“Why did the Model Fail?”: Attributing Model Performance Changes to Distribution Shifts

Haoran Zhang [*1] Harvineet Singh [*2] Marzyeh Ghassemi [1] Shalmali Joshi [3]

Abstract

Machine learning models frequently experience performance drops under distribution shifts. The underlying cause of such shifts may be multiple simultaneous factors such as changes in data quality, differences in specific covariate distributions, or changes in the relationship between label and features. When a model does fail during deployment, attributing performance change to these factors is critical for the model developer to identify the root cause and take mitigating actions. In this work, we introduce the problem of attributing performance differences between environments to distribution shifts in the underlying data generating mechanisms. We formulate the problem as a cooperative game where the players are distributions. We define the value of a set of distributions to be the change in model performance when only this set of distributions has changed between environments, and derive an importance weighting method for computing the value of an arbitrary set of distributions. The contribution of each distribution to the total performance change is then quantified as its Shapley value. We demonstrate the correctness and utility of our method on synthetic, semi-synthetic, and real-world case studies, showing its effectiveness in attributing performance changes to a wide range of distribution shifts.

1. Introduction

Machine learning models are widely deployed in dynamic environments ranging from recommendation systems to personalized clinical care. Such environments are prone to distribution shifts, which may lead to serious degradations in model performance (Guo et al., 2022; Chirra et al., 2018; Koh et al., 2021; Geirhos et al., 2020; Nestor et al., 2019; Yang et al., 2023). Importantly, such shifts are hard to anticipate and reduce the ability of model developers to design reliable systems.

When the performance of a model *does* degrade during deployment, it is crucial for the model developer to know not only which distributions have shifted, but also *how much* a specific distribution shift contributed to model performance degradation. Using this information, the model developer can then take mitigating actions such as additional data collection, data augmentation, and model retraining (Ashmore et al., 2021; Zenke et al., 2017; Subbaswamy et al., 2019).

In this work, we present a method to attribute changes in model performance to shifts in a given set of distributions. Distribution shifts can occur in various marginal or conditional distributions that comprise variables involved in the model. Further, multiple distributions can change simultaneously. We handle this in our framework by defining the effect of changing any set of distributions on model performance, and use the concept of Shapley values (Shapley et al., 1953) to attribute the change to individual distributions. The Shapley value is a co-operative game theoretic framework with the goal of distributing surplus generated by the players in the co-operative game according to their contribution. In our framework, the players correspond to individual distributions, or more precisely, mechanisms involved in the data generating process.

Most relevant to our contributions is the work of Budhathoki et al. (2021), which attributes a shift between two joint distributions to a specific set of individual distributions. The distributions here correspond to the components of the factorization of the joint distribution when the data-generating process is assumed to follow causal structural assumptions. This line of work defines distribution shifts as interventions on causal mechanisms (Pearl & Bareinboim, 2011; Subbaswamy et al., 2019; 2021; Budhathoki et al., 2021; Thams et al., 2022). We build on their framework to justify the choice of players in our cooperative game. We significantly differ from the end goal by attributing a change in *model performance between two environments* to individual distributions. Note that each shifted distribution may influence model performance differently and may result

[*]Equal contribution [1]MIT [2]New York University [3]Columbia University. Correspondence to: Haoran Zhang <haoranz@mit.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

in significantly different attributions than their contributions to the shift in the joint distribution between environments.

In this work, we focus on explaining the discrepancy in model performance between two environments as measured by some metric such as prediction accuracy. We emphasize the non-trivial nature of this problem, as many distribution shifts will have no impact on a particular model or metric, and some distribution shifts may even increase model performance. Moreover, the root cause of the performance change may be due to distribution shifts in variables external to the model input. Thus, explaining performance discrepancy requires us to develop specialized methods. Specifically, we want to quantify the contribution to the performance change of a fixed set of distributions that may change across the environments. Given such a set, we develop a model-free importance sampling approach to quantify this contribution. We then use the Shapley value framework to estimate the attribution for each distribution shift. This framework allows us to expand the settings where our method is applicable.

We make the following contributions[1]:

- We formalize the problem of attributing model performance changes due to distribution shifts.
- We propose a principled approach based on Shapley values for attribution, and show that it satisfies several desirable properties.
- We validate the correctness and utility of our method on synthetic and real-world datasets.

2. Problem Setup

Notation. Consider a learning setup where we have some system variables denoted by V consisting of two types of variables $V = (X, Y)$, which comprises of features X and labels Y such that $V \sim \mathcal{D}$. Realizations of the variables are denoted in lower case. We assume access to samples from two environments. We use $\mathcal{D}^{\text{source}}$ to denote the source distribution and $\mathcal{D}^{\text{target}}$ for the target distribution. Subscripts on $\mathcal{D}$ refer to the distribution of specific variables. For example, $\mathcal{D}_{X_1}$ is the distribution of feature $X_1 \subset X$, and $\mathcal{D}_{Y|X}$ is the conditional distribution of labels given all features X.

Let $X_{\mathtt{M}} \subseteq X$ be the subset of features utilized by a given model f. We are given a loss function $\ell((x, y), f) \mapsto \mathbb{R}$ which assigns a real value to the model evaluated at a specific setting x of the variables. For example, in the case of supervised learning, the model f maps $X_{\mathtt{M}}$ into the label space, and a loss function such as the squared error $\ell((x, y), f) := (y - f(x_{\mathtt{M}}))^2$ can be used to evaluate model performance. We assume that the loss function can be computed separately for each data point. Then, performance of the model in some environment with distribution $\mathcal{D}$ is summarized by the average of the losses:

$$\text{Perf}(\mathcal{D}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell((x, y), f)]$$

This implies that a shift in any variables V in the system may result in performance change across environments, including those that are not directly used by the model, but drive changes to the features $X_{\mathtt{M}}$ used by the model for learning.

Setup. Suppose we are given a *candidate set* of (marginal and/or conditional) distributions $\mathbb{C}_{\mathcal{D}}$ over V that may account for the model performance change from $\mathcal{D}^{\text{source}}$ to $\mathcal{D}^{\text{target}}$: $\text{Perf}(\mathcal{D}^{\text{target}}) - \text{Perf}(\mathcal{D}^{\text{source}})$. **Our goal is to attribute this change to each distribution in the candidate set $\mathbb{C}_{\mathcal{D}}$.** For our method, we assume access to the model f, and samples from $\mathcal{D}^{\text{source}}$ as well as $\mathcal{D}^{\text{target}}$ (see Figure 1).

We assume that dependence between variables V is described by a causal system (Pearl, 2009). For every variable $X_i \in V$, this dependence is captured by a functional relationship between X_i and the so-called "causal parents" of X_i (denoted as $\text{parent}(X_i)$) driving the variation in X_i. The causal dependence induces a Markov distribution over the variables in this system. That is, the joint distribution $\mathcal{D}_V$ can be factorized as, $\mathcal{D}_V = \prod_{X_i \in V} \mathcal{D}_{X_i|\text{parent}(X_i)}$. This dependence can be summarized graphically using a Directed Acyclic Graph (DAG) with nodes corresponding to the system variables and directed edges ($\text{parent}(X_i) \rightarrow X_i$) in the direction of the causal mechanisms in the system (see Figure 1 for an example).

Example. We provide an example that illustrates that the performance attribution problem is ill-specified without knowing how the mechanisms can change to result in the observed performance difference. Suppose we are predicting Y from X with a linear model $f(x) := \phi x$ under the squared loss function. Consider two possible scenarios for data generation – (1) $X \leftarrow Y$ where $\mathcal{D}_Y$ changes from source to target while $\mathcal{D}_{X|Y}$ remains the same, (2) $X \rightarrow Y$ where $\mathcal{D}_X$ changes from source to target while $\mathcal{D}_{Y|X}$ remains the same. The performance difference of $f(x)$ is the same in both the cases. Naturally, we want an attribution method to assign all of the difference to the mechanism for Y in the first case and to the mechanism of X in the second case. Thus, for the same performance difference between source and target data, we would like a method to output different attributions depending on whether the data generating process is case (1) or (2). Note that, in general, it is impossible to find the appropriate attributions by first finding the direction of the causal mechanisms. This follows from the fact that learning the structure is in general, impossible purely from observational data (Peters et al., 2017). Hence knowledge of the data-generating mechanisms is necessary for appropriate attribution.

More concretely, suppose the processes are (1) $Y \sim$

[1] Code: https://github.com/MLforHealth/expl_perf_drop

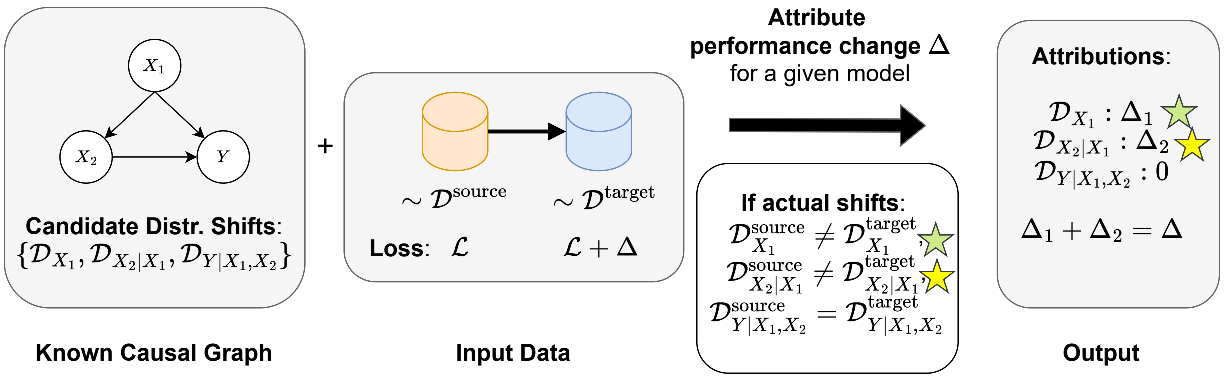

X_1
X_2
Y
Candidate Distr. Shifts:
$\{\mathcal{D}_{X_1}, \mathcal{D}_{X_2|X_1}, \mathcal{D}_{Y|X_1,X_2}\}$
Known Causal Graph
+
$\sim \mathcal{D}^{\text{source}}$
$\sim \mathcal{D}^{\text{target}}$
Loss: $\mathcal{L}$
$\mathcal{L} + \Delta$
Input Data
Attribute performance change Δ for a given model
If actual shifts:
$\mathcal{D}^{\text{source}}_{X_1} \neq \mathcal{D}^{\text{target}}_{X_1}$,
$\mathcal{D}^{\text{source}}_{X_2|X_1} \neq \mathcal{D}^{\text{target}}_{X_2|X_1}$,
$\mathcal{D}^{\text{source}}_{Y|X_1,X_2} = \mathcal{D}^{\text{target}}_{Y|X_1,X_2}$
Attributions:
$\mathcal{D}_{X_1} : \Delta_1$
$\mathcal{D}_{X_2|X_1} : \Delta_2$
$\mathcal{D}_{Y|X_1,X_2} : 0$
$\Delta_1 + \Delta_2 = \Delta$
Output

Figure 1: **Inputs and outputs for attribution.** Input: Causal graph, where all variables are observed providing the candidate distribution shifts we consider. The goal is to attribute the model's performance change Δ between source and target distributions to these candidate distributions. Here, out of the three candidate distributions, the marginal distribution of X_1 and the conditional distribution of X_2 given X_1 change. Our method attributes changes to each one such that the attributions sum to the total performance change Δ. Note that nodes in the causal graph may be vector-valued, which allows our method to be used on high-dimensional data such as images.

$N(\mu_1, 1), X \sim Y + N(0,1)$. The mean of Y shifts to μ_2 in target, and (2) $X \sim N(\mu_1, 1), Y \sim X + N(0,1)$ where the mean of X shifts to μ_2 in target. For the model $f(x) := \phi x$, the performance difference Δ in both cases is $(1-\phi)^2(\mu_2^2 - \mu_1^2)$. This example illustrates the need for specifying how the mechanisms can shift from source to target to solve the attribution problem. In this work, we use partial causal knowledge, **in terms of the causal graph only**, to specify the data-generating mechanisms.

In general, this partial knowledge further allows us to identify potential shifts to consider. Specifically, the number of marginal and conditional shifts that can be defined over (X, Y) is exponential in the dimension of X. The factorization induced by the causal graph or equivalently knowledge of the data-generating mechanism reduces the space of possible shifts to consider for attribution. See Section 3 for additional advantages of using a causal framework.

3. Method

We now formalize our problem setup and motivate a game theoretic method for attributing performance changes to distributions over variable subsets (See Figure 1 for a summary). We proceed with the following Assumptions.

Assumption 3.1. The causal graph corresponding to the data-generating mechanism is known and all variables in the system are observed. Thus, the factorization of the joint distribution $\mathcal{D}_V$ is known.

Assumption 3.2. Distribution shifts of interest are due to (independent) shifts in one or more factors of $\mathcal{D}_V$.

Given these assumptions, we now describe our game theoretic formulation for attribution.

3.1. Game Theoretic Distribution Shift Attribution

We consider the set of candidate distributions $\mathsf{C}_{\mathcal{D}}$ as the *players* in our attribution game. A *coalition* of any subset of players determines the distributions that are allowed to shift (from their source domain distribution to the target domain distribution), keeping the rest fixed. The *value* for the coalition is the model performance change between the resulting distribution for the coalition and the training distribution. See Figure 2 for an overview of the method.

Value of a Coalition. Consider a coalition of distributions $\widetilde{\mathsf{C}} \subseteq \mathsf{C}_{\mathcal{D}}$. This coalition implies a joint distribution over system variables V, where members in the coalition contribute their target domain distribution, and non-members contribute their source domain distribution:

$$\widetilde{\mathcal{D}} = \underbrace{\left(\prod_{i:\mathcal{D}_{X_i|\text{parent}(X_i)} \in \widetilde{\mathsf{C}}} \mathcal{D}^{\text{target}}_{X_i|\text{parent}(X_i)}\right)}_{\text{Coalition}} \cdot \underbrace{\left(\prod_{i:\mathcal{D}_{X_i|\text{parent}(X_i)} \notin \widetilde{\mathsf{C}}} \mathcal{D}^{\text{source}}_{X_i|\text{parent}(X_i)}\right)}_{\text{Not in Coalition}} \tag{1}$$

The above factorization follows from Assumptions 3.1 and 3.2. Note that the coalition only consists of distribu-

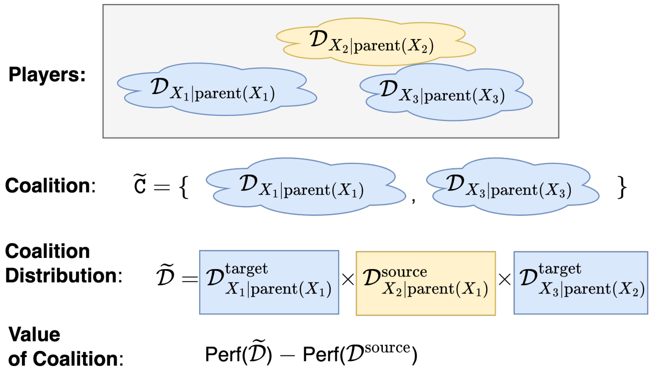

Players:
$\mathcal{D}_{X_2|\text{parent}(X_2)}$
$\mathcal{D}_{X_1|\text{parent}(X_1)}$
$\mathcal{D}_{X_3|\text{parent}(X_3)}$
Coalition: $\widetilde{\mathtt{C}} = \{ \mathcal{D}_{X_1|\text{parent}(X_1)}, \mathcal{D}_{X_3|\text{parent}(X_3)} \}$
Coalition Distribution: $\widetilde{\mathcal{D}} = \mathcal{D}^{\text{target}}_{X_1|\text{parent}(X_1)} \times \mathcal{D}^{\text{source}}_{X_2|\text{parent}(X_1)} \times \mathcal{D}^{\text{target}}_{X_3|\text{parent}(X_2)}$
Value of Coalition: $\text{Perf}(\widetilde{\mathcal{D}}) - \text{Perf}(\mathcal{D}^{\text{source}})$

Figure 2: **Sketch of the game theoretic attribution method.** Each causal mechanism is a player that, if present in the coalition, changes to the target distribution and, if absent, remains fixed at the source distribution. This defines the distribution of the resulting coalition $\widetilde{\mathcal{D}}$. Performance on $\widetilde{\mathcal{D}}$ is estimated using importance sampling from training data samples. After computing values for each possible coalition, Shapley value (Eq. 3) gives the attribution to each player. Thus, we estimate the performance change under all possible ways to shift the mechanisms from source to target and use these to distribute the total performance change among the individual distributions.

tions that are allowed to change across environments. All other relevant mechanisms are indeed fixed to the source distribution. We present an example of a coalition of two players in Figure 2. The value of the coalition $\widetilde{\mathtt{C}}$ with the coalition distribution $\widetilde{\mathcal{D}}$ is now given by

$$\text{Val}(\widetilde{\mathtt{C}}) := \text{Perf}(\widetilde{\mathcal{D}}) - \text{Perf}(\mathcal{D}^{\text{source}}) \tag{2}$$

Thus, our assumptions allow us to represent a factorization where only members of the coalition change, while all other mechanisms correspond to the source distribution. If we consider the change in performance for all combinatorial coalitions, we can estimate the total contribution of a specific distribution by aggregating the value for all possible coalitions a candidate distribution is a part of. This is exactly the Shapley value applied to a set of distributions. The Shapley value framework thus allows us to obtain the attribution of each player $d \in \mathtt{C}_{\mathcal{D}}$ using Equation 3.

Abstractly, the Shapley values framework (Shapley et al., 1953) is a game theoretic framework which assumes that there are $\mathtt{C} := \{1, 2, \ldots, n\}$ players in a co-operative game, achieving some total value (in our case, model performance change). We denote by $\text{Val} : 2^{\mathtt{C}} \mapsto \mathbb{R}$, the value for any subset of players, which is called a coalition. Shapley values correspond to the fair assignment of the value $\text{Val}(\mathtt{C})$ to each player $d \in \mathtt{C}$. The intuition behind Shapley values is to quantify the change in value when a player (here, a distribution) enters a coalition. Since the change in model performance depends on the order in which players (distributions) may join the coalition, Shapley values aggregate the value changes over all permutations of $\mathtt{C}$. Thus the Shapley attribution $\text{Attr}(d)$ for a player d is given by:

$$\text{Attr}(d) = \frac{1}{|\mathtt{C}|} \sum_{\widetilde{\mathtt{C}} \subseteq \mathtt{C} \setminus \{d\}} \binom{|\mathtt{C}|-1}{|\widetilde{\mathtt{C}}|}^{-1} \left(\text{Val}(\widetilde{\mathtt{C}} \cup \{d\}) - \text{Val}(\widetilde{\mathtt{C}}) \right) \tag{3}$$

where we measure the change in model performance (denoted by Val) after adding d to the coalition averaged over all potential coalitions involving d. The computational complexity of estimating Shapley values is exponential in the number of players. Hence we rely on this exact expression only when the number of candidate distributions is small. That is, the causal graph induces a factorization that results in smaller candidate sets. For larger candidate sets, we use previously proposed approximation methods (Castro et al., 2009; Lundberg & Lee, 2017; Janzing et al., 2020) for reduced computational effort.

Choice of Candidate Distribution Shifts. We motivate further the choice of candidate distributions that will inform the coalition. As mentioned before, without the knowledge of the causal graph, many heuristics for choosing the candidate sets are possible. For example, a candidate set could be the set of all marginal distributions on each system variable, $\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_{X_1}, \mathcal{D}_{X_2}, \cdots\}$, or distribution of each variable after conditioning on the rest, $\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_{X_1|V \setminus X_1}, \mathcal{D}_{X_2|V \setminus X_2}, \cdots\}$. Since we have combinatorially many shifts that can be defined on subsets of $V = (X, Y)$, choosing candidate sets that would then inform the coalition is challenging. The causal graph, on the other hand, specifies the factorization of the joint distribution into a set of distributions. We form the candidate set constituting each distribution in this factorization. That is,

$$\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_{X_1|\text{parent}(X_1)}, \cdots, \mathcal{D}_{X_i|\text{parent}(X_i)}, \cdots\}_{i=1,\cdots,|V|}$$

For a node without parents in the causal graph, the parent set can be empty, which reduces $\mathcal{D}_{X_i|\text{parent}(X_i)}$ to the marginal distribution of X_i. This choice of candidate set has three main advantages. First, it is *interpretable* since the candidate shifts are specified by domain experts who constructed the causal graph. Second, it is *actionable* since identifying the causal mechanisms most responsible for performance change can inform mitigating methods for handling distribution shifts (Subbaswamy et al., 2019). Third, it will lead to *succinct* attributions due to the independence property.

Consider the case where only one conditional distribution $\mathcal{D}(X_i|\text{parent}(X_i))$ changes across domains. This will result in a change in distributions of all descendants of X_i (due to the above factorization). In this case, a candidate set defined by all marginals is not succinct, as one would attribute performance changes to all marginals of descendants of X_i. Instead, the candidate set determined by the causal graph will isolate the correct conditional distribution.

Crucially, to compute our attributions, we need estimates of model performance under $\widetilde{\mathcal{D}}$. Note that we only have model performance estimates under $\mathcal{D}^{\text{source}}$ and $\mathcal{D}^{\text{target}}$, but not for any arbitrary coalition where only a subset of the distributions have shifted. To estimate the performance of any coalition, we propose to use importance sampling.

3.2. Importance Sampling to Estimate Performance under a Candidate Distribution Shift

Assumption 3.3. $\texttt{support}(\mathcal{D}^{\text{target}}_{X_i|\text{parent}(X_i)}) \subseteq \texttt{support}(\mathcal{D}^{\text{source}}_{X_i|\text{parent}(X_i)})$ for all $\mathcal{D}^{\text{target}}_{X_i|\text{parent}(X_i)} \in \mathtt{C}_{\mathcal{D}}$.

Importance sampling allows us to re-weight samples drawn from a given distribution, which can be $\mathcal{D}^{\text{source}}$ or $\mathcal{D}^{\text{target}}$, to simulate expectations for a desired distribution, which is the candidate $\widetilde{\mathcal{D}}$ in our case. Thus, we re-write the value as

$$\begin{aligned}\text{Val}(\widetilde{\mathtt{C}}) &= \text{Perf}(\widetilde{\mathcal{D}}) - \text{Perf}(\mathcal{D}^{\text{source}}) \qquad (4)\\ &= \mathbb{E}_{(x,y)\sim\widetilde{\mathcal{D}}}[\ell((x,y),f)] - \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{source}}}[\ell((x,y),f)]\\ &= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{source}}}\left[\frac{\widetilde{\mathcal{D}}((x,y))}{\mathcal{D}^{\text{source}}((x,y))}\ell((x,y),f)\right] -\\ &\qquad \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{source}}}[\ell((x,y),f)]\end{aligned}$$

The importance weights are themselves a product of ratios of source and target distributions corresponding to the causal mechanisms in $\mathtt{C}_{\mathcal{D}}$ as follows:

$$\begin{aligned}w_{\widetilde{\mathtt{C}}}((x,y)) &:= \frac{\widetilde{\mathcal{D}}((x,y))}{\mathcal{D}^{\text{source}}((x,y))} = \prod_{d\in\widetilde{\mathtt{C}}}\frac{\mathcal{D}^{\text{target}}_d((x,y))}{\mathcal{D}^{\text{source}}_d((x,y))}\\ &=: \prod_{d\in\widetilde{\mathtt{C}}} w_d((x,y))\end{aligned} \qquad (5)$$

By Assumption 3.3, we ensure that all importance weights are finite.

Computing Importance Weights. There are multiple ways to estimate importance weights $w_d((x,y))$, which are a ratio of densities (Sugiyama et al., 2012). Here, we use a simple approach for density ratio estimation via training probabilistic classifiers as described in Sugiyama et al. (2012, Section 2.2).

Let D be a binary random variable, such that when $D = 1, Z \sim \mathcal{D}^{\text{target}}_d(Z)$, and when $D = 0, Z \sim \mathcal{D}^{\text{source}}_d(Z)$. Suppose $d = \mathcal{D}_{X_i|\text{parent}(X_i)}$, then

$$w_d = \frac{\mathbb{P}(D=0|\text{parent}(X_i))}{\mathbb{P}(D=1|\text{parent}(X_i))}\cdot\frac{\mathbb{P}(D=1|X_i,\text{parent}(X_i))}{\mathbb{P}(D=0|X_i,\text{parent}(X_i))},$$

where each term is computed using a probabilistic classifier trained to discriminate data points from $\mathcal{D}^{\text{source}}$ and $\mathcal{D}^{\text{target}}$ from the concatenated dataset. We show the derivation of this equation in Appendix A. In total, we need to learn $\mathcal{O}(|\mathtt{C}_{\mathcal{D}}|)$ models for computing all importance weights.

3.3. Properties of Our Method

Under perfect computation of importance weights, the Shapley attributions resulting from the performance-change game have the following desirable properties, which follow from the properties of Shapley values. We provide proofs of these properties in Appendix B.

Property 1. (Efficiency) $\sum_{d\in\mathtt{C}_{\mathcal{D}}} \text{Attr}(d) = \text{Val}(\mathtt{C}_{\mathcal{D}}) = \text{Perf}(\mathcal{D}^{\text{target}}) - \text{Perf}(\mathcal{D}^{\text{source}})$

Property 2.1. (Null Player) $\mathcal{D}^{\text{source}}_d = \mathcal{D}^{\text{target}}_d \implies \text{Attr}(d) = 0$.

Property 2.2. (Relevance) Consider a mechanism d. If $\text{Perf}(\widetilde{\mathtt{C}} \cup \{d\}) = \text{Perf}(\widetilde{\mathtt{C}})$ for all $\widetilde{\mathtt{C}} \subseteq \mathtt{C}_{\mathcal{D}} \setminus d$, then $\text{Attr}(d) = 0$.

Property 3. (Attribution Symmetry) Let $\text{Attr}_{\mathcal{D}_1,\mathcal{D}_2}(d)$ denote the attribution to some mechanism d when $\mathcal{D}_1 = \mathcal{D}^{\text{source}}$ and $\mathcal{D}_2 = \mathcal{D}^{\text{target}}$. Then, $\text{Attr}_{\mathcal{D}_1,\mathcal{D}_2}(d) = -\text{Attr}_{\mathcal{D}_2,\mathcal{D}_1}(d)\ \forall d \in \mathtt{C}_{\mathcal{D}}$.

Thus, the method attributes the overall performance change only to distributions that actually change in a way that affects the specified performance metric. The contribution of each distribution is computed by considering how much they impact the performance if they are made to change in different combinations alongside the other distributions.

3.4. Analysis using a Synthetic Setting

We derive analytical expressions for attributions in a simple synthetic case with the following data generating process.

$$\begin{aligned}\text{Source}: X &\sim \mathcal{N}(\mu_1, \sigma_X^2)\\ Y &\sim \theta_1 X + \mathcal{N}(0, \sigma_Y^2)\\ \text{Target}: X &\sim \mathcal{N}(\mu_2, \sigma_X^2)\\ Y &\sim \theta_2 X + \mathcal{N}(0, \sigma_Y^2)\end{aligned}$$

The model that we are investigating is $f(X) = \phi X$, and $l((x,y),f) = (y - f(x))^2$.

We show the attribution of our method, along with the attribution using the joint method from Budhathoki et al. (2021), in Table 1. The complete derivation, along with experimental verification of the derived expressions, can be found in Appendix C. We highlight several advantages that our method has over the baseline.

First, our attribution takes the model parameter ϕ into account in order to explain model performance changes,

Table 1: Analytical expressions of the attributions for the synthetic case described in Section 3.4. For the full derivation, see Appendix C.

	Attr($\mathcal{D}_X$)	**Attr($\mathcal{D}_{Y\|X}$)**
Ours	$(\frac{1}{2}\mu_2^2 - \frac{1}{2}\mu_1^2)((\theta_2 - \phi)^2 + (\theta_1 - \phi)^2)$	$(\sigma_X^2 + \frac{1}{2}\mu_1^2 + \frac{1}{2}\mu_2^2)((\theta_2 - \phi)^2 - (\theta_1 - \phi)^2)$
Budhathoki et al. (2021)	$\frac{(\mu_2-\mu_1)^2}{2\sigma_X^2}$	$\frac{(\theta_2-\theta_1)^2}{2\sigma_Y^2}(\sigma_X^2 + \mu_2^2)$

whereas Budhathoki et al. (2021) do not, as they only explain shifts in (X, Y), or changes in simple functions such as $\mathbb{E}[X]$ of the variables. Second, we find that our Attr$(\mathcal{D}_X)$ is a function of θ_2. This is desirable, as covariate shift may compound with concept shift to increase loss non-linearly. This also ensures that both attributions always sum to the total shift. Third, we note that our attributions are *signed*, which is particularly important as some shifts may decrease loss. Finally, we note that our attributions are symmetric when the source and target data distributions are swapped by Property 3. This is not true of the baseline method in general, as the KL divergence is asymmetric. Since we assume knowledge of the true causal graph (which provides the factorization that determines the coalition), we also evaluate the attribution when the graph is misspecified. In this case, the coalition will consist of $\{\mathcal{D}_Y, \mathcal{D}_{X|Y}\}$. We include these attribution results in Appendix D.1, specifically, Figure C.2. In this case, as expected, both $\mathcal{D}_Y$ and $\mathcal{D}_{X|Y}$ are attributed the change in model performance (at varying levels depending on the magnitude of concept shift). While this may still be a meaningful attribution, knowledge of the causal graph provides a more succinct interpretation of system behavior.

4. Related Work

Identifying relevant distribution shifts. There has been extensive work that tests whether the data distribution has shifted (e.g. ones evaluated in Rabanser et al. (2019)). Past work has proposed to identify sub-distributions (factors constituting the joint distribution as determined by a generative model for the data) that comprise the shift between two joint distributions and order them by their contribution to the shift (Budhathoki et al., 2021). However, as suggested before, the sub-distributions may have different influence on model performance. Even a small change in some (factors) may have a large effect on model performance (and vice-versa). Thus, a model developer has to filter distributions to identify ones that actually impact model performance (see Property 2.2 and Appendix C). Further, Budhathoki et al. (2021) focuses on changes to the joint distribution as measured by the KL-divergence, which requires assumptions on the class of distributions to leverage closed-form expressions of KL-divergence (such as exponential families), or non-parametric KL estimation which is challenging in high dimensions (Wang et al., 2005; 2006).

Other approaches which aim to localize shifts to individual variables (conditional on the rest of the variables) do not provide a way to identify the ones relevant to performance (Kulinski et al., 2020). In contrast to testing for shifts, Podkopaev & Ramdas (2022) tests for changes in model performance when distribution changes in deployment. Recent work by Wu et al. (2021) decomposes performance change to changes in only marginal distributions using Shapley value framework (Lundberg & Lee, 2017). However, the method as described is restricted to categorical variables. Kulinski & Inouye (2022) propose a method for distribution shift explanation based on transport mappings, though their focus is still on how the distribution has shifted and not its impact on some downstream model. In parallel work, Cai et al. (2023) propose a method for attributing performance degradation to distribution shifts, but their decomposition is limited to $P(X)$ and $P(Y|X)$ and thus has limited granularity. Finally, Jung et al. (2022) propose a framework for measuring causal contributions to expected change in outcomes using Shapley values. This work focuses on attributions to specific feature *values*, by framing causal attribution as hard/do-interventions in a causal system. Further, they only examine the system in a single domain, and focus on the expected causal effect on a target variable Y. On the other hand, our work can be considered to focus on attributing model performance change to *mechanisms* by considering mechanism changes across domains as "soft-interventions". Further, our main goal is to attribute model performance change across two fixed domains where we have access to iid samples from both domains.

Shapley values for attribution. Shapley value-based attribution has recently become popular for interpreting model predictions (Štrumbelj & Kononenko, 2014; Lundberg & Lee, 2017; Wang et al., 2021). In most prior work, Shapley values have been leveraged for attributing a specific model prediction to the input features (Sundararajan & Najmi, 2020). Challenges to appropriately interpreting such attributions and desirable properties thereof have been extensively discussed in (Janzing et al., 2020; Kumar et al., 2021). In this work, we advance the use of Shapley values for interpreting model performance changes to individual distributions at the dataset level.

Detecting data partitions with low model performance. Recent work aims to find subsets of the dataset that have significantly worse (or better) performance (d'Eon et al., 2022; Eyuboglu et al., 2022; Jain et al., 2023; Park et al., 2023). However, they do not study changes in the underlying data distribution. The work by Ali et al. (2022) describes a method to identify and localize a change in model performance, and is applicable under distribution shifts. The main difference in our work is the data representations used for attribution. Instead of identifying subsets of *data* that are relevant to performance change, we find individual *distributions* represented by causal mechanisms.

5. Empirical Evaluation

Table 2: Datasets used to empirically evaluate our method.

Dataset	Modality	Ground Truth Known	Results Section
Synthetic	Tabular	✓	D.1
ColoredMNIST	Images	✓	D.2
CelebA	Images	✓	D.3
eICU	Tabular	✗	5.1
Camelyon-17	Images	✗	5.2

We empirically validate our method on five datasets, shown in Table 2. First, we validate the *correctness* of our method on three synthetic and semi-synthetic datasets where the ground truth shift(s) are known, and show that the baseline methods described below do not attain the correct attributions. Next, we demonstrate the *utility* of our method on two real-world datasets. Each attribution experiment was run on a cluster using 4 cores from an Intel Xeon Gold 5218R Processor and 16 GB of memory. We present two case studies on real-world data here, and the remaining results can be found in Appendix Section D.

Baselines. On datasets with known ground truth shifts (see Appendix Sections D.1, D.2) we evaluate the following baselines:

(a) **Misspecified or unknown causal graph.** When the causal graph is unknown, the user may create a causal graph based on intuition or causal discovery methods (Glymour et al., 2019). Here, we evaluate two simple mis-specified candidate shift sets. First, we evaluate the candidate shifts corresponding to all marginals (i.e. $\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_{X_1}, \cdots, \mathcal{D}_{X_i}, \cdots\}_{i=1,\cdots,|V|}$, which is similar to the method in Wu et al. (2021). Second, we evaluate the candidate shifts corresponding to a fully connected graph (i.e. $\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_{X_1|V\setminus X_1}, \cdots, \mathcal{D}_{X_i|V\setminus X_i}, \cdots\}_{i=1,\cdots,|V|}$).

(b) **KL-based attribution.** We evaluate the joint method from Budhathoki et al. (2021).

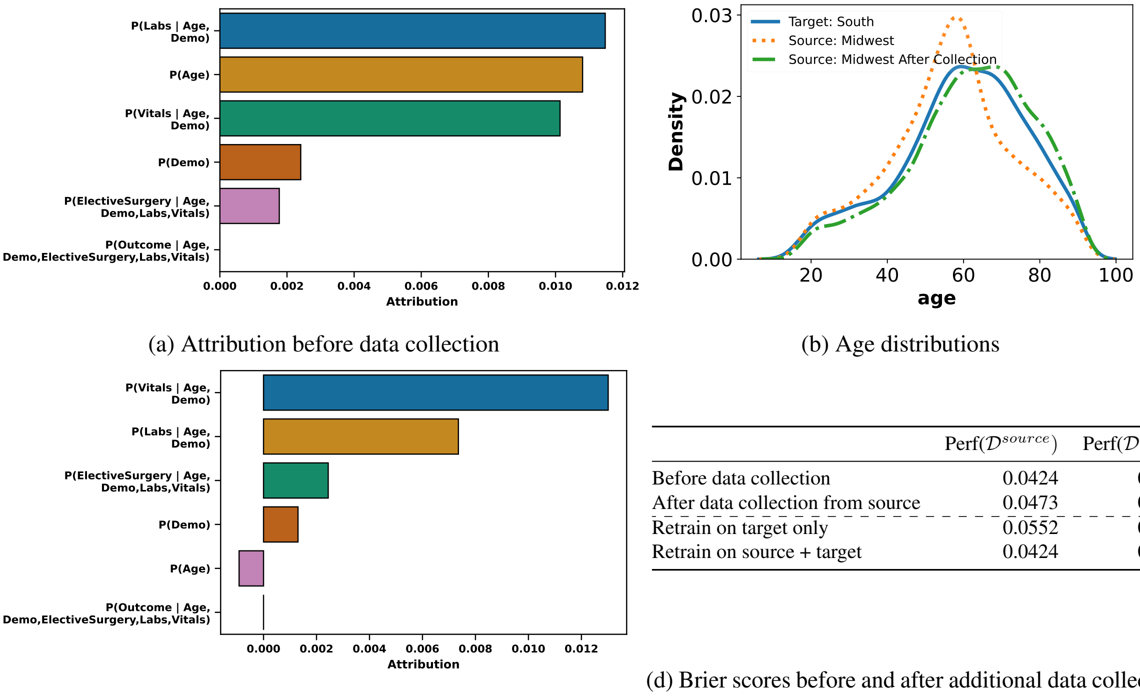
(c) **SHAP-based attribution.** We test a two-stage heuristic devised for this problem setting. First, we use a conditional independence test (Zhang et al., 2011) to find the distributions that are significantly different between source and target. Then, we run Kernel SHAP (Lundberg & Lee, 2017) on all samples in the target domain, taking the mean absolute value of the feature importance only for features that have significant shifts. To create attributions, we normalize these values to sum to the performance drop. Note that this method has several major flaws, namely that it can only attribute to shifts in input features to the model (and not system variables unused by the model), and cannot attribute to the *distribution* generating the target variable.

5.1. Case Study: Mortality Prediction in the ICU

Setup. Clinical machine learning models are being increasingly deployed in the real-world in hospitals, laboratories, and Intensive Care Units (ICUs) (Sendak et al., 2020). However, prior work has shown that such machine learning models are not robust to distribution shifts, and frequently degrade in performance on distributions different than what is seen during training (Singh et al., 2022). Here, we explore a simulated case study where a model which predicts mortality in the ICU is deployed in a different geographical region from where it is trained. We use data from the eICU Collaborative Research Database V2.0 (Pollard et al., 2018), which contains 200,859 de-identified ICU records for 208 hospitals across the United States. We simulate the deployment of a model trained on data from the Midwestern US (source) to the Southern US (target). We subset to 4 hospitals in each geography with the most number of samples. To mimic a realistic deployment scenario with limited sample size, we only observe 250 samples randomly selected from the target domain.

We learn an XGB (Chen & Guestrin, 2016) model to predict mortality given vitals, labs, and demographics data in the source domain. We assume the causal graph in Figure D.12, informed by prior work utilizing causal discovery on this dataset (Singh et al., 2022). As prior work has shown limited performance drops for models in this setting (Zhang et al., 2021), we subsample older population in the source environment to create an additional semi-synthetic distribution shift. We use our method to attribute the increase in Brier score from Midwest to South datasets.

Our method provides actionable attributions. First, we observe from our attributions (Figure 3a) that shift in the age distribution is responsible for 29.5% of the total shift (0.0108 of 0.0366). This confirms the validity of the attributions on a known semi-synthetic shift. Suppose that the practitioner decides to focus on mitigating the shift in age in order to improve target domain performance. To do so, they first plot the age distribution in the source and target environments (Figure 3b), finding that the target domain has dramatically more older patients. Then, they choose to col-



(a) Attribution before data collection

(b) Age distributions

(c) Attribution after data collection

	Perf($\mathcal{D}^{source}$)	Perf($\mathcal{D}^{target}$)	Perf Diff
Before data collection	0.0424	0.0790	0.0366
After data collection from source	0.0473	**0.0705**	**0.0232**
Retrain on target only	0.0552	0.0970	0.0417
Retrain on source + target	0.0424	0.0776	0.0351

(d) Brier scores before and after additional data collection, along with two naive retraining baselines

Figure 3: Attributing Brier score differences to candidate distributions on the eICU dataset for an XGB model trained on either (a) original or (c) age-balanced Midwest, and tested on the South domain.

lect additional data from the older population in the source. Training a new model on this augmented dataset, they find that the target domain performance improves by 10.8%, and the drop in performance is reduced by 36.6% (Figure 3d). In addition, this targeted mitigating action outperforms naively retraining the model on the merged datasets, or only on the target domain, due to the few target samples observed. Now that $\mathcal{D}_{\text{Age}}$ is no longer a significant factor in the performance drop across domains (Figure 3c), the practitioner may next turn their attention to mitigating shifts in more impactful conditional mechanisms such as $\mathcal{D}_{\text{Vitals}|\text{Age, Demo}}$, using methods such as GAN data augmentation (Mariani et al., 2018) or targeted importance weighting (Zhang et al., 2013), but we leave such explorations to future work.

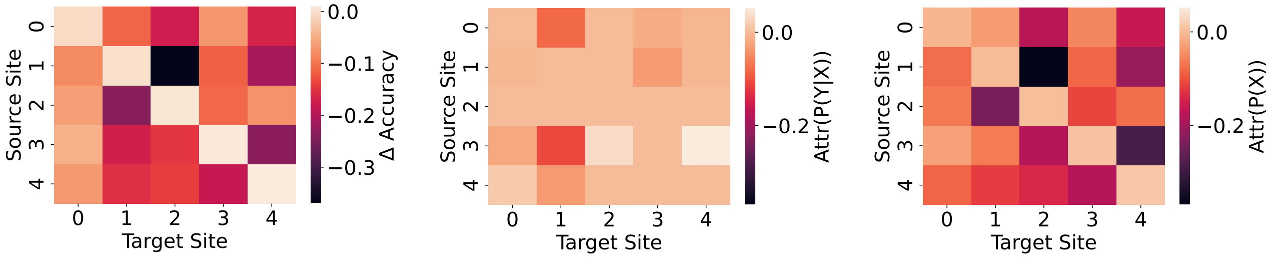
5.2. Case Study: Tumor Prediction from Camelyon17

We evaluate our method on the Camelyon17 dataset (Bandi et al., 2018; Koh et al., 2021), which consists of histopathology images from five hospitals, and the goal is to classify whether the central region contains any tumor tissue. We assume the causal setting (i.e. an $X \to Y$ causal graph) (Bandi et al., 2018), as labels are generated by pathologists from the image. Here, X is a vector-valued node for the image, which we represent using static features extracted from an ImageNet-pretrained ResNet-18. We train linear models on these representations to predict Y separately for each site. We use our method to attribute drops in accuracy of each model to each of the other four sites, to $P(X)$ and $P(Y|X)$.

Results. In Figure 4, we show the attributions from our method for each distribution, as well as the total accuracy drop. We find that our method attributes most of the performance drop to covariate shift $P(X)$ as opposed to concept shift $P(Y|X)$. This aligns with prior work showing that unsupervised domain adaptation methods improve domain robustness in this dataset (Wiles et al., 2022; Ginsberg et al., 2023). Using this result, a practitioner can apply targeted mitigating methods such as targeted data augmentation (Gao et al., 2022) or domain-adversarial training (Ganin et al., 2016).

6. Discussion

We develop a method to attribute changes in performance of a model deployed on a different distribution from the training distribution. We assume that distribution shifts are induced due to interventions in the causal mechanisms which result in model performance changes. We use the knowledge of the causal graph to formulate a game theoretic attribution framework using Shapley values. The coalition members are mechanisms contributing to the change in model performance. We demonstrate the correctness and



(a) Perf($\mathcal{D}^{target}$) − Perf($\mathcal{D}^{source}$)

(b) Attr($\mathcal{D}_{Y|X}$)

(c) Attr($\mathcal{D}_X$)

Figure 4: Attributions made by our method to domain shifts in Camelyon-17, using accuracy as the metric. We show (a) the total change in performance, (b) our attribution to $P(Y|X)$, (c) our attribution to $P(X)$.

utility of our method on synthetic, semi-synthetic, and real-world data.

Limitations and Future Work. Our work assumes knowledge of the causal graph to obtain interpretable and succinct attributions. When the causal graph is unknown, methods in causal discovery (Glymour et al., 2019) can produce a Markov equivalence class of causal graphs for tabular datasets, though these methods often have strict assumptions. While we may still be able to obtain reasonable attributions from a misspecified graph, we argue that such attributions may not be minimal. In addition, we observe some variance in the importance weighting estimates, which may potentially be remedied by using more advanced density estimation techniques (e.g. (Liu et al., 2021)). We note that our experiments on the CelebA dataset are for demonstration purposes only, and do not advocate for deployment of such models. Similarly, while we demonstrate case studies on publicly available health data, our work is only a proof of concept, and we recommend further evaluation before practical deployment. Future work includes relaxing the assumption that all variables are observed, comparing strategies for mitigating conditional shifts, and extending the experiments to additional settings such as unsupervised learning and reinforcement learning.

Acknowledgements

This work was supported in part by a grant from Quanta Computing. We would like to thank Taylor Killian and three anonymous reviewers for their valuable feedback. HS acknowledges support from the National Science Foundation under NSF Award 1922658. HS would like to thank Rumi Chunara and Vishwali Mhasawade for helpful discussions leading up to this work.

References

Ali, A., Cauchois, M., and Duchi, J. C. The lifecycle of a statistical model: Model failure detection, identification, and refitting, 2022. URL https://arxiv.org/abs/2202.04166.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Ashmore, R., Calinescu, R., and Paterson, C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, 54(5): 1–39, 2021.

Aubin, B., Słowik, A., Arjovsky, M., Bottou, L., and Lopez-Paz, D. Linear unit-tests for invariance discovery. *arXiv preprint arXiv:2102.10867*, 2021.

Bandi, P., Geessink, O., Manson, Q., Van Dijk, M., Balkenhol, M., Hermsen, M., Bejnordi, B. E., Lee, B., Paeng, K., Zhong, A., et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

Budhathoki, K., Janzing, D., Bloebaum, P., and Ng, H. Why did the distribution change? In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1666–1674. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/budhathoki21a.html.

Cai, T. T., Namkoong, H., and Yadlowsky, S. Diagnosing model performance under distribution shift, 2023. URL https://arxiv.org/abs/2303.02011.

Castro, J., Gómez, D., and Tejada, J. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Chirra, P., Leo, P., Yim, M., Bloch, B. N., Rastinehad, A. R., Purysko, A., Rosen, M., Madabhushi, A., and Viswanath, S. Empirical evaluation of cross-site reproducibility in radiomic features for characterizing prostate mri. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, pp. 105750B. International Society for Optics and Photonics, 2018.

d'Eon, G., d'Eon, J., Wright, J. R., and Leyton-Brown, K. The spotlight: A general method for discovering systematic errors in deep learning models. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 1962–1981, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533240. URL https://doi.org/10.1145/3531146.3533240.

Eyuboglu, S., Varma, M., Saab, K. K., Delbrouck, J.-B., Lee-Messer, C., Dunnmon, J., Zou, J., and Re, C. Domino: Discovering systematic errors with cross-modal embeddings. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=FPCMqjI0jXN.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

Gao, I., Sagawa, S., Koh, P. W., Hashimoto, T., and Liang, P. Out-of-distribution robustness via targeted augmentations. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022. URL https://openreview.net/forum?id=Bcg0It4i1g.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ginsberg, T., Liang, Z., and Krishnan, R. G. A learning based hypothesis test for harmful covariate shift. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=rdfgqiwz7lZ.

Glymour, C., Zhang, K., and Spirtes, P. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.

Guo, L. L., Pfohl, S. R., Fries, J., Johnson, A. E., Posada, J., Aftandilian, C., Shah, N., and Sung, L. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. *Scientific reports*, 12(1):1–10, 2022.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Jain, S., Lawrence, H., Moitra, A., and Madry, A. Distilling model failures as directions in latent space. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=99RpBVpLiX.

Janzing, D., Minorics, L., and Blöbaum, P. Feature relevance quantification in explainable ai: A causal problem. In *International Conference on artificial intelligence and statistics*, pp. 2907–2916. PMLR, 2020.

Johnson, A. E. W., Pollard, T. J., and Naumann, T. Generalizability of predictive models for intensive care unit patients, 2018. URL https://arxiv.org/abs/1812.02275.

Jung, Y., Kasiviswanathan, S., Tian, J., Janzing, D., Blöbaum, P., and Bareinboim, E. On measuring causal contributions via do-interventions. In *International Conference on Machine Learning*, pp. 10476–10501. PMLR, 2022.

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJE-4xW0W.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Kulinski, S. and Inouye, D. I. Towards explaining distribution shifts. *arXiv preprint arXiv:2210.10275*, 2022.

Kulinski, S., Bagchi, S., and Inouye, D. I. Feature shift detection: Localizing which features have shifted via conditional distribution tests. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 19523–19533. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/e2d52448d36918c575fa79d88647ba66-Paper.pdf.

Kumar, I., Scheidegger, C., Venkatasubramanian, S., and Friedler, S. Shapley residuals: Quantifying the limits of the shapley value for explanations. *Advances in Neural Information Processing Systems*, 34, 2021.

Liu, Q., Xu, J., Jiang, R., and Wong, W. H. Density estimation using deep generative neural networks. *Proceedings of the National Academy of Sciences*, 118(15): e2101344118, 2021.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

Mariani, G., Scheidegger, F., Istrate, R., Bekas, C., and Malossi, C. Bagan: Data augmentation with balancing gan. *arXiv preprint arXiv:1803.09655*, 2018.

Nestor, B., McDermott, M. B., Boag, W., Berner, G., Naumann, T., Hughes, M. C., Goldenberg, A., and Ghassemi, M. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, pp. 381–405. PMLR, 2019.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., and Madry, A. Trak: Attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*, 2023.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pearl, J. and Bareinboim, E. Transportability of causal and statistical relations: A formal approach. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Podkopaev, A. and Ramdas, A. Tracking the risk of a deployed model and detecting harmful distribution shifts. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Ro_zAjZppv.

Pollard, T. J., Johnson, A. E., Raffa, J. D., Celi, L. A., Mark, R. G., and Badawi, O. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018.

Rabanser, S., Günnemann, S., and Lipton, Z. Failing loudly: An empirical study of methods for detecting dataset shift. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/846c260d715e5b854ffad5f70a516c88-Paper.pdf.

Sagawa*, S., Koh*, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Sendak, M. P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., Ratliff, W., and Balu, S. A path for translation of machine learning products into healthcare delivery. *EMJ Innov*, 10:19–00172, 2020.

Shapley, L. S. et al. A value for n-person games. 1953.

Singh, H., Mhasawade, V., and Chunara, R. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. *PLOS Digital Health*, 1(4):e0000023, 2022.

Subbaswamy, A., Schulam, P., and Saria, S. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3118–3127, 2019.

Subbaswamy, A., Adams, R., and Saria, S. Evaluating model robustness and stability to dataset shift. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2611–2619. PMLR, 13–15 Apr 2021. URL http://proceedings.mlr.press/v130/subbaswamy21a.html.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., Von Bünau, P., and Kawanabe, M. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

Sugiyama, M., Suzuki, T., and Kanamori, T. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

Sundararajan, M. and Najmi, A. The many shapley values for model explanation. In *International conference on machine learning*, pp. 9269–9278. PMLR, 2020.

Thams, N., Oberst, M., and Sontag, D. Evaluating robustness to dataset shift via parametric robustness sets. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=OTKJttKN5c.

Štrumbelj, E. and Kononenko, I. 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.

Wang, J., Wiens, J., and Lundberg, S. Shapley flow: A graph-based approach to interpreting model predictions. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 721–729. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/wang21b.html.

Wang, Q., Kulkarni, S. R., and Verdú, S. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory*, 51 (9):3064–3074, 2005.

Wang, Q., Kulkarni, S. R., and Verdú, S. A nearest-neighbor approach to estimating divergence between continuous random vectors. In *2006 IEEE International Symposium on Information Theory*, pp. 242–246. IEEE, 2006.

Wang, Q., Kulkarni, S. R., and Verdú, S. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5):2392–2405, 2009.

Wiles, O., Gowal, S., Stimberg, F., Rebuffi, S.-A., Ktena, I., Dvijotham, K. D., and Cemgil, A. T. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Dl4LetuLdyK.

Wu, E., Wu, K., and Zou, J. Explaining medical ai performance disparities across sites with confounder shapley value analysis, 2021. URL https://arxiv.org/abs/2111.08168.

Yang, Y., Zhang, H., Katabi, D., and Ghassemi, M. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning*, 2023.

Zenke, F., Poole, B., and Ganguli, S. Continual learning through synaptic intelligence. In *International Conference on Machine Learning*, pp. 3987–3995. PMLR, 2017.

Zhang, H., Dullerud, N., Seyyed-Kalantari, L., Morris, Q., Joshi, S., and Ghassemi, M. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, pp. 279–290, 2021.

Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, UAI'11, pp. 804–813, Arlington, Virginia, USA, 2011. AUAI Press. ISBN 9780974903972.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *International conference on machine learning*, pp. 819–827. PMLR, 2013.

A. Derivation of Importance Weights

Let D be a binary random variable, such that when $D = 1, X \sim \mathcal{D}^{\text{target}}(X)$, and when $D = 0, X \sim \mathcal{D}^{\text{source}}(X)$. Suppose $d = \mathcal{D}_{X_i|\text{parent}(X_i)}$, then, for a particular value (x, y):

$$\begin{aligned}
\mathcal{D}_d^{\text{target}}((x,y)) &:= \mathbb{P}(X_i = x|\text{parent}(X_i) = \text{parent}(x_i), D = 1) \\
&= \frac{\mathbb{P}(D = 1, \text{parent}(X_i) = x_i|X_i = x_i) \cdot \mathbb{P}(X_i = x_i)}{\mathbb{P}(D = 1, \text{parent}(X_i) = x_i)} \\
&= \frac{\mathbb{P}(D = 1|\text{parent}(X_i) = x_i, X_i = x_i) \cdot \mathbb{P}(X_i = x_i, \text{parent}(X_i) = X_i)}{\mathbb{P}(D = 1|\text{parent}(X_i) = x_i) \cdot \mathbb{P}(\text{parent}(X_i) = x_i)}
\end{aligned}$$

Then,

$$\begin{aligned}
w_d &= \frac{\mathcal{D}_d^{\text{target}}((x,y))}{\mathcal{D}_d^{\text{source}}((x,y))} \\
&= \frac{\mathbb{P}(D = 0|\text{parent}(X_i) = \text{parent}(x_i))}{\mathbb{P}(D = 1|\text{parent}(X_i) = \text{parent}(x_i))} \cdot \frac{\mathbb{P}(D = 1|X_i = x_i, \text{parent}(X_i) = \text{parent}(x_i))}{\mathbb{P}(D = 0|X_i = x_i, \text{parent}(X_i) = \text{parent}(x_i))} \\
&= \frac{1 - \mathbb{P}(D = 1|\text{parent}(X_i) = \text{parent}(x_i))}{\mathbb{P}(D = 1|\text{parent}(X_i) = \text{parent}(x_i))} \cdot \frac{\mathbb{P}(D = 1|X_i = x_i, \text{parent}(X_i) = \text{parent}(x_i))}{1 - \mathbb{P}(D = 1|X_i = x_i, \text{parent}(X_i) = \text{parent}(x_i))}
\end{aligned}$$

Thus, we learn a model to predict D from X_i, and a model to predict D from $[X_i; \text{parent}(X_i)]$, on the concatenated dataset. In practice, we learn these models on a 75% split of both the source and target data, and use the remaining 25% for Shapley value computation, which only requires inference on the trained models. Therefore, an upper limit on the number of weight models required is $2|\mathbb{C}_{\mathcal{D}}|$, though in practice, this number is often smaller as several nodes may have the same parents.

In the case where X_i is a root node, the expression becomes:

$$w_d = \frac{1 - \mathbb{P}(D = 1)}{\mathbb{P}(D = 1)} \cdot \frac{\mathbb{P}(D = 1|X_i = x_i)}{1 - \mathbb{P}(D = 1|X_i = x_i)}$$

Where we simply compute $P(D = 1)$ as the relative size of the provided source and target datasets.

B. Proof of Properties

Property 1. (Efficiency) $\sum_{d \in \mathtt{C}_{\mathcal{D}}} \text{Attr}(d) = \text{Val}(\mathtt{C}_{\mathcal{D}}) = \text{Perf}(\mathcal{D}^{\text{target}}) - \text{Perf}(\mathcal{D}^{\text{source}})$

By the efficiency property of Shapley values (Shapley et al., 1953), we know that the sum of Shapley values equal the value of the all-player coalition. Thus, we distribute the total performance change due to the shift from source to target distribution to the shifts in causal mechanisms in the candidate set.

Property 2.1. (Null Player) $\mathcal{D}_d^{\text{source}} = \mathcal{D}_d^{\text{target}} \implies \text{Attr}(d) = 0.$

Property 2.2. (Relevance) Consider a mechanism d. If $\text{Perf}(\widetilde{\mathtt{C}} \cup \{d\}) = \text{Perf}(\widetilde{\mathtt{C}})$ for all $\widetilde{\mathtt{C}} \subseteq \mathtt{C}_{\mathcal{D}} \setminus d$, then $\text{Attr}(d) = 0$.

We can verify that our method gives zero attribution to distributions that do not shift between the source and target, and distribution shifts which do not impact model performance. First, we observe that in both cases, $\text{Val}(\widetilde{\mathcal{D}}) = \text{Val}(\widetilde{\mathcal{D}} \cup \{d\})$. For Property 2.1, this is because $\widetilde{\mathcal{D}} = \widetilde{\mathcal{D}} \cup \{d\}$ for any $\widetilde{\mathcal{D}} \subseteq \mathtt{C}_{\mathcal{D}}$ since the factor corresponding to d remains the same between source and target even when it is allowed to change as part of the coalition. For Property 2.2, this is clear from Eq. 4. By definition of Shapley value in Eq. 3, $\text{Attr}(d) = 0$.

Property 3. (Attribution Symmetry) Let $\text{Attr}_{\mathcal{D}_1,\mathcal{D}_2}(d)$ denote the attribution to some mechanism d when $\mathcal{D}_1 = \mathcal{D}^{\text{source}}$ and $\mathcal{D}_2 = \mathcal{D}^{\text{target}}$. Then, $\text{Attr}_{\mathcal{D}_1,\mathcal{D}_2}(d) = -\text{Attr}_{\mathcal{D}_2,\mathcal{D}_1}(d)\ \forall d \in \mathtt{C}_{\mathcal{D}}$.

We overload $\text{Perf}_{src \to tar}(\widetilde{\mathtt{C}})$ for some coalition $\widetilde{\mathtt{C}}$ to denote $\text{Perf}(\widetilde{\mathcal{D}})$ where $\widetilde{\mathcal{D}}$ is given by Equation 1. Analogously, we denote $\text{Perf}_{tar \to src}(\widetilde{\mathtt{C}})$ to be $\text{Perf}(\widetilde{\mathcal{D}}')$ when $\widetilde{\mathcal{D}}'$ is given by

$$\widetilde{\mathcal{D}}' = \left(\prod_{i:\mathcal{D}_{X_i|\text{parent}(X_i)} \in \widetilde{\mathtt{C}}} \mathcal{D}^{\text{source}}_{X_i|\text{parent}(X_i)} \right) \left(\prod_{i:\mathcal{D}_{X_i|\text{parent}(X_i)} \notin \widetilde{\mathtt{C}}} \mathcal{D}^{\text{target}}_{X_i|\text{parent}(X_i)} \right)$$

Note that $\text{Perf}_{src \to tar}(\widetilde{\mathtt{C}})$ = $\text{Perf}_{tar \to src}(\mathtt{C}_{\mathcal{D}} \setminus \widetilde{\mathtt{C}})$ for all $\widetilde{\mathtt{C}} \subseteq \mathtt{C}_{\mathcal{D}}$.

We can use Equation 2 to rewrite Equation 3 as:

$$\begin{aligned}
\text{Attr}_{\mathcal{D}_1,\mathcal{D}_2}(d) &= \frac{1}{|\mathtt{C}_{\mathcal{D}}|} \sum_{\widetilde{\mathtt{C}} \subseteq \mathtt{C}_{\mathcal{D}} \setminus \{d\}} \binom{|\mathtt{C}_{\mathcal{D}}| - 1}{|\widetilde{\mathtt{C}}|}^{-1} \left(\text{Perf}_{src \to tar}(\widetilde{\mathtt{C}} \cup \{d\}) - \text{Perf}_{src \to tar}(\widetilde{\mathtt{C}}) \right) \\
&= \frac{-1}{|\mathtt{C}_{\mathcal{D}}|} \sum_{\widetilde{\mathtt{C}} \subseteq \mathtt{C}_{\mathcal{D}} \setminus \{d\}} \binom{|\mathtt{C}_{\mathcal{D}}| - 1}{|\widetilde{\mathtt{C}}|}^{-1} \left(\text{Perf}_{tar \to src}(\mathtt{C}_{\mathcal{D}} \setminus \widetilde{\mathtt{C}}) - \text{Perf}_{tar \to src}(\mathtt{C}_{\mathcal{D}} \setminus (\widetilde{\mathtt{C}} \cup \{d\})) \right) \\
&= \frac{-1}{|\mathtt{C}_{\mathcal{D}}|} \sum_{\widetilde{\mathtt{C}}' \subseteq \mathtt{C}_{\mathcal{D}} \setminus \{d\}} \binom{|\mathtt{C}_{\mathcal{D}}| - 1}{|\widetilde{\mathtt{C}}'|}^{-1} \left(\text{Perf}_{tar \to src}(\widetilde{\mathtt{C}}' \cup \{d\}) - \text{Perf}_{tar \to src}(\widetilde{\mathtt{C}}') \right) \\
&= -\text{Attr}_{\mathcal{D}_2,\mathcal{D}_1}(d)
\end{aligned}$$

C. Shapley Values for A Synthetic Setting

C.1. Derivation

Suppose that we have the following data generating process for the source environment:

$$\begin{aligned} X &\sim \mathcal{N}(\mu_1, \sigma_X^2) \\ Y &\sim \theta_1 X + \mathcal{N}(0, \sigma_Y^2) \end{aligned}$$

And for the target environment:

$$\begin{aligned} X &\sim \mathcal{N}(\mu_2, \sigma_X^2) \\ Y &\sim \theta_2 X + \mathcal{N}(0, \sigma_Y^2) \end{aligned}$$

The model that we are investigating is $\hat{Y} = f(X) = \phi X$, and $l((x,y), f) = (y - f(x))^2$. Then,

$$\begin{aligned} \text{Perf}(\mathcal{D}^{\text{source}}) &= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{source}}}[l((x,y),f)] \\ &= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{source}}}[\big(\theta_1 X + \mathcal{N}(0,\sigma_Y^2) - \phi X\big)^2] \\ &= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{source}}}[\big(\mathcal{N}((\theta_1-\phi)\mu_1, (\theta_1-\phi)^2\sigma_X^2) + \mathcal{N}(0,\sigma_Y^2)\big)^2] \\ &= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{source}}}[\big(\mathcal{N}((\theta_1-\phi)\mu_1, (\theta_1-\phi)^2\sigma_X^2 + \sigma_Y^2)\big)^2] \\ &= (\theta_1-\phi)^2\sigma_X^2 + \sigma_Y^2 + (\theta_1-\phi)^2\mu_1^2 \end{aligned}$$

$$\begin{aligned} \text{Perf}(\mathcal{D}^{\text{target}}) &= \mathbb{E}_{(x,y)\sim\mathcal{D}^{\text{target}}}[l((x,y),f)] \\ &= (\theta_2-\phi)^2\sigma_X^2 + \sigma_Y^2 + (\theta_2-\phi)^2\mu_2^2 \end{aligned}$$

$$\begin{aligned} \Delta &= \text{Perf}(\mathcal{D}^{\text{target}}) - \text{Perf}(\mathcal{D}^{\text{source}}) \\ &= \sigma_X^2((\theta_2-\phi)^2 - (\theta_1-\phi)^2) + (\theta_2-\phi)^2\mu_2^2 - (\theta_1-\phi)^2\mu_1^2 \\ &= \text{Val}(\mathbb{C}_{\mathcal{D}}) \end{aligned}$$

$$\begin{aligned} \text{Val}(\{\mathcal{D}_X\}) &= (\theta_1-\phi)^2(\mu_2^2 - \mu_1^2) && (\theta_2 := \theta_1) \\ \text{Val}(\{\mathcal{D}_{Y|X}\}) &= (\sigma_X^2 + \mu_1^2)((\theta_2-\phi)^2 - (\theta_1-\phi)^2) && (\mu_2 := \mu_1) \end{aligned}$$

$$\begin{aligned} \text{Attr}(\mathcal{D}_X) &= \frac{1}{2}\left(\text{Val}(\mathbb{C}_{\mathcal{D}}) - \text{Val}(\{\mathcal{D}_{Y|X}\}) + \text{Val}(\{\mathcal{D}_X\}) - \text{Val}(\{\})\right) \\ &= \frac{1}{2}\left((\theta_2-\phi)^2(\mu_2^2-\mu_1^2) + (\theta_1-\phi)^2(\mu_2^2-\mu_1^2)\right) \\ &= (\frac{1}{2}\mu_2^2 - \frac{1}{2}\mu_1^2)((\theta_2-\phi)^2 + (\theta_1-\phi)^2) \end{aligned}$$

$$\begin{aligned} \text{Attr}(\mathcal{D}_{Y|X}) &= \frac{1}{2}\left(\text{Val}(\mathbb{C}_{\mathcal{D}}) - \text{Val}(\{\mathcal{D}_X\}) + \text{Val}(\{\mathcal{D}_{Y|X}\}) - \text{Val}(\{\})\right) \\ &= \frac{1}{2}\left((\sigma_X^2+\mu_2^2)((\theta_2-\phi)^2 - (\theta_1-\phi)^2) + (\sigma_X^2+\mu_1^2)((\theta_2-\phi)^2 - (\theta_1-\phi)^2)\right) \\ &= (\sigma_X^2 + \frac{1}{2}\mu_1^2 + \frac{1}{2}\mu_2^2)((\theta_2-\phi)^2 - (\theta_1-\phi)^2) \end{aligned}$$

Note that $\text{Attr}(\mathcal{D}_X) + \text{Attr}(\mathcal{D}_{Y|X}) = \Delta$.

Using the method proposed by Budhathoki et al. (2021), we get that:

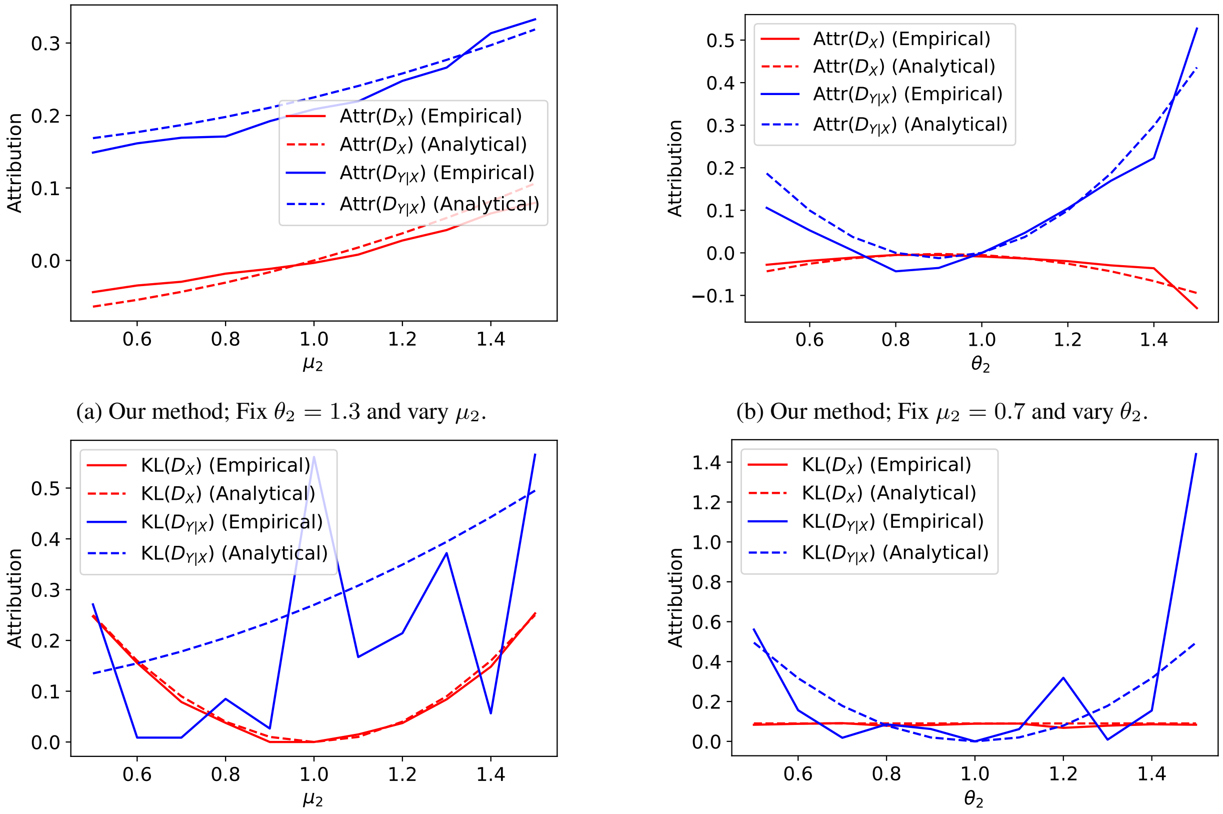
$$\begin{aligned} D(\tilde{P}_X || P_X) &= \frac{(\mu_2 - \mu_1)^2}{2\sigma_X^2} \\ D(\tilde{P}_{Y|X} || P_{Y|X}) &= \mathbb{E}_{X \sim \tilde{P}_X}[D(\tilde{P}_{Y|X=x} || P_{Y|X=x})] \\ &= \mathbb{E}_{X \sim \tilde{P}_X}\left[\frac{((\theta_2 - \theta_1)X)^2}{2\sigma_Y^2}\right] = \frac{(\theta_2 - \theta_1)^2}{2\sigma_Y^2}(\sigma_X^2 + \mu_2^2) \end{aligned}$$

C.2. Experiments

Now, we verify the correctness of our method by conducting a simulation of this setting, using $\mu_1 = 0$, $\theta_1 = 1$, $\sigma_X^2 = 0.5$, $\sigma_Y^2 = 0.25$, $\phi = 0.9$, and varying μ_2 (the level of covariate shift), and θ_2 (the level of concept drift). We generate $10,000$ samples from the source environment, and, for each setting of μ_2 and θ_2, we generate $10,000$ samples from the corresponding target environment. We then apply our method to attribute shifts to $\{\mathcal{D}_X, \mathcal{D}_{Y|X}\}$, using XGB to estimate importance weights. We also apply the joint method in Budhathoki et al. (2021).

In Figure C.1, we compare our attributions with the baseline, when both covariate and concept drift are present. We find that for our method, the empirical results match with the previously derived analytical expressions, where any deviations can be attributed to variance in the importance weight computations. For Budhathoki et al. (2021), we find that there appears to be very high variance in the attribution the attribution to $\mathcal{D}_{Y|X}$, which is likely a product of the nearest-neighbors KL estimator (Wang et al., 2009) used in their work.

In Figure C.2, we explore the case where we have a misspecified causal graph. Specifically, we examine the case where only concept drift is present, for the actual graphical model ($\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_X, \mathcal{D}_{Y|X}\}$), and for a misspecified graphical model ($\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_Y, \mathcal{D}_{X|Y}\}$). We find that using the mechanisms from the true data generating process results in a *minimal* attribution (i.e. Attr$(\mathcal{D}_X) = 0$), whereas the the misspecified causal graph gives non-zero attribution to both distributions.



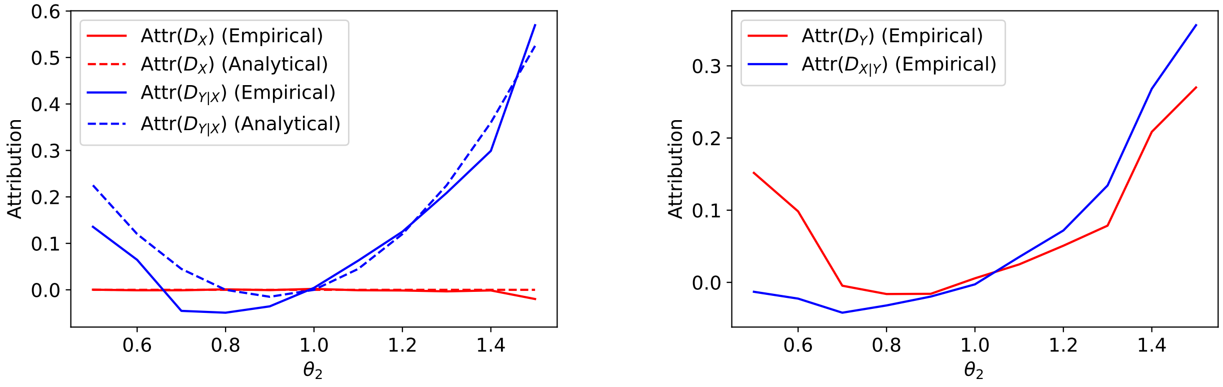
(a) Our method; Fix $\theta_2 = 1.3$ and vary μ_2.

(b) Our method; Fix $\mu_2 = 0.7$ and vary θ_2.

(c) Joint method from Budhathoki et al. (2021); Fix $\theta_2 = 1.3$ and vary μ_2.

(d) Joint method from Budhathoki et al. (2021); Fix $\mu_2 = 0.7$ and vary θ_2.

Figure C.1: Mean squared error differences attributed by our model and Budhathoki et al. (2021) in the synthetic setting described in Appendix C



(a) Our method; Fix $\mu_2 = 1$ and vary θ_2, with $\mathbb{C}_\mathcal{D} = \{\mathcal{D}_X, \mathcal{D}_{Y|X}\}$, the actual causal graph

(b) Our method; Fix $\mu_2 = 1$ and vary θ_2, with $\mathbb{C}_\mathcal{D} = \{\mathcal{D}_Y, \mathcal{D}_{X|Y}\}$, a mis-specified causal graph

Figure C.2: Mean squared error differences attributed by our model when there is only concept drift, for the actual causal graph (a), and a mis-specified causal graph (b).

D. Additional Experimental Results

D.1. Synthetic Data

Setup. We generate a synthetic binary classification dataset with five variables according to the following data generating process, corresponding to the causal graph shown in Figure D.1. Here, $\xi_p : \{0,1\} \to \{0,1\}$ is a function that randomly flips the input with probability p.

$$\begin{aligned} G &\sim Ber(0.5), \qquad X_2 = \mathcal{N}(\xi_{0.25}(Y) + G, 1) \\ Y &= \xi_q(G), \qquad X_1 = \mathcal{N}(\omega\xi_{0.25}(Y), 1) \\ X_3 &= \mathcal{N}(\xi_{0.25}(Y) + \mu G, 1) \end{aligned}$$

Where q, ω and μ are parameters of the data generating process. Here, G represents a spurious correlation (Aubin et al., 2021; Arjovsky et al., 2019) that is highly correlated with Y, and is easily inferred from (X_2, X_3). By selecting a large value for q (the spurious correlation strength) on the source environment, we can create a dataset where models rely more heavily on using X_2 and X_3 to infer G and then Y, instead of inferring $\xi_{0.25}(Y)$ across the three features to estimate Y directly.

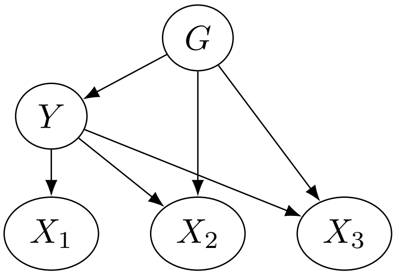


Figure D.1: Causal Graph for Synthetic data

In the source environment, we set $q = 0.9, \omega = 1$ and $\mu = 3$. We generate 20,000 samples using these parameters, and train logistic regression (LR) models on (X_1, X_2, X_3) to predict Y, using 3-fold cross-validation to select the best model. We attribute performance changes for this model using the proposed method. We explore four data settings for the target environment:

(a) Label Shift: Vary $q \in [0, 1]$. Keep ω and μ at their source values. Only $P(Y|G)$ changes. This represents a label shift for the model across domains (which does not have access to G).

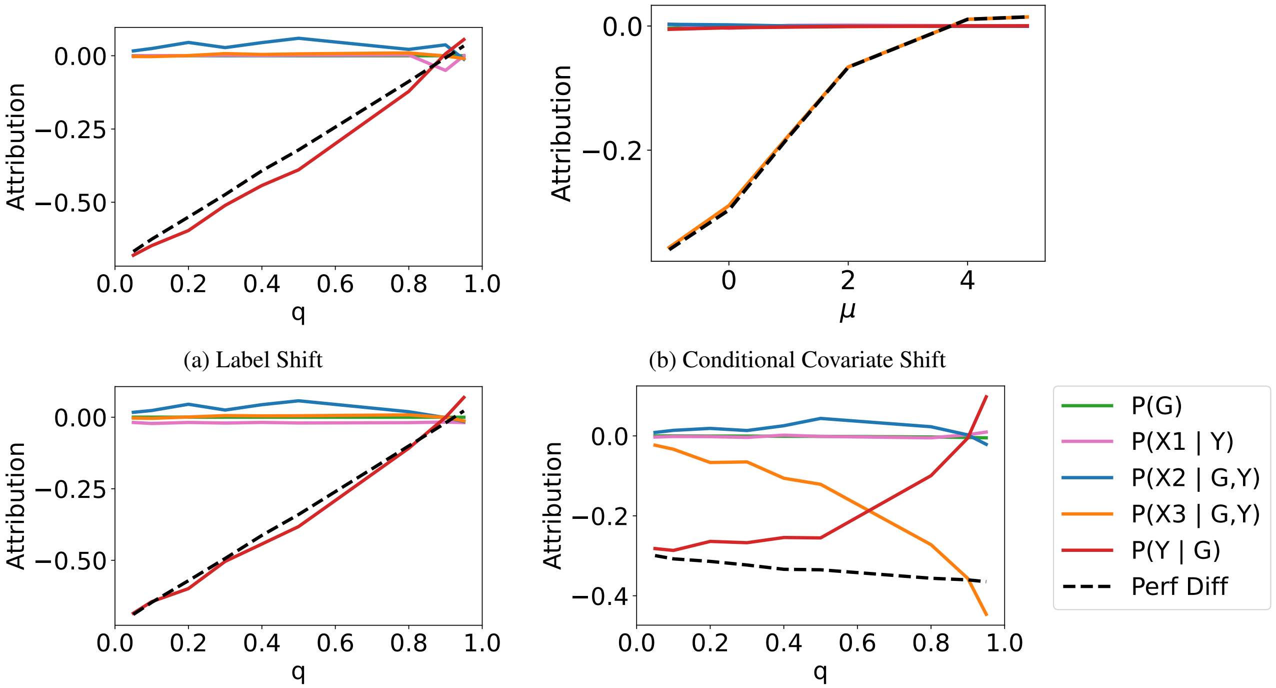
(b) Covariate Shift: Vary $\mu \in [0, 5]$. Keep q and ω at their source values. Only $P(X_3|G, Y)$ changes across domains.

(c) Combined Shift 1: Set $\omega = 0$ in the target environment and vary $q \in [0, 1]$. Keep μ at its source value. Both $P(X_1|Y)$ and $P(Y|G)$ change across domains, but the shift should be largely attributed to $P(Y|G)$ as the model relies on this correlation much more than X_1.

(d) Combined Shift 2: Set $\mu = -1$ in the target environment. Further, vary $q \in [0, 1]$. Keep ω at its source value. Both $P(X_3|Y)$ and $P(Y|G)$ change across domains, but their specific contribution to model performance degradation is not known exactly.

We use our method to explain performance changes in accuracy and Brier score for each model on target environments generated within each setting (with $n = 20,000$), computing density ratios using XGB (Chen & Guestrin, 2016) models. Note that the causal graph shown in Figure D.1 implies five potential distribution in the candidate set: $\mathtt{C}_{\mathcal{D}} = \{\mathcal{D}_G, \mathcal{D}_{Y|G}, \mathcal{D}_{X_1|Y}, \mathcal{D}_{X_2|G,Y}, \mathcal{D}_{X_3|G,Y}\}$.

Our method correctly identifies distribution shifts. First, we focus on the output of our method with LR as the model of interest and accuracy as the metric, shown in Figure D.2. We find that our method attributes all of the performance changes to the correct ground truth shifts, both when there is a single shift (Settings (a) and (b)) and when there are multiple shifts (Settings (c) and (d)). In the case of Setting (c), we find that our method attributes all of the performance drop to a shift in $P(Y|G)$. This is because the model relies largely on the spurious information (G inferred from X_2 and X_3) in the source environment. We verify this by examining the overall feature importance for both models (see Table D.2). Further, in the



(a) Label Shift

(b) Conditional Covariate Shift

(c) Combined Shift 1

(d) Combined Shift 2

Figure D.2: Attributions by our method using the correct causal graph for the change in accuracy to five potential distributional shifts on the synthetic dataset for the LR model. Further from 0 implies higher (signed) attribution. We observe that the overall change (Perf Diff) is attributed to the true shift(s) in all cases.

presence of multiple shifts which simultaneously impact model performance (Setting (d)), we find that our method is able to attribute a meaningful fraction of the performance shift to each distribution.

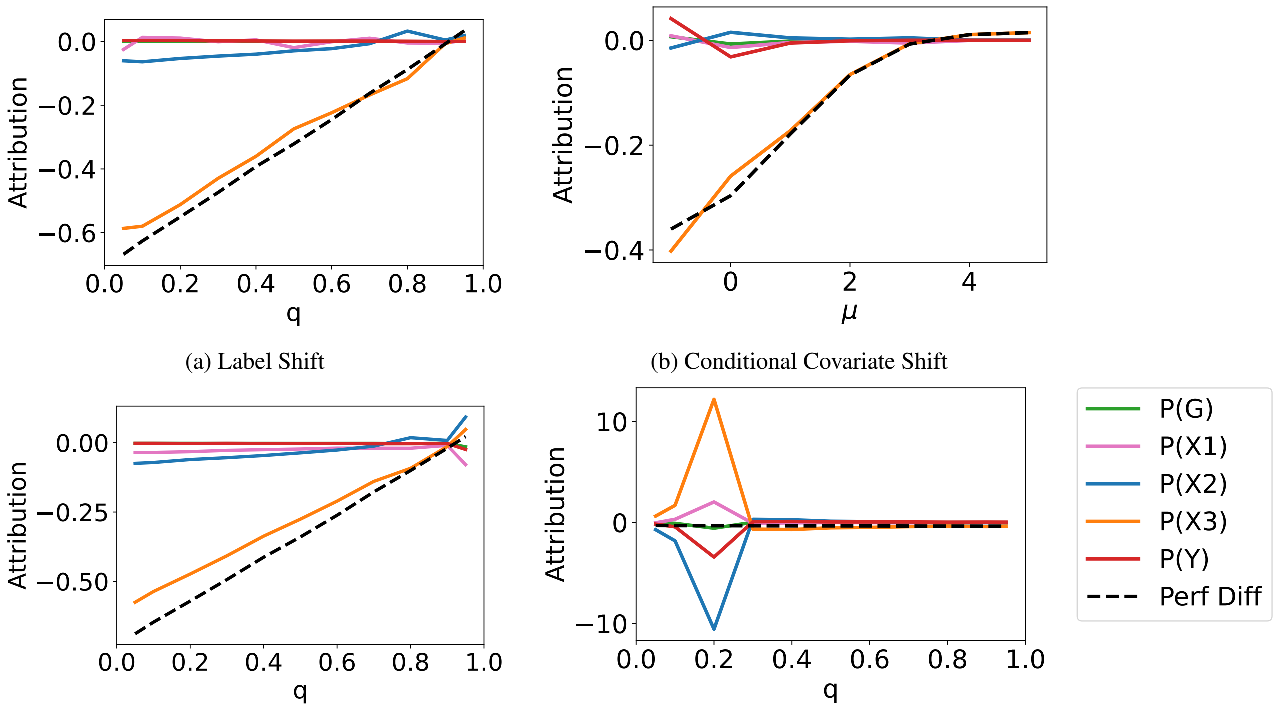
Baseline methods have multiple flaws. We find that the baselines methods all have several flaws which result in inadequate attributions in this setting. For example, the marginal candidate set (Figure D.3) does not provide meaningful attributions as it does not examine conditional relationships, especially as it attributes large shifts to $P(X_3)$. Similarly, the fully connected graph (Figure D.4) demonstrates a large degree of noise, particularly in the combined shift, though the dominant distribution appears to largely be correct. Next, the SHAP baseline (Figure D.5) completely fails in this setting, as it is not able to attribute any shift to the mechanism for Y. Finally, we find that the attributions provided by the joint method in (Budhathoki et al., 2021) (Figure D.6) are not meaningful, as the magnitude of the KL divergence varies wildly between distributions when multiple shifts are present.

Table D.1: Performance of each model on the source environment for the synthetic dataset.

	Accuracy	Brier Score
LR	0.871	0.102
XGB	0.870	0.099

Table D.2: Feature importances of each model on the synthetic dataset. For LR, the model coefficient is shown, and for XGB, the total information gain from each feature.

	LR (Coefficient)	XGB (Gain)
X_1	0.400	31.1
X_2	0.381	29.2
X_3	1.994	358.2



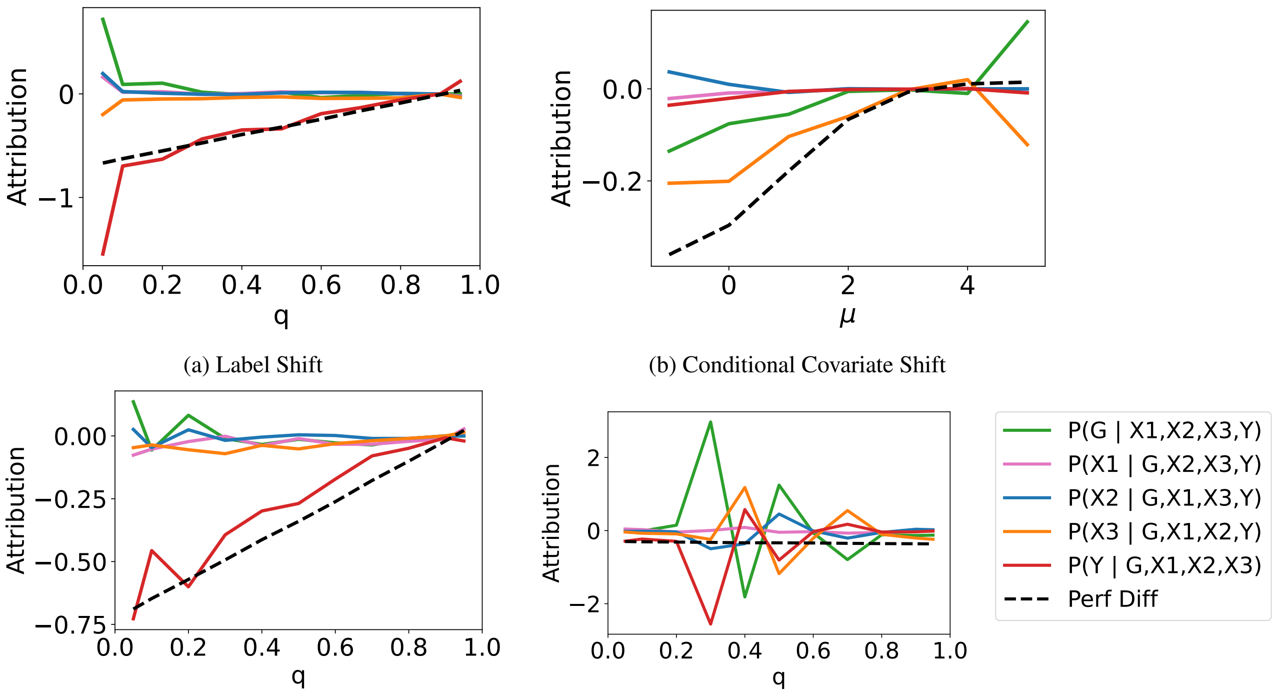
(a) Label Shift

(b) Conditional Covariate Shift

(c) Combined Shift 1

(d) Combined Shift 2

Figure D.3: Accuracy differences attributed by our method using the marginal graph to five potential distributional shifts on the synthetic dataset for the LR model.



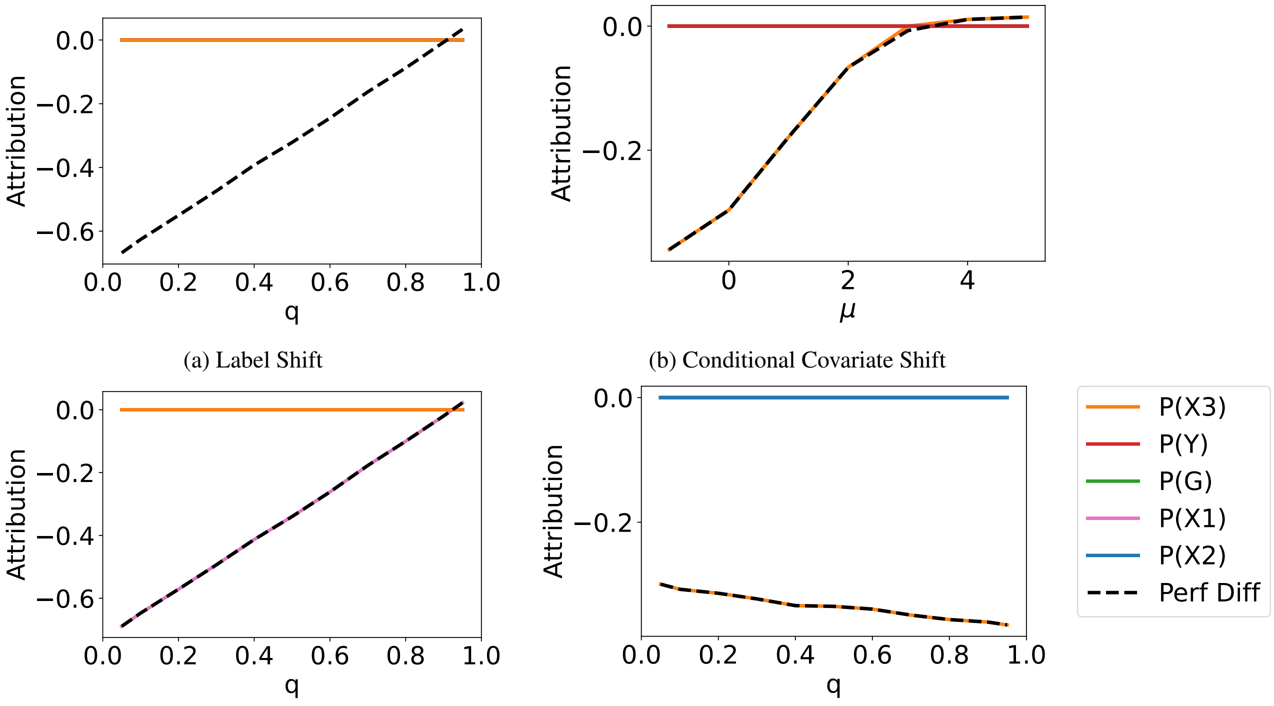
(a) Label Shift

(b) Conditional Covariate Shift

(c) Combined Shift 1

(d) Combined Shift 2

Figure D.4: Accuracy differences attributed by our method using the fully connected graph to five potential distributional shifts on the synthetic dataset for the `LR` model.



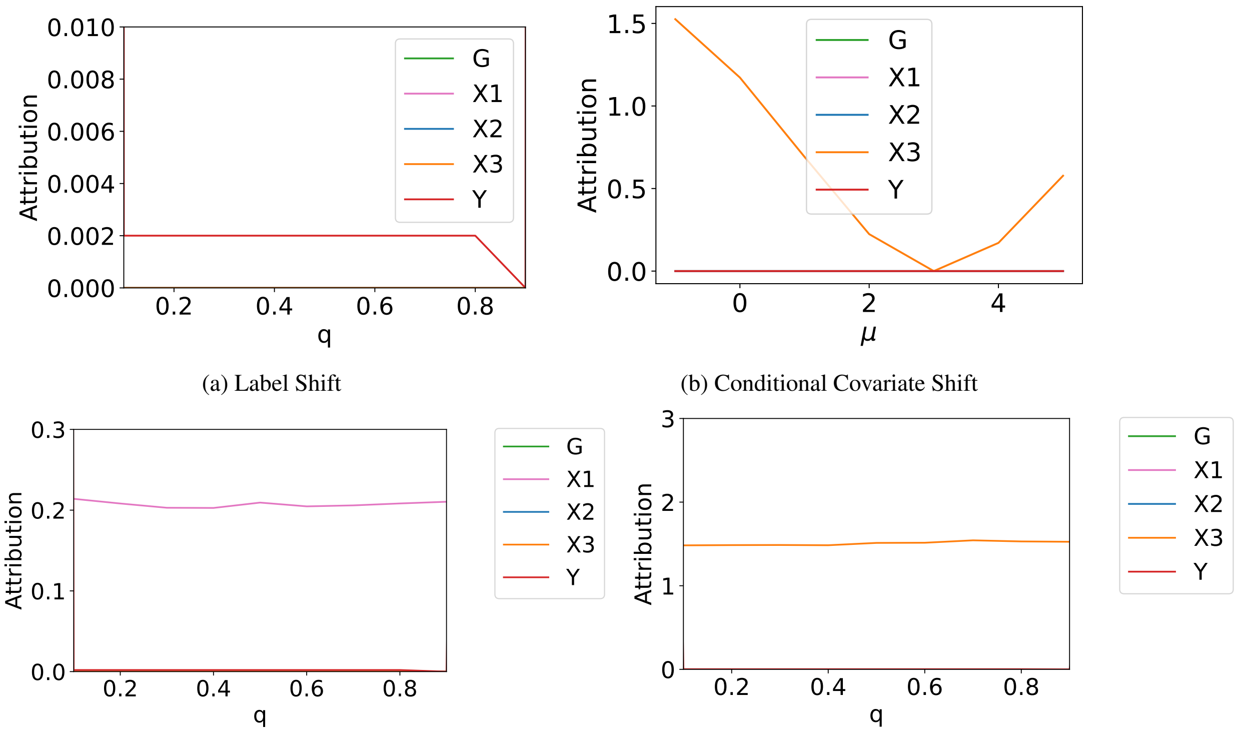
(a) Label Shift

(b) Conditional Covariate Shift

(c) Combined Shift 1

(d) Combined Shift 2

Figure D.5: Accuracy differences attributed by the SHAP baseline to five potential distributional shifts on the synthetic dataset for the `LR` model.



(a) Label Shift

(b) Conditional Covariate Shift

(c) Combined Shift 1

(d) Combined Shift 2

Figure D.6: Attributions by the joint method in Budhathoki et al. (2021) to five potential distributional shifts on the synthetic dataset. We note that the magnitude of the attribution is not informative in interpreting model performance changes, particularly when multiple shifts are present.

D.2. ColoredMNIST

Setup. We evaluate our method on the popular ColoredMNIST dataset (Arjovsky et al., 2019). We generalize the data generating process for this dataset to include several tunable dataset parameters, using the following procedure:

1. Generate a binary label y_{obs} from the MNIST label y_{num} by assigning $y_{obs} = 0$ if $y_{num} \in \{0, 1, 2, 3, 4\}$, and $y_{obs} = 1$ otherwise.
2. Flip y_{obs} with probability η to obtain y.
3. Generate the color a by flipping y with probability ρ.
4. Construct X as $[X_{fig} \cdot (1 - a), X_{fig} \cdot a]$.
5. Subsample the majority class so that the fraction of samples with $y = 1$ in the dataset is equal to β.

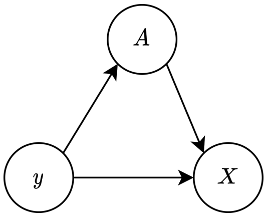
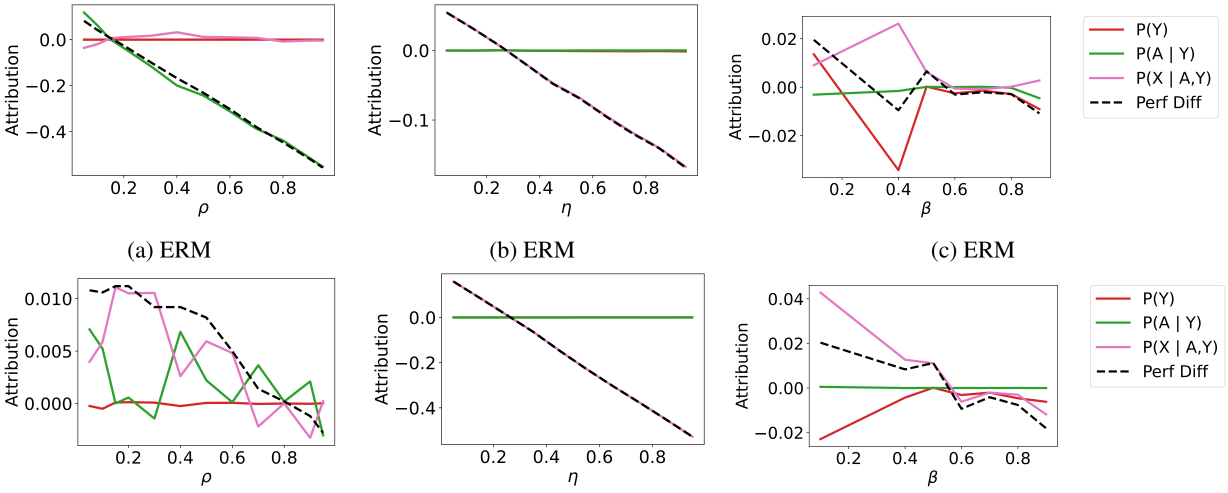


Figure D.7: Causal graph for the ColoredMNIST dataset.

This corresponds to the causal graph in Figure D.7. Note that X is a vector-valued node representing the image. We split the MNIST data equally into source and target environments. On the source environment, we set β = 0.5, $\eta = 0.25$, and $\rho = 0.15$. We vary these two parameters independently on the target environment, keeping the other fixed at their source value. Note that shifting β, η, and ρ correspond to shifting $P(Y)$, $P(X|A, Y)$ and $P(A|Y)$ respectively.

Following (Arjovsky et al., 2019), we use 3-layer MLPs to predict y from X on the source dataset. We train this MLP with standard ERM, as well as with GroupDRO (Sagawa* et al., 2020). The network trained with ERM should rely on the spurious correlation (i.e. the color), while the GroupDRO network should be invariant to the color. We experiment with using both the correct causal graph, and the all marginal causal graph.

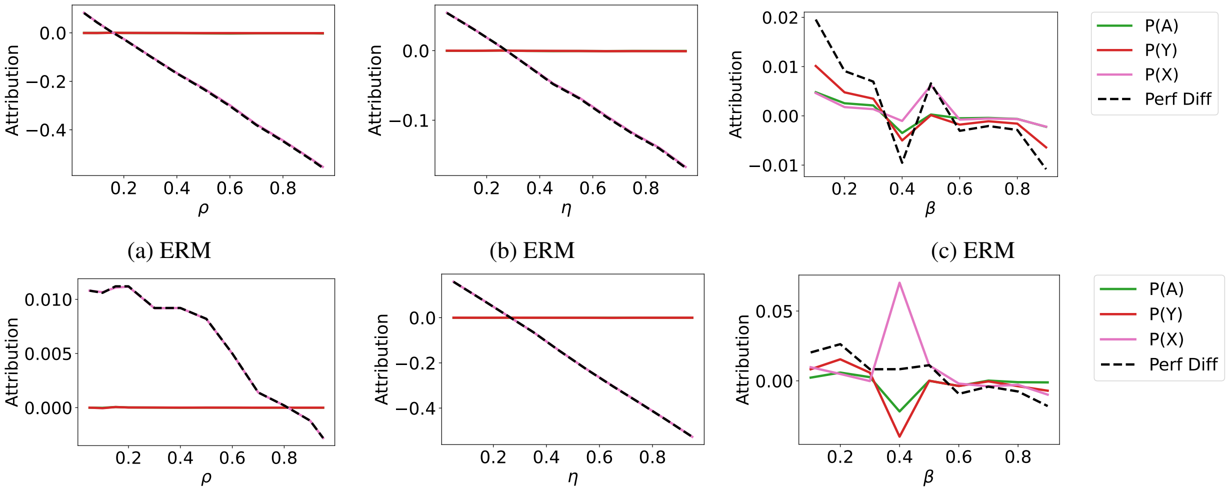
Results. In Figure D.8, we show the attributions provided by our method for the correct causal graph. We observe that the ERM model is highly susceptible to shifts in ρ and correctly attributes all of the shift to $P(A|Y)$ in that setting. In contrast, the GroupDRO model receives almost no attribution to $P(A|Y)$ as it does not use the attribute spuriously. However, since it makes use of the invariant signal, shifting η results in large performance drops. Both models do not receive a significant attribution for $P(Y)$, as the error rate tends to be similar between the two classes. Looking at the results for the marginal causal graph in Figure D.9, we find that attributions created using this candidate set are not meaningful as it deviates too far from the actual causal mechanism. For example, no shifts are ever attributed to $P(A)$, as this distribution is not changed by any parameters.



(a) ERM (b) ERM (c) ERM

(d) GroupDRO (e) GroupDRO (f) GroupDRO

Figure D.8: Accuracy differences attributed by our model with the correct causal graph to three potential distribution shifts in ColoredMNIST.



(a) ERM (b) ERM (c) ERM

(d) GroupDRO (e) GroupDRO (f) GroupDRO

Figure D.9: Accuracy differences attributed by our model with the all marginals causal graph to three potential distribution shifts in ColoredMNIST. We observe that using a causal graph that does not match the underlying shifting mechanisms may lead to attributions that are not meaningful.

D.3. Gender Classification in CelebA

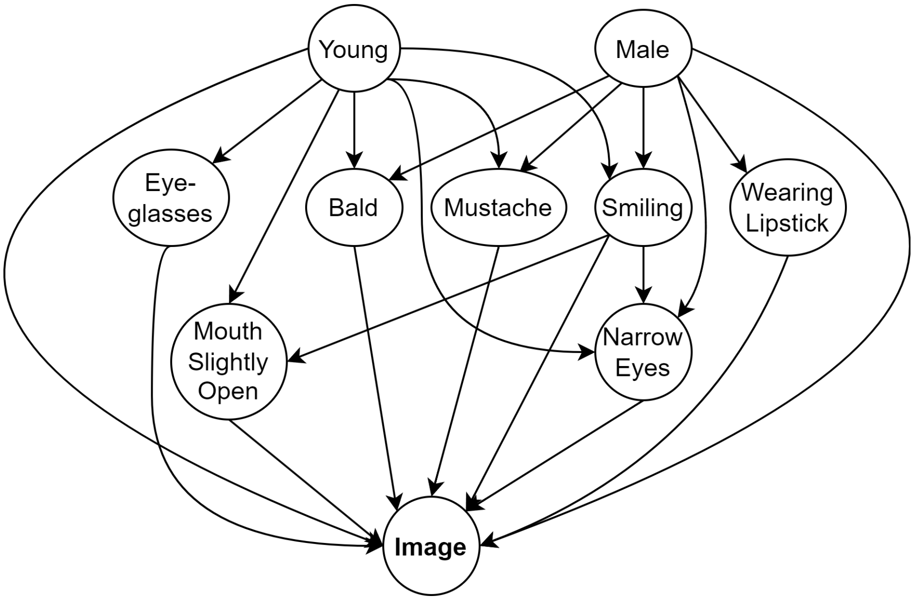


Figure D.10: Causal graph for the celebA dataset.

We use the CelebA dataset (Liu et al., 2015), where the goal is to predict gender from facial images. We adopt a setup similar to the one presented in Thams et al. (2022). We assume this data is generated from the causal graph shown in Figure D.10. We train a CausalGAN (Kocaoglu et al., 2018), a generative model that allows us to synthesize images faithful to the graph. CausalGAN allows to train attribute nodes (young, bald, etc) which are binary-valued, and then synthesize images conditioned on specific attributes. This allows us to simulate known distribution shifts (in attributes and hence images) across environments. We assume that the causal mechanisms in the source environment have log-odds equal to the ones shown in Table D.3. We omit $\mathcal{D}_{\text{Image}|\text{Pa(Image)}}$ from $\mathbb{C}_{\mathcal{D}}$, as 1) this distribution is parameterized by the CausalGAN and does not change, and 2) it is high-dimensional and difficult to work with. We investigate attribution to distribution shift of an ImageNet-pretrained ResNet-18 (He et al., 2016) finetuned to predict gender from the image using frozen representations. Note that the model is only given access to the image itself, but not any of the binary attributes in the causal graph. We conduct the following two experiments for evaluation.

Experiment 1. The purpose of this experiment is to demonstrate that our method provides the correct attributions for a wide range of random shifts. To create the target environment, we first select the number of mechanisms to perturb, $n_p \in \{1, 2, ..., 6\}$. We select n_p mechanisms from the causal graph, which we define as the ground truth shift. For each mechanism, we perturb one of the log odds by a quantity uniformly selected from $[-2.0, -1.0] \cup [1.0, 2.0]$. We then use the CausalGAN to simulate a dataset of $10,000$ images based on the modified mechanisms, and use our method to attribute the accuracy change between source and target. We select the n_p distributions from our method with the largest attribution magnitude, and compare this set with the set of ground truth shifts to calculate an accuracy score. We repeat this experiment 20 times for each value of $n_p \in \{1, 2, ..., 6\}$, and only select experiments with a non-trivial change in model performance (change in accuracy $\geq 1\%$).

Experiment 2. The purpose of this experiment is to investigate the magnitude of our model attributions in the presence of multiple shifts. We perturb the log odds for $P(\text{Wearing Lipstick}|\text{Male})$ and $P(\text{Mouth Slightly Open}|\text{Smiling})$ jointly by $[-3.0, 3.0]$. We compare the magnitude of the attributions for the two associated mechanisms, relative to the total shift in accuracy.

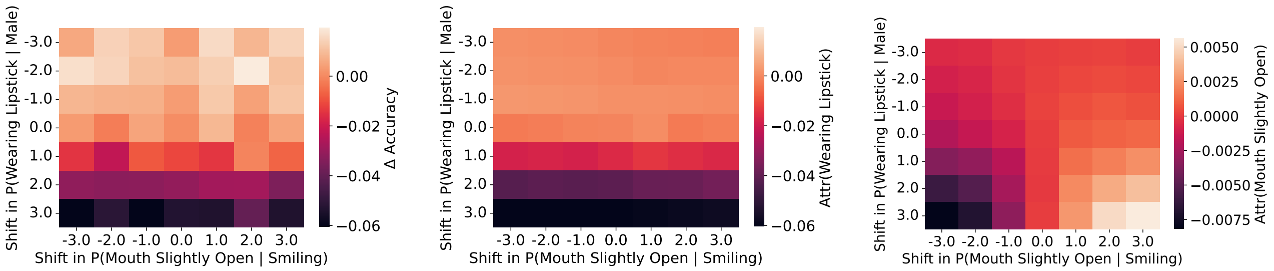
Results. In Table D.4, we show the average accuracy of our method for each value of n_p. We find that our method achieves roughly 90% accuracy at this task. However, we note that this is not the ideal scenario to validate our method, as not all shifts in the ground truth set will result in a decrease in the model performance. As our method will not attribute a significant value to shifts which do not impact model performance, this explains the accuracy discrepancy observed.

Table D.3: Data generating process for the causal graph shown in Figure D.10

Variable	**Log Odds**
Young	Base: 0.0
Male	Base: 0.0
Eyeglasses	Base: 0.0, Young: -0.4
Bald	Base: -3.0, Male: 3.5, Young: -1.0
Mustache	Base: -2.5, Male: 2.5, Young: 0.5
Smiling	Base: 0.25, Male: -0.5, Young: 0.5
Wearing Lipstick	Base: 3.0, Male: -5.0
Mouth Slightly Open	Base: -1.0, Young: 0.5, Smiling: 1.0
Narrow Eyes	Base: -0.5, Male: 0.3, Young: 0.2, Smiling: 1.0

Table D.4: Average accuracy of our method in attributing shifts to the ground truth shift in CelebA for each number of perturbed mechanisms (n_p).

n_p	Avg Accuracy
1	1.00 ± 0.00
2	0.72 ± 0.36
3	0.90 ± 0.16
4	0.85 ± 0.13
5	0.93 ± 0.10
6	0.91 ± 0.09



(a)

(b)

(c)

Figure D.11: We vary the perturbation in log odds in the target environment for the "wearing lipstick" and "mouth slightly open" attributes. We show (a) the total shift in accuracy, (b) our attribution to $P(\text{Wearing Lipstick}|\text{Male})$, (c) our attribution to $P(\text{Mouth Slightly Open}|\text{Young, Smiling})$.

Table D.5: Predictive performance of XGB models trained to predict attributes from the source environment in CelebA, and the correlation of each attribute the gender label, as measured by the Matthews Correlation Coefficient (MCC).

	Predictive Performance		Correlation
	AUROC	AUPRC	MCC
Wearing Lipstick	0.968	0.976	-0.837
Mouth Slightly Open	0.927	0.924	-0.036

In Figure D.11, we show the output of our method in Experiment 2. First, we find that shifting these two attributes causes a large decrease in the accuracy (up to 6%), and that $P(\text{Wearing Lipstick}|\text{Male})$ seem to be the stronger factor responsible for the decrease. Looking at our attributions, we find that we indeed attribute the large majority of the shift to $P(\text{Wearing Lipstick}|\text{Male})$. Here, the relative attribution to $P(\text{Wearing Lipstick}|\text{Male})$ is relatively unaffected by the shift in the other variable, as its effect on the total shift is so minuscule. However, looking at the attribution to $P(\text{Mouth Slightly Open}|\text{Young, Smiling})$, in addition to the small magnitude, we do observe an interesting effect, where the attributed accuracy drop is greater when the two shifts are combined.

To justify the magnitude of our attributions, we use an ad-hoc heuristic that attempts to approximate the model reliance on each attribute in making its prediction. First, we train XGBoost models on the ResNet-18 embeddings from the source environment to predict the two attributes. From Table D.5, we find that “Wearing Lipstick” is easier to infer from the representations than “Mouth Slightly Open”. Next, we measure the correlation of each attribute to the label (gender), finding that the magnitude of the correlation is also much higher for “Wearing Lipstick”. As “Wearing Lipstick” is both easier to detect from the image, and is also a stronger predictor of gender, it seems reasonable to conclude that the model trained on the source would utilize it more in its predictions, and thus our method should attribute more of the performance drop to the “Wearing Lipstick” distribution when it shifts.

D.4. eICU Data

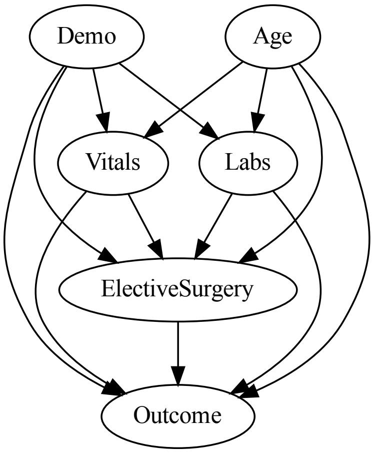


Figure D.12: Causal Graph for eICU data

Table D.6 lists the features that comprise the nodes in the causal graph. Please refer to (Singh et al., 2022, Supporting Information Table C) for descriptions. Code for preprocessing the eICU database for the mortality prediction task is made available at the Github repository by Johnson et al. (2018).

Table D.6: Features comprising the nodes of the causal graph in Figure D.12.

Variable	Features
Demo	is_female, race_black, race_hispanic, race_asian, race_other
Vitals	heartrate, sysbp, temp, bg_pao2fio2ratio, urineoutput
Labs	bun, sodium, potassium, bicarbonate, bilirubin, wbc, gcs
Age	age
ElectiveSurgery	electivesurgery
Outcome	death

The Midwest domain has 10,056 samples, and the South domain has 7,836. Both domains have 20 features and a binary outcome. We randomly split each into 50% for training the XGBoost model and 50% for evaluation (and estimation of Shapley values). To create the resampled Midwest dataset, we subsample 67% of the training set but selectively sample records with age less than 63 (which is the median age in Midwest) with probability 5 times that of the probability of sampling the rest of the records.

E. Convergence Analysis

We present a preliminary analysis to study the impact of errors resulting from estimating importance weights on the properties of the Shapley values. The theoretical analysis is informal and presented here with the goal of motivating further study. Importantly, we experimentally evaluate the error in a synthetic setup.

E.1. Sketch for a Theoretical Analysis

Remark E.1. Under bounded estimation error and for a bounded loss, Property 2.1 holds asymptotically.

Proof. Suppose $1-\epsilon_d^n \leq w_d \leq 1+\epsilon_d^n$, that is we get approximate importance weights from finite samples. Then for a bounded loss,

$$-\frac{\epsilon_d^n l}{|\mathtt{C}_{\mathcal{D}}|}\sum_{\widetilde{\mathtt{C}}\subseteq\mathtt{C}_{\mathcal{D}}\setminus\{d\}}\binom{|\mathtt{C}_{\mathcal{D}}|-1}{|\widetilde{\mathtt{C}}|}^{-1}\mathbb{E}_{\mathcal{D}^{\text{source}}}[\prod_{\tilde{d}\in\widetilde{\mathtt{C}}\cup d} w_{\tilde{d}}] \leq \text{Attr}^n(d) \leq \frac{\epsilon_d^n l}{|\mathtt{C}_{\mathcal{D}}|}\sum_{\widetilde{\mathtt{C}}\subseteq\mathtt{C}_{\mathcal{D}}\setminus\{d\}}\binom{|\mathtt{C}_{\mathcal{D}}|-1}{|\widetilde{\mathtt{C}}|}^{-1}\mathbb{E}_{\mathcal{D}^{\text{source}}}[\prod_{\tilde{d}\in\widetilde{\mathtt{C}}\cup d} w_{\tilde{d}}] \tag{6}$$

Suppose that in the worst case all other distributions shift except d, but the shifts are bounded, i.e. $\frac{1}{\eta} < w_{\tilde{d}} < \eta$ where $\eta > 1$ for all $\tilde{d} \in \mathtt{C}_{\mathcal{D}} \setminus \{d\}$. Then,

$$\begin{aligned}
&-\frac{\epsilon_d^n l}{|\mathtt{C}_{\mathcal{D}}|}\sum_{\widetilde{\mathtt{C}}\subseteq\mathtt{C}_{\mathcal{D}}\setminus\{d\}}\binom{|\mathtt{C}_{\mathcal{D}}|-1}{|\widetilde{\mathtt{C}}|}^{-1}\left(\frac{1}{\eta}\right)^{|\widetilde{\mathtt{C}}|-1} \leq \text{Attr}^n(d) \leq \frac{\epsilon_d^n l}{|\mathtt{C}_{\mathcal{D}}|}\sum_{\widetilde{\mathtt{C}}\subseteq\mathtt{C}_{\mathcal{D}}\setminus\{d\}}\binom{|\mathtt{C}_{\mathcal{D}}|-1}{|\widetilde{\mathtt{C}}|}^{-1}\eta^{|\widetilde{\mathtt{C}}|-1} \\
&-\frac{\epsilon_d^n l}{|\mathtt{C}_{\mathcal{D}}|}\left(1+\frac{1}{\eta}\right)^{|\widetilde{\mathtt{C}}|-1} \leq \text{Attr}^n(d) \leq \frac{\epsilon_d^n l}{|\mathtt{C}_{\mathcal{D}}|}(1+\eta)^{|\widetilde{\mathtt{C}}|-1}
\end{aligned} \tag{7}$$

The error in attribution is dominated by the shifts in other distribution and the error in estimating the weight for distribution d. Thus as $n \to \infty$, so long as $\epsilon_d^n \to 0$, the attribution $\text{Attr}^n(d) \to 0$. □

The above suggests that our Property 2.1 may not hold exactly in finite samples due to estimation error.

E.2. Empirical Analysis

To empirically examine the estimation error as a function of the number of samples, we adopt the synthetic setup described in Appendix C with $\theta_2 = 0.5$ and $\mu_2 = 0.5$. We choose this setup because the exact importance weights can be computed analytically, and thus allows us to quantify the error of an importance weight estimator. We experiment with the KLIEP method (Sugiyama et al., 2008), as well as using logistic regression (LR) and XGBoost (XGB) as probabilistic estimators, the last of which we use in the paper. Given n samples, we randomly choose $n/2$ samples to train the importance weight estimator, and the remaining $n/2$ samples to evaluate the attribution. In each run, we compute the mean squared error between empirical and exact importance weights for $\mathcal{D}_X$ and $\mathcal{D}_{Y|X}$, as well as the mean squared error between the empirical and exact attributions $Attr(\mathcal{D}_X)$ and $Attr(\mathcal{D}_{Y|X})$. Note that the analytical attributions are $Attr(\mathcal{D}_X) = -0.06375$ and $Attr(\mathcal{D}_{Y|X}) = 0.16875$. We display the result in Table E.7.

We first find that KLIEP did not converge for smaller values of n, and takes prohibitively long to run for larger values of n. It was for these reasons that we did not select it for use in the main paper. Next, we note that a linear model is not complex enough to differentiate between the source and target, and thus results in a biased estimator (non-zero MSE for large n), though this still results in relatively small MSE for the attributions. Finally, we observe that the errors of XGB converge to zero both in the estimated importance weights as well as in the resulting attributions.

Table E.7: Estimation error in Shapley attributions for finite samples, using the synthetic setup described in Appendix C.

Model	n	$MSE(w_X)$	$MSE(w_{X,Y})$	$MSE(Attr(\mathcal{D}_X))$	$MSE(Attr(\mathcal{D}_Y))$
KLIEP	100	0.557 ± 0.339	1.093 ± 0.422	0.016 ± 0.003	0.029 ± 0.015
	200	0.344 ± 0.194	1.612 ± 0.809	0.004 ± 0.000	0.037 ± 0.002
	1000	0.212 ± 0.145	0.588 ± 0.203	0.003 ± 0.001	0.041 ± 0.012
LR	20	0.237 ± 0.133	0.428 ± 0.221	0.004 ± 0.000	0.028 ± 0.000
	50	0.748 ± 0.707	14.175 ± 22.402	0.004 ± 0.000	0.023 ± 0.005
	100	0.555 ± 0.064	0.549 ± 0.342	0.004 ± 0.000	0.033 ± 0.009
	200	0.543 ± 0.083	1.861 ± 1.864	0.004 ± 0.000	0.036 ± 0.012
	1000	0.600 ± 0.042	4.549 ± 3.196	0.004 ± 0.000	0.010 ± 0.007
	5000	0.583 ± 0.061	3.208 ± 0.269	0.004 ± 0.000	0.008 ± 0.003
	10000	0.566 ± 0.038	5.887 ± 3.753	0.004 ± 0.000	0.007 ± 0.004
	50000	0.325 ± 0.008	3.631 ± 0.430	0.004 ± 0.000	0.009 ± 0.002
XGB	20	1.191 ± 1.034	0.974 ± 0.740	0.003 ± 0.003	0.068 ± 0.038
	50	1.003 ± 0.848	40.087 ± 66.979	0.010 ± 0.008	0.022 ± 0.004
	100	13.151 ± 19.926	30.083 ± 25.079	0.017 ± 0.016	0.013 ± 0.020
	200	10.053 ± 8.450	7.491 ± 9.321	0.001 ± 0.001	0.031 ± 0.015
	1000	0.418 ± 0.378	39.845 ± 66.373	0.005 ± 0.003	0.031 ± 0.023
	5000	0.419 ± 0.502	3.485 ± 2.964	0.003 ± 0.003	0.030 ± 0.025
	10000	0.065 ± 0.062	0.903 ± 0.429	0.001 ± 0.001	0.014 ± 0.018
	50000	0.083 ± 0.050	0.396 ± 0.095	0.000 ± 0.000	0.002 ± 0.001