) \mid do(X_1 = x_1)] = \mathbb{E}[f(x_1, X_2)] = x_1,$$

for any x_2. It straightforwardly follows that, in this case, the Interventional Shapley value for feature x_2, $\phi_2 = 0$, i.e. Interventional Shapley satisfies the Sensitivity property.

However, if we use CES instead, we get that

$$\begin{aligned} \phi_2 =& \frac{1}{2}\left(\mathbb{E}[f(X_1, X_2) \mid X_2 = x_2] - \mathbb{E}[f(X_1, X_2)] + \mathbb{E}[f(X_1, X_2) \mid X_1 = x_1, X_2 = x_2] - \mathbb{E}[f(X_1, X_2) \mid X_1 = x_1]\right) \\ =& \frac{1}{2}\left(\mathbb{E}[X_1 \mid X_2 = x_2] - 1/2 + x_1 - x_1\right) \\ =& \frac{1}{2}\left(\mathbb{P}(X_1 = 1 \mid X_2 = x_2) - 1/2\right) \\ =& \frac{1}{2}\left(\frac{1/2p\mathbb{1}(x_2 = 1) + 1/2(1-p)\mathbb{1}(x_2 = 0)}{1/2} - 1/2\right) \\ =& \frac{1}{2}\left(p\mathbb{1}(x_2 = 1) + (1-p)\mathbb{1}(x_2 = 0) - 1/2\right). \end{aligned}$$

which is non-zero when $p \neq 1/2$.

This example illustrates that CES value function can lead to misleading Shapley values, especially when the features are highly correlated. Interventional Shapley value function, on the other hand, incorporates the causal effect of *fixing* a set of features S, and therefore, yields Shapley values which are unaffected by correlations within the data.

E COMPUTING CONDITIONAL EXPECTATION SHAPLEY USING SUPERVISED APPROACH

In this work, when the conditional distribution is not tractable analytically, we use the supervised learning approach as in Frye et al. [2021] to estimate the conditional expectation in CES. We present this methodology in this section for completeness.

Here, we use the fact that the conditional expectation $g(\mathbf{x}_S) \coloneqq \mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$ minimises the mean squared error $\mathcal{L}_S(\tilde{g}) = \mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}[f(\mathbf{X}) - \tilde{g}(\mathbf{X}_S)]^2$. Using this, we can define a surrogate model $g_\theta(\mathbf{x}_S)$ that takes as input coalition of features $\mathbf{x}_S$ (e.g., by masking features in $\bar{S}$) and that is trained to minimise the loss:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{X}\sim p(\mathbf{X})}\mathbb{E}_{S\sim \text{Shapley}}[f(\mathbf{X}) - g_\theta(\mathbf{X}_S)]^2$$

Here, $S \sim$ Shapley corresponds to sampling coalitions from the distribution where the probability assigned to each coalition is the combinatorial factor $|S|!(n - |S| - 1)!/n!$. As the surrogate model $g_\theta(\mathbf{x}_S)$ approaches the CES value function $\mathbb{E}[f(\mathbf{X}) \mid \mathbf{X}_S = \mathbf{x}_S]$, the loss $\mathcal{L}(\theta)$ is minimised.

F COMPUTING THE MANIFOLD $\mathcal{D}_\epsilon$

If we choose the manifold $\mathcal{Z}$ based on probability density/mass as outlined in Definitions 2 and 3, we must estimate the region $\mathcal{D}_\epsilon \coloneqq \{\mathbf{x} : p(\mathbf{x}) > \epsilon\}$. There are various ways to esimate this set, and the methodology used depends on dataset properties, such as the data dimensions, as well as the degree of accuracy sought. Below, we outline two such solutions, which can be used depending on the dataset dimensions:

F.1 Using Variational Auto-Encoders for manifold estimation

Variational Auto-Encoders (VAEs) have been a popular method of density estimation [Kingma and Welling, 2013, Kingma et al., 2019]. Instead of maximising the log likelihood, which may be intractable in general, VAE training involves maximising a lower bound of the log likelihood, called the Evidence Lower Bound (ELBO). In order to do so, the VAEs assume that data are generated from some random process, involving latent random variables $\mathbf{z}$, and that a value $\mathbf{x}$ is sampled from a conditional $p_\theta(\mathbf{x} \mid \mathbf{z})$, also referred to as the likelihood [Kingma et al., 2019].

Let $q_\phi(\mathbf{z} \mid \mathbf{x})$ be a parametrized posterior. Then, we have that

$$\log p_\theta(\mathbf{x}) = \log \int \frac{p_\theta(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} q_\phi(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} \geq \int \log \left(\frac{p_\theta(\mathbf{x} \mid \mathbf{z})p(\mathbf{z})}{q_\phi(\mathbf{z} \mid \mathbf{x})} \right) q_\phi(\mathbf{z} \mid \mathbf{x}) d\mathbf{z} =: \text{ELBO}_{\theta,\phi}(\mathbf{x})$$

The VAE training therefore involves maximising the expected lower bound $\mathbb{E}[\text{ELBO}_{\theta,\phi}(\mathbf{X})]$ over θ, ϕ. Let $(\theta^*, \phi^*) \coloneqq \arg\max_{(\theta,\phi)} \mathbb{E}[\text{ELBO}_{\theta,\phi}(\mathbf{X})]$, then we have that

$$\log p_{\theta^*}(\mathbf{x}) \geq \text{ELBO}_{\theta^*,\phi^*}(\mathbf{x}).$$

Therefore, $\exp(\text{ELBO}_{\theta^*,\phi^*}(\mathbf{x})) \geq \epsilon$ implies that $p_{\theta^*}(\mathbf{x}) \geq \epsilon$. We can use the ELBO to approximate the manifold:

$$\mathcal{D}_\epsilon \approx \{\mathbf{x} : \exp(\text{ELBO}_{\theta^*,\phi^*}(\mathbf{x})) \geq \epsilon\}.$$

Assuming that $p_{\theta^*}(\mathbf{x})$ is an accurate density model, we get that the above approximation of the ϵ-manifold is going to be conservative in the sense that it will be a subset of the true $\mathcal{D}_\epsilon$.

F.2 Thresholded Kernel Density Classification (tKDC)

Alternatively, we can use Kernel Density Estimation (KDE) to estimate the manifold $\mathcal{D}_\epsilon$. KDE provides a way of estimating normalized density $\hat{p}(\mathbf{x})$ using a finite dataset. This can be used to approximate most well-behaved smooth densities. Given n datapoints, KDE provides an estimate $\hat{p}_n(\mathbf{x})$ with mean squared error that shrinks at rate $O(n^{-\frac{4}{4+d}})$, where d is the dimension of $\mathbf{x}$. This means that with enough data, KDE will identify an accurate density. The same may not be true for parametric methods of density estimation. However, evaluating the density $\hat{p}_n(\mathbf{x})$ at a point $\mathbf{x}$ is prohibitively expensive when n is large, as it involves the kernel contributions from every point in dataset.

To circumvent this problem, Gan and Bailis [2017] propose tKDC, a computationally efficient algorithm for classifying points with $p(\mathbf{x}) \geq \epsilon$ using KDE, where classification errors are allowed for densities within $\pm\epsilon\delta$ of the density threshold ϵ (for a given $\delta > 0$).

When the dimension of data, d, is large, the convergence of $\hat{p}_n(\mathbf{x})$ is extremely slow. This is reflected in the error term $O(n^{-\frac{4}{4+d}})$ which shrinks slowly with increasing n when d is large. Therefore, while tKDC provides an asymptotically accurate density classification methodology, the convergence can be slow for large d, and in this case alternative methodologies like using VAEs may be more feasible.

F.3 Choosing the threshold ϵ

Our manifold $\mathcal{D}_\epsilon$ is parameterized by a density threshold ϵ. In practice, the probability density may depend on the dataset size, dimensionality and distribution, and as a result the range of density values may vary substantially among different datasets. It is therefore not possible to a priori define threshold values of ϵ. Instead, we specify thresholds in terms of a probability mass $\alpha \in [0, 1]$. That is, we pick a fraction of the data that we choose to classify as having low density and set the threshold accordingly. This corresponds to picking the value of ϵ to be $\epsilon^{(\alpha)}$, where $\epsilon^{(\alpha)} \coloneqq \sup\{\epsilon \geq 0 : \mathbb{P}(\mathbf{X} \in \mathcal{D}_\epsilon) \geq \alpha\}$. In practice, since we do not have access to the true density model, the estimation of $\epsilon^{(\alpha)}$ can be difficult, and we pick the threshold $\epsilon^{(\alpha)}$ based on the quantiles of the observed density estimates $\hat{p}(\mathbf{x})$ for $\mathbf{x}$ in some held out dataset. The authors in Cadre et al. [2013] show that for kernel density estimators this quantile converges to the ideal $\epsilon^{(a)}$ defined above.

G EXPERIMENTAL RESULTS

G.1 Experimental details for real-world dataset experiments

Here, we explicitly define the models used for the real-world dataset experiments.

G.1.1 COMPAS dataset

Biased classifier. For COMPAS dataset, the biased classifier f is defined as:

$$f(\mathbf{x}) \coloneqq \mathbb{1}(\mathbf{x}_{\text{race}} = \text{African American}).$$

Here, $\mathbf{x}_{\text{race}}$ denotes the race feature of the for the datapoint $\mathbf{x}$.

Out of manifold perturbation. To perturb the model outside the manifold $\mathcal{Z}$, we construct 2 synthetic binary features (referred to as 'unrelated columns') positively correlated with race.

Let Z_i be i.i.d. random variables with distribution $\text{Bern}(0.90)$, then the 'unrelated columns' feature correponding to the i'th datapoint, $\mathbf{X}^{(i)}_{\text{unrelated column}}$, is defined as follows:

$$\mathbf{X}^{(i)}_{\text{unrelated column}} \coloneqq \mathbb{1}(\mathbf{X}^{(i)}_{\text{race}} = \text{African American}) \times Z_i.$$

Finally, the perturbed classifier model $g_{\mathcal{Z}} : \mathcal{X} \to \{0, 1\}$ is defined as follows:

$$g_{\mathcal{Z}}(\mathbf{x}) \coloneqq \mathbb{1}(\mathbf{x} \in \mathcal{Z})\, f(\mathbf{x}) + \mathbb{1}(\mathbf{x} \notin \mathcal{Z})\, \mathbb{1}(\mathbf{x}_{\text{unrelated column}} > 0).$$

G.1.2 Communities and crime dataset

Biased classifier. Likewise, for the CC dataset, the biased classifier f is defined as:

$$f(\mathbf{x}) \coloneqq \mathbb{1}(\mathbf{x}_{\text{percentage of caucasian population}} > 0).$$

Out of manifold perturbation. We again construct 2 synthetic features (referred to as 'unrelated columns'). Using the same random variables Z_i as defined above, the 'unrelated columns' feature corresponding to the i'th datapoint, $\mathbf{X}^{(i)}_{\text{unrelated column}}$, is defined as follows:

$$\mathbf{X}^{(i)}_{\text{unrelated column}} \coloneqq \mathbf{X}^{(i)}_{\text{percentage of caucasian population}} \times Z_i.$$

Just like in COMPAS dataset experiments, the perturbed classifier model $g_{\mathcal{Z}} : \mathcal{X} \to \{0, 1\}$ is defined as follows:

$$g_{\mathcal{Z}}(\mathbf{x}) \coloneqq \mathbb{1}(\mathbf{x} \in \mathcal{Z})\, f(\mathbf{x}) + \mathbb{1}(\mathbf{x} \notin \mathcal{Z})\, \mathbb{1}(\mathbf{x}_{\text{unrelated column}} > 0).$$

G.2 Additional Experiments

For all experiments in this section, we consider the causal structure in Janzing et al. [2020] (see Figure 2), where the true features are formally distinguished from the input features. In this setting, the Interventional Shapley is equivalent to Marginal Shapley.

G.2.1 Off-manifold perturbation

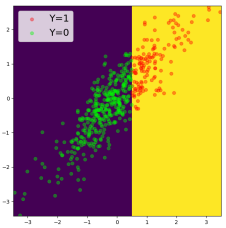
In this experiment we investigate the effect of model perturbation in low density regions on Shapley values obtained using our methodology as well as other baselines. We do so by defining adversarial models, which agree with the ground truth model on the manifold $\mathcal{P}_\alpha$, but have been perturbed outside the manifold.

First, we define a ground truth data generating mechanism as described below.

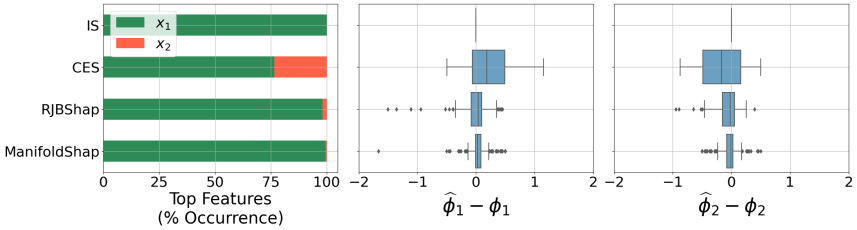
Data generating mechanism. In this experiment, $\mathcal{Y} = \{0, 1\}$ and $\mathcal{X} \subseteq \mathbb{R}^2$, where:

$$\mathbf{X} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0.90\\0.90 & 1\end{pmatrix}\right)$$
$$Y \coloneqq \mathbb{1}(X_1 > 1/2).$$

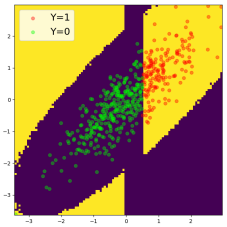
Next, for the adversarial models, we define the following family of perturbed models.



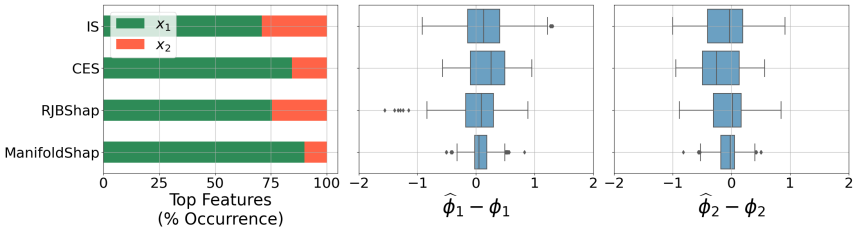
(a) Classifier g_δ for $\delta = 0$.



(b) $\delta = 0$



(c) Classifier g_δ for $\delta = 10$.



(d) $\delta = 10$

Figure 6: **Left:** 6a and 6c show classifier decision boundaries for $\delta = 0, 10$ respectively. **Right:** 6b and 6d show boxplots of $\hat{\phi}_i - \phi_i$ for $i \in \{1, 2\}$ and different off-manifold perturbations.

Perturbed models. We define the following family of perturbed models $g_\delta : \mathcal{X} \to \{0, 1\}$, parameterised by $\delta \in \mathbb{R}$.

$$g_\delta(\mathbf{X}) \coloneqq Y \mathbb{1}(\mathbf{X} \in \mathcal{P}_\alpha) + \mathbb{1}((1-\delta)X_1 > 1/2)\mathbb{1}(\mathbf{X} \notin \mathcal{P}_\alpha).$$

Here, we use VAEs to estimate $\mathcal{P}_\alpha$ as described in Section F, and choose $\alpha = 1 - 10^{-3}$.

By construction, the classifiers g_δ should agree with the ground truth on the α-manifold, i.e. $g_\delta(\mathbf{X}) = Y$ when $\mathbf{X} \in \mathcal{P}_\alpha$. However, these classifiers differ from the ground truth outside the α-manifold. Figures 6a and 6c show the classifier decision boundaries, along with the original data (X_1, X_2). Each of these classifiers have a test accuracy of at least 99.5%, and therefore it is impossible to distinguish between them on the data manifold. However, as we will show next, the Interventional Shapley values computed for these classifiers are drastically different.

Estimating conditional expectation for CES. For the data generating mechanism described above, the conditional distributions $p(X_2 \mid X_1)$ and $p(X_1 \mid X_2)$ are tractable. In fact, it is straightforward to get that $X_2 \mid X_1 \sim \mathcal{N}(0.90 * X_1, 1 - 0.90^2)$, and similarly for $X_1 \mid X_2$. We use this to estimate the conditional expectation using m Monte Carlo samples from the conditional distributions:

$$\mathbb{E}[f(\mathbf{X}) \mid X_1 = x_1] \approx \frac{1}{m} \sum_{i=1}^{m} f(x_1, X_2^i) \quad \text{where}$$
$$X_2^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0.90 * x_1, 1 - 0.90^2),$$

and similarly for $\mathbb{E}[f(\mathbf{X}) \mid X_2 = x_2]$. In this experiment, we use $m = 500$.

Results. We compute the Shapley values for the models on 500 datapoints from a held-out dataset. In Figure 6 we plot the difference between estimated Shapley values and the ground truth Interventional Shapley values, for different methodologies. For a fair comparison between different value functions, we normalise the Shapley values so that $\sum_{i \in \{1,2\}} |\phi_i| = 1$.

To compute the ground truth IS values $\{\phi_i\}_{i \in \{1,2\}}$, we use the ground truth function $f(\mathbf{X}) = \mathbb{1}(X_1 > 1/2)$ instead of the perturbed model g_δ, and therefore, the ground truth IS values do not change with increasing off-manifold perturbations δ. Moreover, note that since the ground truth model f is independent of X_2, the Shapley values for feature 2, $\phi_2 = 0$.

The figure shows that when $\delta = 0$, i.e., there is no off-manifold perturbation, the errors for IS values, i.e., $\hat{\phi}_i - \phi_i$, are 0. This is because the IS values are equal to the ground truth in this case, as the perturbed model g_δ is equal to the ground truth model f everywhere. CES, on the other hand gives biased Shapley values, as can be seen from the errors $\hat{\phi}_i - \phi_i$ being concentrated away from 0. This happens because of the high positive correlation between the features – conditional expectation is highly sensitive to feature correlations, unlike marginal expectation.

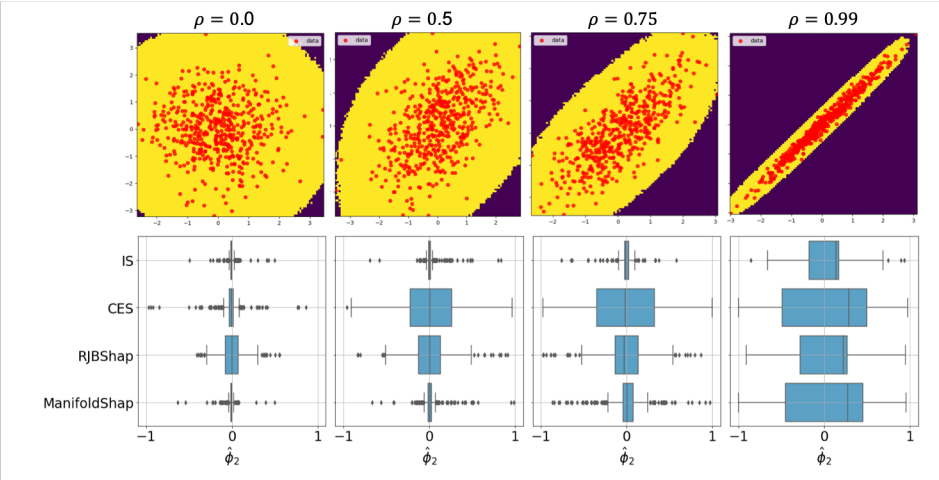


Figure 7: **Fig (a) (above):** Manifolds $\mathcal{D}_\epsilon$ for different values of ρ. The manifold $\mathcal{D}_\epsilon$ is denoted by the yellow region. **Fig (b) (below):** Boxplots of $\hat{\phi}_2$ for increasing values of ρ.

For ManifoldShap, the errors are less concentrated around 0 than for IS values. This highlights the reduced causal accuracy in ManifoldShap as a result of restricting function evaluations to the manifold $\mathcal{P}_\alpha$. However, ManifoldShap values are more accurate than the CES and RJBShap values.

It can be seen that for $\delta = 10$, the errors in IS values are highest among all the baselines. This highlights the off-manifold nature of IS values, i.e., perturbing model in low-density regions can significantly change the computed Shapley values. Additionally, CES values are biased as the errors are concentrated away from 0. ManifoldShap errors remain largely restricted between -0.2 and 0.2, with error distribution concentrated around 0.

The barplots in Figures 6b and 6d show the most important features as per different value functions for $\delta = 0, 10$ respectively. The figure shows that for ground truth model ($\delta = 0$), IS values attribute the greatest importance to feature 1 for all datapoints. This is expected as the ground truth model does not depend on x_2. For CES, on the other hand, feature 2 receives greater importance for roughly 25% datapoints. This is again due to the positive correlation between the features X_1, X_2. For $\delta = 10$, Figure 6d shows that ManifoldShap attributes least importance to feature 2, among all baselines considered.

The results show that CES and ManifoldShap are in practice less sensitive to off-manifold manipulation compared to IS and RJBShap, however, CES values may be biased when features are highly correlated, whereas ManifoldShap remain closer to the ground truth IS values overall.

G.2.2 Sensitivity to correlations

In this experiment we investigate the sensitivity of ManifoldShap to increasing correlation among the features, as compared to the other baselines. To this end, we define the following family of data distributions:

Data generating mechanism. In this experiment, $\mathcal{X} \subseteq \mathbb{R}^2$ and $\mathcal{Y} \subseteq \mathbb{R}$. Specifically,

$$\mathbf{X} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right) \quad \text{where, } \rho \in (-1, 1).$$

Moreoever, the ground truth model under consideration is $Y := f(\mathbf{X}) = X_1$. The parameter ρ corresponds to the correlation between X_1 and X_2. When $\rho = 0$, X_1 and X_2 are independent random variables. As ρ increases (decreases), the features get more positively (negatively) correlated. Figure 7a shows the data generated for different values of ρ, along with the ϵ-manifold. Here, we choose ϵ to be the 1st percentile of density values on a held out dataset, i.e. $\epsilon \approx \epsilon^{(0.99)}$.

Using the data generating mechanism described above, we generate data for a given ρ, which is then used to estimate Shapley values. Note that since f is independent of X_2, we would ideally expect the Shapley values corresponding to feature X_2, i.e., $\hat{\phi}_2$, to be close to 0.

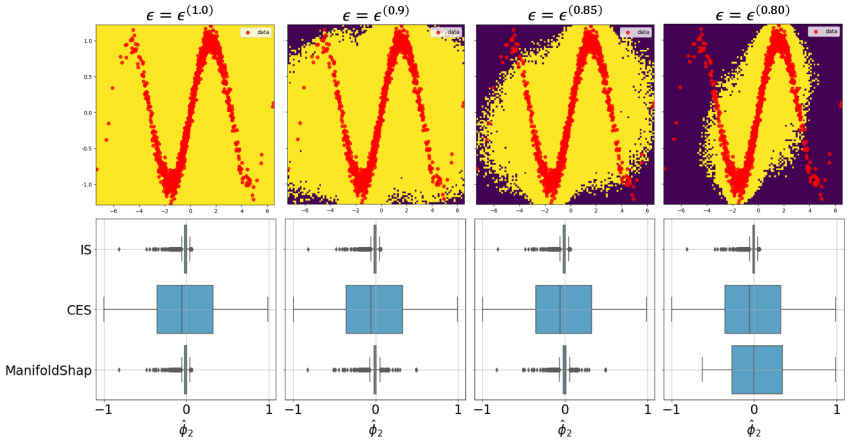


Figure 8: **Fig (a) (above):** Manifolds $\mathcal{D}_\epsilon$ for different values of ϵ. The manifold $\mathcal{D}_\epsilon$ is denoted by the yellow region. **Fig (b) (below):** Boxplots of $\hat{\phi}_2$ for increasing values of ϵ.

Estimating conditional expectation for CES. Like the previous experiment, the conditional distributions $p(X_2 \mid X_1)$ and $p(X_1 \mid X_2)$ are tractable for the data generating mechanism described above. In fact, $X_2 \mid X_1$ follows the Gaussian distributions $\mathcal{N}(\rho X_1, 1 - \rho^2)$ (and similarly for $X_1 \mid X_2$). We use this to estimate the conditional expectation using $m = 500$ Monte Carlo samples like described for previous experiment.

Results. Figure 7b shows the boxplots of $\hat{\phi}_2$ values for the different methodologies and values of correlation ρ. For $\rho = 0$, the features are independent, and in this case the CES and IS values are expected to be equal, as the conditional expectation is same as marginal expectation in this case. Therefore, we observe in Figure 7b that CES values are close to IS values when $\rho = 0$.

As ρ increases, the distribution of CES values $\hat{\phi}_2$ gets more spread out, away from the ground truth value of 0. In comparison, both ManifoldShap and IS values remain concentrated around 0, with IS remaining closer to 0. This happens because IS values are not sensitive to feature correlations. Furthermore, ManifoldShap values are significantly less sensitive to increasing ρ as compared to CES values. In comparison, it can be seen that the distribution of RJBShap values gets wider as ρ increases, showing that RJBShap is more sensitive to increasing correlation than ManifoldShap values. This is because RJBShap values explicitly depend on the density values $p(\mathbf{x})$ which changes with changing values of ρ.

Finally, when $\rho = 0.99$, the features are highly correlated. In this case, the manifold $\mathcal{D}_\epsilon$ is sparse, and as a result the ManifoldShap and CES behave similarly. This is evident from the fact that the boxplots of $\hat{\phi}_2$ for ManifoldShap and CES in Figure 7b are very similar when $\rho = 0.99$. This also highlights a potential failure mode of ManifoldShap: when the manifold $\mathcal{Z}$ is sparse, the ManifoldShap may behave similarly to CES, leading to unintuitive explanations.

G.2.3 Dependence on manifold size

In this experiment, we investigate how the ManifoldShap values change as the size of $\mathcal{Z}$ decreases. In particular, we investigate the relationship between ManifoldShap, IS and CES as the manifold $\mathcal{Z} = \mathcal{D}_\epsilon$ gets smaller. To do so, we consider $\epsilon \in \{\epsilon^{(1.0)}, \epsilon^{(0.9)}, \epsilon^{(0.85)}, \epsilon^{(0.80)}\}$, where $\epsilon^{(\alpha)}$ is as defined in definition 3. We carry out this experiment on the following data generating mechanism, with $\mathcal{X} \subseteq \mathbb{R}^2$ and $\mathcal{Y} \subseteq \mathbb{R}$:

Sine Wave.

$$X_1 \sim \mathcal{N}(0, 4); \quad X_2 \mid X_1 \sim \mathcal{N}(\sin(X_1), 0.01).$$

Moreoever, the ground truth model under consideration is $Y := f(\mathbf{X}) = X_1$. Using the data generating mechanism described above, we generate data which is then used to compute Shapley values of f. Figure 8a shows how the ϵ-manifolds shrinks as ϵ increases from 0. Here, we use the supervised approach described in Section E to compute conditional expectation for CES values since the conditional $X_1 \mid X_2$ is not easily tractable.

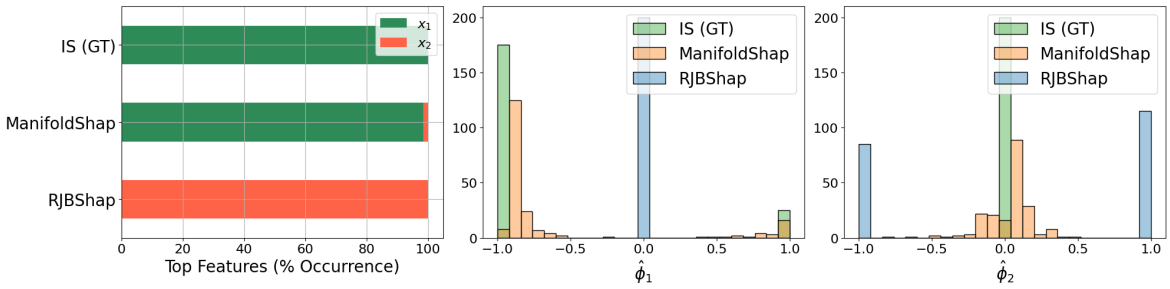


Figure 9: **Left (a):** Top features according to different Shapley value functions. **Right (b):** Histograms of computed Shapley values using different value functions. 'IS (GT)' refers to the ground truth Interventional Shapley values.

Results for Sine Wave. Figure 8 shows the boxplots of $\hat{\phi}_2$ for different values of ϵ. Recall that since the ground truth function is independent of X_2, the ground truth IS value $\phi_2 = 0$. When $\epsilon = \epsilon^{(1.0)} = 0$, ManifoldShap is equivalent to IS, and therefore the values in Figure 8 are identical. CES values, on the other hand are concentrated away from 0. This happens because the features are highly coupled in this experiment.

As ϵ increases, the ManifoldShap values for $\hat{\phi}_2$ get increasingly spread out. This shows that increasing ϵ may reduce the causal accuracy of computed ManifoldShap values, despite making them more robust to off-manifold perturbations. However, it is important to note that relative to CES, ManifoldShap values are closer to the ground truth than Interventional Shapley values for $\epsilon \in \{\epsilon^{(1.0)}, \epsilon^{(0.9)}, \epsilon^{(0.85)}\}$. When $\epsilon = \epsilon^{(0.80)}$, the ManifoldShap values for $\hat{\phi}_2$ are no longer concentrated around the ground truth value 0, as the manifold $\mathcal{D}_\epsilon$ excludes a significant number of input samples. This shows that ManifoldShap values may become inaccurate when the set $\mathcal{Z}$ becomes 'small' relative to the true data manifold.

G.2.4 ManifoldShap vs RJBShap

In this experiment, we demonstrate that RJBShap provides explanations for $\tilde{f}_p(\mathbf{x}) := f(\mathbf{x})p(\mathbf{x})$ which is fundamentally different from $f(\mathbf{x})$. The explanations obtained can therefore be misleading.

Data generating mechanism. In this experiment, $\mathcal{X} \subseteq \mathbb{R}^2$ and $\mathcal{Y} \subseteq \mathbb{R}$. Specifically,

$$\mathbf{X} \sim \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right)$$

Moreover, the ground truth model under consideration is $Y := f(\mathbf{X}) = \exp(X_1^2/2)$. The model is deliberately chosen so that $f(\mathbf{X})p(\mathbf{X})$ only depends on X_2, while $f(\mathbf{X})$ only depends on X_1. Therefore, the Interventional Shapley value for feature 2 is 0, i.e., $\phi_2 = 0$. In contrast, the RJBShap value for feature 1 is 0, since $\tilde{f}_p(\mathbf{X})$ is independent of feature 1. Therefore, if we use RJBShap to explain the function $f(\mathbf{X})$, we would be misled into concluding that the function is independent of feature 1, when in reality the function is independent of feature 2.

For this experiment, we use $\mathcal{Z} = \mathcal{P}_\alpha$ with $\alpha = 1 - 10^{-3}$ to compute ManifoldShap, and use the ground truth data density to compute RJBShap values.

Results. Using the data-generating mechanism defined above, we generate datapoints $\{\mathbf{x}^{(i)}\}_{i=1}^{500}$, and compute Shapley values for these datapoints. Figure 9a shows the top features according to different value functions. The results confirm that, for RJBShap the top feature is feature 2 for all datapoints, whereas for ground truth Interventional Shapley value feature 1 is most important for all datapoints. ManifoldShap values remain significantly closer to ground truth IS values, as over 98% datapoints have feature 1 as the most important feature.

Figure 9b shows the histograms of Shapley values for different value functions. For a fair comparison, we normalise the Shapley values so that $\sum_{i\in\{1,2\}} |\phi_i| = 1$. It can be seen that ManifoldShap values are very close to the ground truth IS values, while the RJBShap values provide a stark contrast to the ground truth IS values. For IS values, $|\phi_1| = 1$ and $\phi_2 = 0$ which accurately reflects the fact that the function f only depends on feature 1. In contrast, for RJBShap, $\phi_1 = 0$ and $|\phi_2| = 1$.

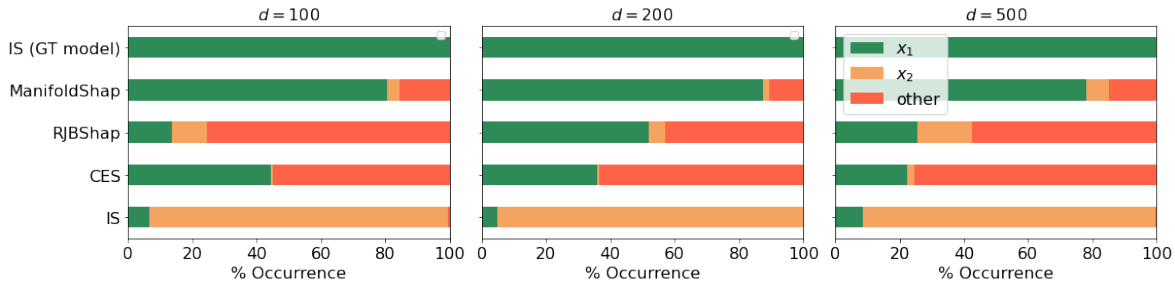


Figure 10: Top features according to different Shapley value functions for increasing feature dimensions d. 'IS (GT)' refers to the ground truth Interventional Shapley values.

G.2.5 Accuracy with increasing feature space dimensions

In this experiment, we illustrate how the accuracy of computed Shapley values varies with increasing dimensions of the feature space. Here, we consider dimensions of feature space $d \in \{100, 200, 500\}$.

Data generating mechanism. Here, $\mathbf{X} \in \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. Specifically,

$$\mathbf{X} \sim \mathcal{N}(\mathbf{0}_d, \Sigma^d), \quad \text{where} \quad \Sigma^d_{ij} = \mathbb{1}(i = j) + 0.9\mathbb{1}(i \neq j)$$
$$Y = X_1.$$

In this example, the correlation between any two features is 0.90. This high positive correlation among features restricts support size of the data. Additionally, the model under consideration is the perturbed model

$$f(\mathbf{X}) = Y + 10\, X_2 \mathbb{1}(\mathbf{X} \notin \mathcal{P}_\alpha).$$

We use VAEs to estimate $\mathcal{P}_\alpha$ as described in Section F, and choose $\alpha = 1 - 10^{-3}$.

Like in previous experiments, the model only depends on the first feature X_1 on the α-manifold, i.e., $f(\mathbf{X}) = X_1$ when $\mathbf{X} \in \mathcal{P}_\alpha$. Since $\mathcal{P}_\alpha$ contains 99.9% of the data, the mean squared error of $f(\mathbf{X})$ is very small (of the order $O(10^{-3})$).

Results. Using the data generating mechanism described above, we generate datapoints $\{\mathbf{x}^{(i)}\}_{i=1}^{500}$, and compute Shapley values for these datapoints. In this example, we use the supervised approach in Section E to compute the conditional expectation for CES values, and we use the rejection sampling procedure in Section C.2 to estimate ManifoldShap values.

Figure 10 shows the top features according to different value functions for $d \in \{100, 200, 500\}$. The results show that, IS values attribute greater importance to feature 2 for more than 90% of the datapoints, for all values of d under consideration, while the remaining datapoints have feature 1 as the most important feature. Similarly, both RJBShap and CES attribute greatest importance to feature 1 for less than 50% of the datapoints. Moreover, the importance that CES attributes to feature 1 decreases with increasing dimensions. Intuitively, this happens because the number of features highly correlated with feature 1 increases as d increases, leading CES to divide the attributions among increasing number of features. Among all the baselines considered, ManifoldShap remains closest to the ground truth Shapley values as it attributes greatest importance to feature 1 for more than 80% of the datapoints even as d increases.

G.2.6 Sensitivity of ManifoldShap and RJBShap to density estimation errors

In this experiment we investigate the sensitivity of computed Shapley values using ManifoldShap and RJBShap, to increasing density estimation errors. Here, $\mathbf{X} \in \mathbb{R}^{50}$ and $\mathcal{Y} \subseteq \mathbb{R}$. Specifically, we use the same data generating mechanism and model used in Section G.2.5 with $d = 50$. We use VAEs to obtain a density estimate $\hat{p}(\mathbf{x})$, which is subsequently used to estimate RJBShap and $\mathcal{P}_\alpha$ for ManifoldShap estimation. We generate datapoints $\{\mathbf{x}^{(i)}\}_{i=1}^{500}$, and compute ManifoldShap and RJBShap values for these datapoints, using density estimates of differing quality obtained by training VAE for different number of epochs.

Table 1 shows the percentage of datapoints with feature 1 as the most important feature as per each value function, for different density estimates. We use the oracle density of $\mathbf{X}$ to estimate density mean squared error in table 1. It can be seen that ManifoldShap is significantly less sensitive to density estimation errors as compared to RJBShap. This is because

Table 1: The percentage of datapoints with feature 1 as the most important feature as per each value function, for different density estimates obtained by training VAE for different number of epochs.

No. of epochs	10	50	200	Oracle density
ManifoldShap	**79.9**	**78.5**	**80.2**	**81.0**
RJBShap	17.5	15.0	10.0	13.1
Density MSE	395.1	220.0	63.7	0.0

ManifoldShap only depends on the density estimate via the indicator $\mathbb{1}(\hat{p}(\mathbf{x}) \geq \epsilon^{(\alpha)})$, whereas RJBShap depends on the density explicitly.