Robust and Stable Black Box Explanations

Himabindu Lakkaraju [1] **Nino Arsov** [2] **Osbert Bastani** [3]

Abstract

As machine learning black boxes are increasingly being deployed in real-world applications, there has been a growing interest in developing post hoc explanations that summarize the behaviors of these black boxes. However, existing algorithms for generating such explanations have been shown to lack stability and robustness to distribution shifts. We propose a novel framework for generating robust and stable explanations of black box models based on adversarial training. Our framework optimizes a minimax objective that aims to construct the highest fidelity explanation with respect to the worst-case over a set of adversarial perturbations. We instantiate this algorithm for explanations in the form of linear models and decision sets by devising the required optimization procedures. To the best of our knowledge, this work makes the first attempt at generating post hoc explanations that are robust to a general class of adversarial perturbations that are of practical interest. Experimental evaluation with real-world and synthetic datasets demonstrates that our approach substantially improves robustness of explanations without sacrificing their fidelity on the original data distribution.

1. Introduction

Over the past decade, there has been an increasing interest in leveraging machine learning (ML) models to aid decision making in critical domains such as healthcare and criminal justice. However, the successful adoption of these models in the real world relies heavily on how well decision makers are able to understand and trust their functionality (Doshi-Velez & Kim, 2017; Lipton, 2016). Decision makers must have a clear understanding of the model behavior so they can diagnose errors and potential biases in these models,

and decide when and how to employ them. However, the proprietary nature and increasing complexity of machine learning models poses a severe challenge to understanding these complex *black boxes*, motivating the need for tools that can explain them in a faithful and interpretable manner.

Several different kinds of approaches have been proposed to produce interpretable *post hoc explanations* of black box models. For instance, LIME and SHAP (Ribeiro et al., 2016; Lundberg & Lee, 2017b) explain individual predictions of any given black box classifier via *local* approximations. On the other hand, approaches such as MUSE (Lakkaraju et al., 2019b) focus on explaining the high-level *global* behavior of any given black box.

However, recent work has shown that post hoc explanation methods are *unstable* (i.e., small perturbations to the input can substantially change the constructed explanations), as well as *not robust to distribution shifts* (i.e., explanations constructed using a given data distribution may not be valid on others) (Ghorbani et al., 2019; Lakkaraju & Bastani, 2020). A key reason why many post hoc explanation methods are not robust is that they construct explanations by optimizing fidelity on a given covariate distribution $p(x)$ (Ribeiro et al., 2018; 2016; Lakkaraju et al., 2019b)—i.e., choose the explanation that makes the same predictions as the black box on $p(x)$. To see why these approaches may fail to be robust, consider a covariate distribution $p(x_1, x_2)$ where x_1 and x_2 are perfectly correlated, and an outcome $y = \mathbb{I}[x_1 \geq 0]$. Suppose we have a black box $B^*(x_1, x_2) = \mathbb{I}[x_2 \geq 0]$, and an explanation $\hat{E}(x_1, x_2) = \mathbb{I}[x_1 \geq 0]$. Since x_1 and x_2 are perfectly correlated, the explanation has perfect fidelity—i.e.,

$$\mathbb{P}_{p(x_1, x_2)}[\hat{E}(x_1, x_2) = B^*(x_1, x_2)] = 1. \qquad (1)$$

Thus, $\hat{E}$ appears to be a good explanation of B^*. However, if the underlying covariate distribution changes—e.g., to $p'(x_1, x_2)$ where x_1 and x_2 are independent—then $\hat{E}$ no longer has high fidelity.

The lack of robustness is problematic because many of the undesirable behaviors of black box models that can be diagnosed using interpretability relate to distribution shifts. For instance, it has been shown that interpretability can help users in assessing whether a model would transfer well to a new domain (Ribeiro et al., 2016)—e.g., from one

---

[1]Harvard University [2]Macedonian Academy of Arts & Sciences [3]University of Pennsylvania. Correspondence to: Himabindu Lakkaraju <hlakkaraju@hbs.edu>.

hospital to another (Bastani, 2018); Caruana et al. (2015) show that experts use interpretable models to identify spurious relationships which do not hold if the underlying data changes—e.g., if a patient has asthma, he is not likely to die from pneumonia; these are intrinsically distribution shift issues. Thus, for the explanations to shed light on these kinds of issues in the black box, high fidelity on the original distribution alone may be insufficient; instead, it also needs to achieve high fidelity on the relevant shifted distributions. To further complicate the problem, we often do not know in advance what are the relevant distribution shifts. Therefore, constructing explanations that are robust to a general class of possible shifts is of great importance.

We propose a novel algorithmic framework, RObust Post hoc Explanations (ROPE) for constructing black box explanations that are not only stable but also robust to shifts in the underlying data distribution. To the best of our knowledge, our work is the first attempt at generating robust post hoc explanations for black boxes. ROPE focuses on two notions of robustness. The first is *adversarial robustness* (Ghorbani et al., 2019), which intuitively says that if the inputs are adversarially perturbed (by small amounts), then the explanation should not change significantly. The second is *distributional robustness* (Namkoong & Duchi, 2016), which is similar to adversarial robustness but considers perturbations to the input distribution rather than individual inputs. While ROPE considers distributional and adversarial robustness, these properties also improve stability. This is due to the fact that explanations designed to be robust to input perturbations are not likely to vary drastically with small changes in inputs.

First, we propose a novel minimax objective that can be used to construct robust black box explanations for a given family of interpretable models. This objective encodes the goal of returning the highest fidelity explanation with respect to the worst-case over a set of distribution shifts.

Second, we propose a set of distribution shifts that captures our intuition about the kinds of shifts to which interpretations should be robust. In particular, this set includes shifts that contain perturbations to a small number of covariates. For instance, robustness to these shifts ensure that the marginal dependence of the black box on a single covariate is preserved in the explanation, since the explanation must be robust to changes in that covariate alone.

Third, we propose algorithms for optimizing this objective in two settings: (i) explanations such as linear models with continuous parameters that can be optimized using gradient descent, in which case we use adversarial training (Goodfellow et al., 2015), and (ii) explanations such as decision sets with discrete parameters, in which case we use a sampling-based approximation in conjunction with submodular optimization (Lakkaraju et al., 2016).

We evaluated our approach ROPE on real-world data from healthcare, criminal justice, and education, focusing on datasets that include some kind of distribution shift—i.e., individuals from two different subgroups (e.g., patients from two different counties). Our results demonstrate that the explanations constructed using ROPE are substantially more robust to distribution shifts than those generated by state-of-the-art post hoc explanation techniques such as LIME, SHAP, and MUSE. Furthermore, the fidelity of ROPE explanations is equal or higher than the fidelity of the explanations generated by state-of-the-art methods even on the original data distribution, thus demonstrating that ROPE improves robustness of explanations without sacrificing their fidelity on the original data distributions. In addition, we used synthetic data to analyze how the degree of distribution shift affects fidelity of the explanations constructed by our approach and other baselines. Finally, we performed an experiment where the "black box" models are themselves interpretable, and showed that ROPE explanations constructed based on shifted data are substantially more similar to the black box than the explanations output by other baselines.

2. Related Work

Post hoc explanations. Many approaches have been proposed to directly learn interpretable models (Breiman, 2017; Tibshirani, 1997; Letham et al., 2015; Lakkaraju et al., 2016; Caruana et al., 2015; Kim & Bastani, 2019); however, complex models such as deep neural networks and random forests typically achieve higher accuracy than simpler interpretable models (Ribeiro et al., 2016); thus, it is often desirable to use complex models and then construct post hoc explanations to understand their behavior.

A variety of post hoc explanation techniques have been proposed, which differ in their access to the complex model (i.e., black box vs. access to internals), scope of approximation (e.g., global vs. local), search technique (e.g., perturbation-based vs. gradient-based), explanation families (e.g., linear vs. non-linear), etc. For instance, in addition to LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017a), several other *local explanation* methods have been proposed that compute *saliency maps* which capture importance of each feature for an individual prediction by computing the gradient with respect to the input (Simonyan et al., 2014; Sundararajan et al., 2017; Selvaraju et al., 2017; Smilkov et al., 2017). An alternate approach is to provide a global explanation summarizing the black box as a whole (Lakkaraju et al., 2019a; Bastani et al., 2017), typically using an interpretable model.

There has also been recent work on exploring vulnerabilities of black box explanations (Adebayo et al., 2018; Slack et al., 2020; Lakkaraju & Bastani, 2020; Rudin, 2019; Dombrowski et al., 2019)—e.g., Ghorbani et al. (2019) demon-

strated that post hoc explanations can be unstable, changing drastically even with small perturbations to inputs. However, none of the prior work has studied the problem of constructing robust explanations.

Distribution shift. *Distribution shift* refers to settings where there is a mismatch between the training and test distributions. A lot of work in this space has focused on *covariate shift*, where the covariate distribution $p(x)$ changes but the outcome distribution $p(y \mid x)$ remains the same. This problem has been studied in the context of learning predictive models (Quionero-Candela et al., 2009; Jiang & Zhai, 2007). Proposed solutions include importance weighting (Shimodaira, 2000), invariant representation learning (Ben-David et al., 2007; Tzeng et al., 2017), online learning (Cesa-Bianchi & Lugosi, 2006), and learning adversarially robust models (Teo et al., 2007; Graepel & Herbrich, 2004; Decoste & Schölkopf, 2002). However, none of these approaches are applicable in our setting since they assume either that the underlying predictive model is not a black box, that data from the shifted distribution is available, or that the black box can be adaptively retrained.

Adversarial robustness. Due to the discovery that deep neural networks are not robust (Szegedy et al., 2014), there has been recent interest in *adversarial training* (Goodfellow et al., 2015; Bastani et al., 2016; Sinha et al., 2018; Shaham et al., 2018), which optimizes a minimax objective that captures the worst-case over a given set of perturbations to the input data. At a high level, these algorithms are based on gradient descent; at each gradient step, they solve an optimization problem to find the worst-case perturbation, and then compute the gradient at this perturbation. For instance, for L_∞ robustness (i.e., perturbations of bounded L_∞ norm), Goodfellow et al. (2015) propose to approximate the optimization problem using a single gradient step, called the *signed-gradient* update; Shaham et al. (2018) generalizes this approach to arbitrary norms. We propose a set of perturbations that capture our intuition about the kinds of distribution shifts that explanations should be robust to; for this set of shifts, we show how approximations along the lines of these previous approaches correspond to solving a linear program on every step to compute the gradient.

3. Our Framework

Here, we describe our framework for constructing robust explanations. We assume we are given a black box model $B^* : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^n$ is the space of covariates and $\mathcal{Y}$ is the space of labels. Our goal is to construct a global explanation for the computation performed by B^*. To construct such an explanation, one approach would be to learn an interpretable model that approximates B^*. In particular, given a family $\mathcal{E}$ of interpretable models, a distribution $p(x)$ over $\mathcal{X}$, and a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, this approach

constructs an explanation as follows:

$$\hat{E}(x) = \arg\min_{E \in \mathcal{E}} \mathbb{E}_{p(x)}[\ell(E(x), B^*(x))]. \qquad (2)$$

In other words, $\hat{E}$ minimizes the error (as defined by ℓ) relative to the black box B^*. Intuitively, if $\hat{E}$ is a good approximation of B^*, then the computation performed by B^* should be mirrored by the computation performed by $\hat{E}$.

The problem with Eq. 2 is that it only guarantees that $\hat{E}$ is a good approximation of B^* according to the distribution $p(x)$. If the underlying data distribution changes, then $\hat{E}$ may no longer be a good approximation of B^*.

3.1. Robust & Stable Explanations

To construct explanations that are robust to shifts in the data distribution p, we first consider the general setting where we are given a set of distribution shifts that we want our explanations to be robust to; we describe a practical choice in Section 3.2. We initially focus on distributional robustness; we connect it to adversarial robustness below.

Definition 3.1. Let p be a distribution over $\mathcal{X}$, and let $\delta \in \mathbb{R}^n$. The δ-*shifted distribution* is $p_\delta(x) = p(x - \delta)$.

In other words, p_δ places probability mass on covariates that are shifted by δ compared to p.

Definition 3.2. Let p be a distribution over $\mathcal{X}$. Given $\Delta \subseteq \mathbb{R}^n$, the set of Δ-*small shifts* is the set $\{p_\delta \mid \delta \in \Delta\}$ of δ-shifted distributions.

For computational tractability, we assume:

Assumption 3.3. The set Δ of shifts is a convex polytope.

Given a set of distribution shifts, our goal is to compute the best explanation that is robust to these shifts:

Definition 3.4. Given $\Delta \subseteq \mathbb{R}^n$, $s_0 \in \mathbb{N}$, and $\delta_{\max} \in \mathbb{R}_{>0}$, the *optimal robust explanation* for $(s_0, \delta_{\max})$-small shifts is

$$\hat{E} = \arg\min_{E \in \mathcal{E}} \max_{\delta \in \Delta} \mathbb{E}_{p_\delta(x)}[\ell(E(x), B^*(x))]. \qquad (3)$$

That is, $\hat{E}$ optimizes the worst-case loss over shifts p_δ. Computing the worst-case over shifts p_δ can be intractable; instead, we use an upper bound on the objective in Eq. 3.

Lemma 3.5. *We have*

$$\max_{\delta \in \Delta} \mathbb{E}_{p_\delta(x)}[\ell(E(x), B^*(x))]$$

$$\leq \mathbb{E}_{p(x)}\left[\max_{\delta \in \Delta} \ell(E(x + \delta), B^*(x + \delta))\right].$$

Proof: Note that

$$\max_{\delta \in \Delta} \mathbb{E}_{p_\delta(x)}[\ell(E(x), B^*(x))]$$

$$= \max_{\delta \in \Delta} \int_\mathcal{X} \ell(E(x), B^*(x))p(x - \delta)dx$$

$$= \max_{\delta \in \Delta} \int_\mathcal{X} \ell(E(x' + \delta), B^*(x' + \delta))p(x')dx'$$

$$\leq \int_\mathcal{X} \max_{\delta \in \Delta} \ell(E(x' + \delta), B^*(x' + \delta))p(x')dx',$$

$$\leq \mathbb{E}_{p(x)} \left[\max_{\delta \in \Delta} \ell(E(x + \delta), B^*(x + \delta)) \right] \quad \square$$

This lemma gives us a surrogate objective that we can optimize in place of the one in Eq. 3—i.e.,

$$\hat{E} = \arg\min_{E \in \mathcal{E}} \mathbb{E}_{p(x)} \left[\max_{\delta \in \Delta} \ell(E(x + \delta), B^*(x + \delta)) \right]. \quad (4)$$

In particular, this approach connects distributional robustness to adversarial robustness—Eq. 4 is the standard objective used to achieve adversarial robustness to input perturbations $\delta \in \Delta$ (Goodfellow et al., 2015).

3.2. General Class of Distribution Shifts

Next, we propose a choice of Δ that captures distributions shifts we believe to be of importance in practical applications. We begin with a concrete setting that motivates our choice, but our choice includes shifts beyond this setting.

In particular, consider the case where $\mathcal{X} = \{0, 1\}^d$ is a vector of indicators. Our intuition is that when examining an explanation, users often want to understand how the model predictions change when a handful of components of an input $x \in \mathcal{X}$ change.

For instance, this intuition captures the case of counterfactual explanations, where the goal is to identify a small number of covariates that can be changed to affect the outcome (Zhang et al., 2018). It also captures certain intuitions underlying fairness and causality, where we care about how the model changes when a covariate such as gender or ethnicity changes (Lakkaraju & Bastani, 2020; Rosenbaum & Rubin, 1983; Pearl, 2009). Finally, it also encompasses the shifts considered in measures of variable importance (Hastie et al., 2001)—in particular, variable importance measures how the explanation changes when a *single* component of the input x is changed.

We can use the following choice to capture our intuition:

$$\Delta_1 = \{\delta \in \{-1, 0, 1\}^n \mid \|\delta\|_0 \leq s_0\}$$

for $s_0 \in \mathbb{N}$. However, this set is nonconvex. We can approximate this constraint using the following set:

$$\Delta_2 = \{\delta \in \mathbb{R}^n \mid \|\delta\|_0 \leq s_0 \wedge \|\delta\|_\infty \leq 1\}.$$

In particular, the constraint $\|\delta\|_\infty \leq 1$ ensures that $-1 \leq \delta_i \leq 1$ for each $i \in \{1, ..., n\}$. Finally, we can replace the L_0 norm with the L_1 norm:

$$\Delta_3 = \{\delta \in \mathbb{R}^n \mid \|\delta\|_1 \leq s_0 \wedge \|\delta\|_\infty \leq 1\}. \quad (5)$$

This overapproximation is a heuristic based on the fact that the L_1 loss induces sparsity in regression (Tibshirani, 1997).

More generally, we consider a shift from p to a distribution p' such that p' places probability mass on the same inputs x as p, except a small number of components of x are systematically changed by a small amount:

$$\tilde{\Delta}(s_0, \delta_{\max}) = \{\delta \in \mathbb{R}^n \mid \|\delta\|_0 \leq s_0 \wedge \|\delta\|_\infty \leq \delta_{\max}\},$$

where $s_0 \in \mathbb{N}$ and $\delta_{\max} \in \mathbb{R}$—i.e., $\delta \in \tilde{\Delta}(s_0, \delta_{\max})$ is a sparse vector whose components are not too large. However, $\tilde{\Delta}(s_0, \delta_{\max})$ is nonconvex. As above, for computational tractability, we approximate it using

$$\Delta(s_0, \delta_{\max}) = \{\delta \in \mathbb{R}^n \mid \|\delta\|_1 \leq s_0 \wedge \|\delta\|_\infty \leq \delta_{\max}\}.$$

It is easy to see that $\tilde{\Delta}(s_0, \delta_{\max}) \subseteq \Delta(s_0, \delta_{\max})$, so this choice overapproximates the set of shifts. In particular, this choice $\Delta(s_0, \delta_{\max})$ is a polytope, so it satisfies Assumption 3.3. The set defined in Eq. 5 is $\Delta_3 = \Delta(s_0, 1)$.

A particular benefit of $\Delta(s_0, \delta_{\max})$ is that the marginal dependencies of B^* on a component x_i of an input $x \in \mathcal{X}$ is preserved in $\hat{E}$—i.e., if we unilaterally change x_i by a small amount, B^* and $\hat{E}$ change in the same way. Formally:

Proposition 3.6. *Suppose* $\mathcal{Y} = \mathbb{R}$, $\ell(y, y') = |y - y'|$, *and* $\Delta = \Delta(s_0, \delta_{max})$, *and consider an explanation* $\hat{E}$ *with error*

$$\mathbb{E}_{p(x)} \left[\max_{\delta \in \Delta} \ell(\hat{E}(x + \delta), B^*(x + \delta)) \right] \leq \epsilon.$$

Then, letting α *be the the one-hot encoding of* i *(i.e.,* $\alpha_i = 1$ *and* $\alpha_j = 0$ *if* $i \neq j$*), for any* $c \in \mathbb{R}$ *such that* $|c| \leq \delta_{max}$*,*

$$\mathbb{E}_{p(x)} \left[\left| (\hat{E}(x + c\alpha) - \hat{E}(x)) - (B^*(x + c\alpha) - B^*(x)) \right| \right]$$

$$\leq 2\epsilon.$$

Proof: Note that

$$\mathbb{E}_{p(x)} \left[\left| (\hat{E}(x + c\alpha) - \hat{E}(x)) - (B^*(x + c\alpha) - B^*(x)) \right| \right]$$

$$\leq \mathbb{E}_{p(x)} \left[\left| \hat{E}(x + c\alpha) - B^*(x + c\alpha) \right| \right]$$

$$+ \mathbb{E}_{p(x)} \left[\left| \hat{E}(x) - B^*(x) \right| \right]$$

$$\leq 2\epsilon \text{ since } c\alpha \in \Delta \quad \square$$

As shown in Section 1, this property is not satisfied by standard measures of fidelity, since an explanation with perfect fidelity (i.e., Eq. 1) may use completely different covariates from the black box.

3.3. Constructing Robust Linear Explanations

We consider the case where $\mathcal{E}$ is the space of linear functions, or more generally, any model family that can be optimized using gradient descent. Then, we can use *adversarial training* to optimize Eq. 4 (Goodfellow et al., 2015; Shaham et al., 2018). The key idea behind adversarial training is to learn a model $f^* \in \mathcal{F}$ that is robust with respect to a worst-case set of perturbations to the input data—i.e.,

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{p(x,y)} \left[\max_{\delta \in \Delta} \ell(f(x + \delta), y) \right].$$

We can straightforwardly adapt this formalism to our setting by replacing $\mathcal{F}$ with $\mathcal{E}$ and y with $B^*(x)$. In particular, suppose that $E_\theta \in \mathcal{E}$ is parameterized by $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^d$ and $J(\theta; x)$ is defined as follows:

$$J(\theta; x) = \ell(E_\theta(x), B^*(x)).$$

Then, Eq. 4 becomes

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \mathbb{E}_{p(x)} \left[\max_{\delta \in \Delta} J(\theta; x + \delta)) \right]. \qquad (6)$$

The adversarial training approach optimizes Eq. 6 by using stochastic gradient descent (Goodfellow et al., 2015; Shaham et al., 2018)—for a single sample $x \sim p(x)$, the stochastic gradient estimate of the objective in Eq. 6 is

$$\nabla_\theta \max_{\delta \in \Delta} J(\theta; x + \delta) \approx \nabla_\theta J(\theta; x + \delta^*),$$

where
$$\delta^* = \arg\max_{\delta \in \Delta} J(\theta, x + \delta). \qquad (7)$$

To solve Eq. 7, we use the Taylor approximation

$$J(\theta; x + \delta) \approx J(\theta; x) + \nabla_x J(\theta; x)^\top \delta.$$

Using this approximation, Eq. 7 becomes

$$\begin{aligned}
\delta^* &= \arg\max_{\delta \in \Delta} J(\theta, x + \delta) \\
&\approx \arg\max_{\delta \in \Delta} \left\{ J(\theta; x) + \nabla_x J(\theta; x)^\top \delta \right\} \\
&= \arg\max_{\delta \in \Delta} \nabla_x J(\theta; x)^\top \delta, \qquad (8)
\end{aligned}$$

where in the last line, we dropped the term $J(\theta; x)$ since it is constant with respect to δ. Since we have assumed Δ is a polytope, Eq. 8 is a linear program with free variables δ.

3.4. Constructing Robust Rule-Based Explanations

Here, we describe how we can construct robust rule-based explanations (Lakkaraju et al., 2016; Letham et al., 2015; Lakkaraju et al., 2019b)—e.g., decision sets (Lakkaraju et al., 2016; 2019b), decision lists (Letham et al., 2015),

decision trees (Quinlan, 1986). Any rule based model can be expressed as a decision set (Lakkaraju & Rudin, 2017), so we focus on these models.

Unlike explanations with continuous parameters, we can no longer use gradient descent to optimize Eq. 4. Instead, we optimize it using a sampling-based heuristic. We assume we are given a distribution $p_0(\delta)$ over shifts $\delta \in \Delta$. Then, we approximate the maximum in Eq. 4 using k samples:

$$\max_{\delta \in \Delta} F(\delta) \approx \max_{\delta^j \sim p_0(\delta)} F(\delta^j),$$

where $F(\delta)$ is a general objective and $j \in \{1, ..., k\}$. In particular, our optimization problem becomes

$$\hat{E} = \arg\min_{E \in \mathcal{E}} \mathbb{E}_{p(x)} \left[\max_{\delta^j \sim p_0(\delta)} \ell(E(x + \delta^j), B^*(x + \delta^j)) \right]. \tag{9}$$

Next, a decision set
$$E = \{(s_1, c_1), (s_2, c_2) \cdots (s_m, c_m)\} \subseteq S \times C$$
is a set of rules of the form (s, c) where s is a conjunction of predicates of the form (feature, operator, value) (e.g., age ≥ 45) and $c \in \mathcal{Y}$ is a label. Typically, we consider the case where $\mathcal{Y}$ is a finite set. Existing algorithms (Lakkaraju et al., 2019b; 2016) for constructing decision set explanations primarily optimize for the following three goals: (i) maximizing the *coverage* of E—i.e., for $x \in \mathcal{X}$, maximizing the probability that one of the rules $(s, c) \in E$ has a condition s that is satisfied by x, (ii) minimizing the *disagreement* between E and B^*—i.e., minimizing the probability that $E(x) \neq B^*(x)$, and (iii) minimizing the complexity of E—e.g., E has fewer rules. In particular, these algorithms optimize the following objective:

$$\hat{E} = \arg\max_{E \subseteq S \times C} \{-\text{disagree}(E) + \lambda \cdot \text{cover}(E)\} \qquad (10)$$
$$\text{subj. to } |E| \leq \alpha,$$

where
$$\text{disagree}(E) = \sum_{i=1}^m \mathbb{P}_{p(x)}(s_i(x) \to B^*(x) \neq c_i)$$
$$\text{cover}(E) = \mathbb{P}_{p(x)}(\exists(s, c) \in E \text{ s.t. } s(x) = \text{true}).$$

Here, we let $s(x) = \text{true}$ if x satisfies s and $s(x) = \text{false}$ otherwise. In disagree(E), the event in the probability says if predicate s_i applies to x, then $B^*(x) \neq c_i$.

To adapt this approach to solving Eq. 9, we modify the disagreement to take the worst-case over $\delta^j \sim p_0(\delta)$:

$$\text{disagree}(E)$$
$$= \sum_{i=1}^m \mathbb{P}_{p(x)} \left(s_i(x) \to \exists \delta^j \sim p_0(\delta) . B^*(x + \delta^j) \neq c_i \right).$$

where $j \in \{1, ..., k\}$. Here, we have used an approximation where we only check if s_i applies to the unperturbed input x; this choice enables our submodularity guarantee.

Theorem 3.7. *Suppose that $p(x) = Uniform(X_{train})$, where $X_{train} \subseteq \mathcal{X}$ is a training set, is the empirical training distribution. Then, the optimization problem Eq. 10 is non-monotone and submodular with cardinality constraints.*

Proof: To show non-monotonicity, it suffices to show that at least one term in the objective Eq. 10 is non-monotone. Every time a new rule is added, the value of disagree either remains the same or increases, since the newly added rule may potentially label new instances incorrectly, but does not decrease the number of instances already labeled incorrectly by previously chosen rules. Therefore, disagree$(A) \leq$ disagree(B) if $A \subseteq B$, so $-$disagree$(A) \geq -$disagree(B), which implies disagree term is non-monotone. Thus, the entire linear combination is non-monotone.

To prove that the objective in Eq. 10 is submodular, we need to: (i) introduce a (large enough) constant C into the objective function to ensure that $C -$ disagree(E) is never negative,[1] and (ii) prove that each of the its terms are submodular. The *cover* term is clearly submodular—i.e., more data points will be covered when we add a new rule to a smaller set of rules compared to a larger set. It is also easy to check that the *disagree* term is modular/additive (and therefore submodular). Lastly, the constraint in Eq. 10 is a cardinality constraint. □

Since the objective of Eqn. 10 is non-monotone and submodular with cardinality constraints (Theorem 3.7), exactly solving it is NP-Hard (Khuller et al., 1999). So, we use approximate local search algorithm (Lee et al., 2009) to optimize Eq. 10. This algorithm provides the best known theoretical guarantees for this class of problems—i.e., $(k + 2 + 1/k + \delta)^{-1}$, where k is the number of constraints ($k = 1$ in our case) and $\delta > 0$.

4. Experiments

As part of our evaluation, we first use real-world data to assess the robustness of the post hoc explanations constructed using our algorithm and compare it to state-of-the-art baselines. Second, on synthetic data, we analyze how varying the degree of distribution shift impacts the fidelity of our explanations. Third, we ascertain the correctness of explanations generated using our framework—in particular, in cases where the black box is also an interpretable model $B^* \in \mathcal{E}$, we study how closely the constructed explanations resemble the ground truth black box model.

4.1. Experimental Setup

Datasets. We analyzed three real-world datasets from criminal justice, healthcare, and education domains (Lakkaraju

---

[1]Note that adding such a constant does not impact the solution to the optimization problem.

et al., 2016). Our first dataset contains **bail outcomes** from two different state courts in the U.S. 1990-2009. It includes criminal history, demographic attributes, information about current offenses, and other details on 31K defendants who were released on bail. Each defendant in the dataset is labeled as either high risk or low risk depending on whether they committed new crimes when released on bail. Our second dataset contains **academic performance** records of about 19K students who were set to graduate high school in 2012 from two different school districts in the U.S. It includes information about grades, absence rates, suspensions, and tardiness scores from grades 6 to 8 for each of these students. Each student is assigned a class label indicating whether the student graduated high school on time. Our third dataset contains **electronic health records** of about 22K patients who visited hospitals in two different counties in California between 2010-2012. It includes demographic information, symptoms, current and past medical conditions, and family history of each patient. Each patient is assigned a class label which indicates whether the patient has been diagnosed with diabetes.

Distribution shifts. Each of our datasets contains two different subgroups—e.g., our bail outcomes dataset contains defendants from two different states. We randomly choose data from one of these subgroups (e.g., a particular state) to be the *training data*, and data from the other subgroup to be the *shifted data*. In particular, we apply each algorithm on the training data to construct explanations, and evaluate these explanations on the shifted data.

Our explanations. Our framework ROPE can be applied in a variety of configurations. We consider four: (i) ROPE logistic: We construct a single global logistic regression model using our framework to approximate any given black box. (ii) ROPE dset: We construct a single global decision set using our framework to approximate any given black box. (iii) ROPE logistic multi: We construct multiple *local* explanations. In particular, we first cluster the data into K subgroups (details below), and use ROPE to fit a robust logistic regression model to approximate the given black box for each subgroup. We also compute the centroid of each subgroup to serve as a representative sample. (iv) ROPE dset multi: Similar to ROPE logistic multi, except that we fit a decision set.

Baselines. We compare our framework to the following state-of-the-art post hoc explanation techniques: (i) LIME (Ribeiro et al., 2016), (ii) SHAP (Lundberg & Lee, 2017a), and (iii) MUSE (Lakkaraju et al., 2019b). LIME and SHAP are *model-agnostic*, *local explanation* techniques that explain an individual prediction of a black box by training a linear model on data near that prediction. LIME and SHAP can be adapted to produce global explanations of any given black box using a submodular pick procedure (Ribeiro

Algorithms	Bail			Academic			Health		
	Train	Shift	% Drop	Train	Shift	% Drop	Train	Shift	% Drop
LIME	0.79	0.64	18.99%	0.68	0.57	16.18%	0.81	0.69	14.81%
SHAP	0.76	0.66	13.16%	0.67	0.59	11.94%	0.83	0.68	18.07%
MUSE	0.75	0.59	21.33%	0.66	0.51	22.73%	0.79	0.61	22.78%
ROPE logistic	0.61	0.59	**3.28%**	0.57	0.57	**0.00%**	0.70	0.68	**2.86%**
ROPE dset	0.64	0.61	4.69%	0.65	0.63	3.08%	0.73	0.69	5.48%
ROPE logistic multi	0.79	0.74	6.33%	0.70	0.69	1.43%	0.82	0.76	7.32%
ROPE dset multi	**0.82**	**0.77**	6.1%	**0.73**	**0.71**	2.74%	**0.84**	**0.78**	7.14%

Table 1. Fidelity values of all the explanations are reported on both training data and shifted data, along with percentage drop in fidelity from training data to shifted data. Smaller values of percentage drop correspond to more robust explanations.
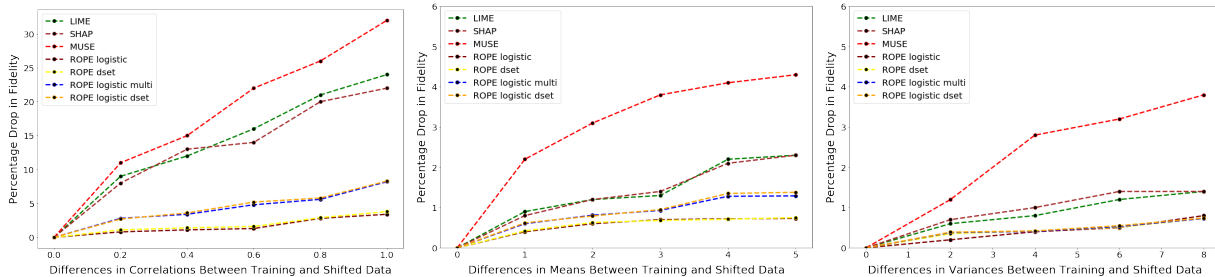

Figure 1. Impact of changes in covariate correlations (left), means (middle), and variances (right) on percentage drop in fidelities. Lower values of percentage drop indicate higher robustness. Standard errors too small to be included.

et al., 2016), which chooses a few representative points from the dataset and combines their corresponding local models to form a global explanation. In our evaluation, we use the global explanations of LIME and SHAP constructed using this technique. MUSE is a *model-agnostic*, *global explanation* technique; it provides global explanations in the form of two-level decision sets.

Parameters. In case of LIME, SHAP, ROPE logistic multi, and ROPE dset multi, there is a parameter K which corresponds to the number of local explanations that need to be generated; K can also be thought as the number of subgroups in the data. We use Bayesian Information Criterion (BIC) to choose K. For a given dataset, we use the same K for all these techniques to ensure they construct explanations of the same size. For MUSE, we set all the parameters using the procedure in Lakkaraju et al. (2019b); to ensure these explanations are similar in size to the others, we fix the number of outer rules to be K. Finally, when using ROPE to construct rule-based explanations, there is a term λ in our objective (Eq. 10); we fix $\lambda = 5$.

Black boxes. We generate post hoc explanations of deep neural networks (DNNs), gradient boosted trees, random forests, and SVMs. Here, we present results for a 5-layer DNN; remaining results are included in the Appendix. Results presented below are representative of those for other model families.

Metrics. We use fidelity to measure performance—i.e., the fraction of inputs x in the given dataset for which

$\hat{E}(x) = B^*(x)$ (Lakkaraju et al., 2019b). Fidelity is straightforward to compute for MUSE, ROPE logistic, and ROPE dset since they construct an explanation in the form of a single interpretable model. However, the explanations constructed by LIME, SHAP, ROPE logistic multi, and ROPE dset multi consist of a collection of local models. In these cases, we need to determine which local model to use for each input x. By construction, each local model $\hat{E}_i$ is associated with a representative input x_i, for $i \in \{1, ..., K\}$. Thus, we compute the distance $\|x - x_i\|$ for each i, and return $\hat{E}_{i*}(x)$ where x_{i*} is closest to x.

4.2. Robustness to Real Distribution Shifts

We assess the robustness of explanations constructed using each approach on real-world datasets. In particular, we compute the fidelity of the explanations on both the training data and the shifted data, as well as the percentage change between the two. A large drop in fidelity from the training data to the shifted data indicates that the explanation is not robust. Ideally, explanations should have high fidelity on both the training data (indicating it is a good approximation of the black box model) and on the shifted data (indicating it is robust to distribution shift).

Results for all three real-world datasets are shown in Table 1. As can be seen, all the explanations constructed using our framework ROPE have a much smaller drop in fidelity (0% to 7%) compared to those generated using the baselines. These results demonstrate that our approach significantly

improves robustness. MUSE explanations have the largest percentage drop (21% to 23%), likely because MUSE relies entirely on the training data. In contrast, both LIME and SHAP employ input perturbations when constructing explanations (Ribeiro et al., 2016; Lundberg & Lee, 2017b), resulting in somewhat increased robustness compared to MUSE. Nevertheless, LIME and SHAP still demonstrate a considerable drop (13% to 19%), so they are still not very robust. The reason is because these approaches do not optimize a minimax objective that encodes robustness such as ours. Thus, these results validate our approach.

In addition, Table 1 shows the actual fidelities on both training data and shifted data. As can be seen, the fidelities of ROPE logistic and ROPE dset are lower than the other approaches; these results are expected since ROPE logistic and ROPE dset only use a single logistic regression and a single decision set model, respectively, to approximate the entire black box. On the other hand, ROPE logistic multi and ROPE dset multi achieve fidelities that are equal or better than the other baselines. These results demonstrate that ROPE achieves robustness without sacrificing fidelity on the original training distribution. Thus, our approach *strictly outperforms* the baseline approaches.

4.3. Impact of Degree of Distribution Shift on Fidelity

Next, we assess how different kinds of distribution shifts impact the fidelity of explanations constructed using our framework and the baselines using synthetic data. We study the effects of three different kinds of shifts: (i) changes in the correlations between different components of the covariates, (ii) changes in the means of the covariates, and (iii) changes in the variances of the covariates.

Shifts in correlation. We first describe our study for shifted data of type (i) above. We generate a synthetic dataset with 5K samples. The covariate dimension is randomly chosen between 2 and 10. Each data point is sampled $x \sim \mathcal{N}(\mu, \Sigma)$, where $\mu_i = 0$, $\Sigma_{ii} = 1$ and $\Sigma_{ij} = \beta$, where β is uniformly random in $[-1, 1]$—i.e., the correlation between any two components of the covariates is β. The label for each data point is chosen randomly. We train a 5 layer DNN B^* on this dataset, and construct explanations for B^*.

To generate shifted data, we generate a new dataset with the same approach as above but using a different correlation $\beta' = \beta + \alpha$, where we vary α. Then, we compute the percentage drop in fidelity of the explanations from the training data to each of the shifted datasets. We show results averaged over 100 runs in Figure 1 (left); the x-axis shows $|\alpha|$, and the y-axis shows the percentage drop. As can be seen, MUSE exhibits the highest drop in fidelity, followed closely by LIME and SHAP. In contrast, the ROPE explanations are substantially more robust, incurring less than a 10% drop in fidelity.

Mean shifts. For shifts of type (ii) above, we follow the same procedure, except we use $\beta = 0$ for both the training and shifted datasets (i.e., uncorrelated covariates), and choose μ randomly in $[-5, 5]$. To generate shifted data, we use a different $\mu' = \mu + \alpha$. Results averaged across 100 runs are shown in Figure 1 (middle). ROPE is still the most robust, though LIME and SHAP are closer to ROPE than to MUSE. Explanations generated by MUSE are not robust even to small changes in covariate means.

Variance shifts. For shifts of type (iii) above, we follow the same procedure, except we use $\beta = 0$, and choose $\Sigma_{ii} = \sigma$, where σ is randomly chosen from $[1, 10]$. To generate shifted data, we use a different $\sigma' = \sigma + \alpha$. Results averaged across 100 runs are shown in Figure 1 (right). The results are similar to the case of mean shifts.

4.4. Evaluating Correctness of Explanations

Here, we evaluate the *correctness* of the constructed explanations—i.e., how closely an explanation resembles the black box. To this end, we first train "black box" models $B^* \in \mathcal{E}$ that are interpretable using the *training data* from each of our real-world datasets. Then, we construct an explanation $\hat{E}$ for B^* using the *shifted data*. If $\hat{E}$ resembles B^* structurally, then the underlying explanation technique is generating explanations that are correct despite being constructed based on shifted data.

Logistic regression black box. We first train a logistic regression (**LR**) "black box" B^*, and then use LIME, SHAP, ROPE logistic, and ROPE logistic multi to construct explanations $\hat{E}$ for B^*. We define the *coefficient mismatch* to measure the correctness. For ROPE logistic, it is computed as $\|\hat{E} - B^*\|$—i.e., the L_2 distance between the weight vectors of $\hat{E}$ and B^*; smaller distances mean the explanation more closely resembles the black box. The remaining approaches construct multiple logistic regression models—one $\hat{E}_i$ for each representative input x_i, for $i \in \{1, ..., K\}$. To measure the coefficient mismatch, we assign a weight w_i to each x_i that equals the fraction of inputs x that are assigned to x_i (i.e., x_i is the closest representative). Then, we measure coefficient mismatch as $\sum_{i=1}^{K} w_i \cdot \|\hat{E}_i - B^*\|$.

We also consider the case where B^* is a collection of multiple logistic regression (**Multiple LR**) models—one B_i^* for each of the K subgroups. We construct explanations using LIME, SHAP, ROPE logistic, and ROPE logistic multi, and measure the coefficient mismatch as $\sum_{i=1}^{K} w_i \cdot \|\hat{E}_i - B_e^*\|$; In case of ROPE logistic, $\hat{E}_i = \hat{E}$.

Results for the bail dataset are shown in Table 2. When B^* is a single logistic regression (LR), ROPE logistic and ROPE logistic multi explanations achieve the best performance and are about 38.2% more structurally similar to B^* than the baselines. When B^* is multiple logistic regressions (Mul-

Algorithms	Black Boxes					
	LR Coefficient Mismatch	**Multiple LR** Coefficient Mismatch	**DS** Rule Match	**DS** Feature Match	**Multiple DS** Rule Match	**Multiple DS** Feature Match
LIME	4.37	5.01	–	–	–	–
SHAP	4.28	4.96	–	–	–	–
MUSE	–	–	4.39	11.81	4.42	9.23
ROPE logistic	**2.69**	4.73	–	–	–	–
ROPE dset	–	–	**6.23**	15.87	4.78	11.23
ROPE logistic multi	2.70	**2.93**	–	–	–	–
ROPE dset multi	–	–	6.25	**16.18**	**7.09**	**16.78**

Table 2. Correctness of explanations on the bail dataset: Smaller coefficient mismatch and larger rule/feature match are better—i.e., the explanation more closely resembles the black boxes. ROPE dset multi and ROPE logistic multi uniformly outperform all the baselines.

tiple LR), the coefficient mismatch of ROPE logistic multi is at least 38.05% lower than the baselines. We obtained similar results for the academic and health datasets.

Decision set black box. As before, we train a decision set (**DS**) "black box" B^* on the real-world training data, and then construct an explanation $\hat{E}$ based on the shifted data using MUSE, ROPE dset, and ROPE dset multi. We consider two measures of correctness for ROPE dset: (i) *rule match*: the number $d_r(\hat{E}, B^*)$ of rules present in both $\hat{E}$ and B^*, and (ii) *feature match*: the number of features $d_f(\hat{E}, B^*)$ present in both $\hat{E}$ and B^*. As before, for ROPE dset multi and MUSE, we use the weighted measure $\sum_{i=1}^{K} w_i \cdot d(\hat{E}_i, B^*)$, where $d = d_r$ and $d = d_f$ for rule match and feature match respectively. Higher rule and feature matches indicate that $\hat{E}$ better resembles B^*. We also consider the case where B^* consists of multiple decision sets (**Multiple DS**)—one B_i^* for each of the K subgroups.

On the bail dataset, ROPE dset multi has 42.3% (resp., 60.4%) higher rule match than MUSE when B^* corresponds to DS (resp., Multiple DS), and has at least 37% higher feature match than the baselines.

4.5. Evaluating Stability of Explanations

Finally, we evaluate the *stability* of the constructed explanations—i.e., how much do the explanations change if the input data is perturbed by a small amount. To this end, we first generate a synthetic dataset D with 5000 samples as described in Section 4.3. Then, we generate the perturbed dataset D' by adding a small amount of Gaussian noise to data points—i.e., $x' = x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 0.2)$.[2] We then train LR, Multiple LR, DS, and Multiple DS "black boxes" B^*, and use LIME, SHAP, ROPE logistic, ROPE logistic multi, ROPE dset, ROPE dset multi to construct explanations $\hat{E}$ for the corresponding B^*. We use the orig-

inal dataset D both to train each black box B^* as well as to construct its explanation $\hat{E}$. Then, for each black box B^*, we use the perturbed dataset D' to construct an additional explanation $\hat{E}'$. Since D' is obtained by making small changes to instances in D, E and E' should be structurally similar if the explanation technique used to construct them generates stable explanations.

We measure structural similarity of E and E'—similar to the results in Table 2, we compute their coefficient mismatch in the case of LR and Multiple LR, and rule and feature match in case of DS and Multiple DS. We find that explanations E and E' constructed using ROPE are 18.21% to 21.08% more structurally similar than those constructed using LIME, SHAP, or MUSE. Thus, our results demonstrate that ROPE explanations are much more stable than those constructed using baselines.

5. Conclusions & Future Work

In this paper, we proposed a novel framework based on adversarial training for constructing explanations that are robust to distribution shifts and are stable. Experimental results have demonstrated that our framework can be used to construct explanations that are far more robust to distribution shifts than those constructed using other state-of-the-art techniques. Our work paves way for several interesting future research directions. First, it would be interesting to extend our techniques to other classes of explanations such as saliency maps. Second, it would also be interesting to design adversarial attacks that can potentially exploit any vulnerabilities in our framework to generate unstable and incorrect explanations.

Acknowledgements

---

[2]We experimented with other choices of variance in the range $(0.2, 1.0]$ and found similar results.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

Bastani, H. Predicting with proxies: Transfer learning in high dimension. *arXiv preprint arXiv:1812.11097*, 2018.

Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A., and Criminisi, A. Measuring neural net robustness with constraints. In *Advances in neural information processing systems*, pp. 2613–2621, 2016.

Bastani, O., Kim, C., and Bastani, H. Interpretability via model extraction. *arXiv preprint arXiv:1706.09773*, 2017.

Ben-David, S., Blitzer, J., Crammer, K., and Pereira, F. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pp. 137–144, 2007.

Breiman, L. *Classification and regression trees*. Routledge, 2017.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Knowledge Discovery and Data Mining (KDD)*, 2015.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Decoste, D. and Schölkopf, B. Training invariant support vector machines. *Machine learning*, 46(1-3):161–190, 2002.

Dombrowski, A.-K., Alber, M., Anders, C. J., Ackermann, M., Müller, K.-R., and Kessel, P. Explanations can be manipulated and geometry is to blame. *arXiv preprint arXiv:1906.07983*, 2019.

Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Graepel, T. and Herbrich, R. Invariant pattern recognition by semidefinite programming machines. In *NIPS*, pp. 33, 2004.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer New York Inc., 2001.

Jiang, J. and Zhai, C. A two-stage approach to domain adaptation for statistical classifiers. In *CIKM*, pp. 401–410, 2007.

Khuller, S., Moss, A., and Naor, J. S. The budgeted maximum coverage problem. *Information Processing Letters*, 70(1):39–45, 1999.

Kim, C. and Bastani, O. Learning interpretable models with causal guarantees. *arXiv preprint arXiv:1901.08576*, 2019.

Lakkaraju, H. and Bastani, O. "how do i fool you?": Manipulating user trust via misleading black box explanations. In *AIES*, 2020.

Lakkaraju, H. and Rudin, C. Learning cost-effective and interpretable treatment regimes. In *Artificial Intelligence and Statistics*, pp. 166–175, 2017.

Lakkaraju, H., Bach, S. H., and Leskovec, J. Interpretable decision sets: A joint framework for description and prediction. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 1675–1684, 2016.

Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. Faithful and customizable explanations of black box models. In *AAAI Conference on Artificial Intelligence, Ethics, and Society (AIES)*, 2019a.

Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. Faithful and customizable explanations of black box models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 131–138. ACM, 2019b.

Lee, J., Mirrokni, V. S., Nagarajan, V., and Sviridenko, M. Non-monotone submodular maximization under matroid and knapsack constraints. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, pp. 323–332, 2009.

Letham, B., Rudin, C., McCormick, T. H., and Madigan, D. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 2015.

Lipton, Z. C. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

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.), *Neural Information Processing Systems (NIPS)*, pp. 4765–4774. Curran Associates, Inc., 2017a.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017b.

Namkoong, H. and Duchi, J. C. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in neural information processing systems*, pp. 2208–2216, 2016.

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

Quinlan, J. R. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., and Lawrence, N. D. *Dataset Shift in Machine Learning*. The MIT Press, 2009. ISBN 0262170051, 9780262170055.

Ribeiro, M. T., Singh, S., and Guestrin, C. "why should i trust you?": Explaining the predictions of any classifier. In *Knowledge Discovery and Data Mining (KDD)*, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206, 2019.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Shaham, U., Yamada, Y., and Negahban, S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. *Neurocomputing*, 307:195–204, 2018.

Shimodaira, H. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations (ICLR)*, 2014.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *ICLR*, 2018.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. How can we fool lime and shap? adversarial attacks on post hoc explanation methods. 2020.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., and Wattenberg, M. SmoothGrad: removing noise by adding noise. In *ICML Workshop on Visualization for Deep Learning*, 2017.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning (ICML)*, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks, 2014.

Teo, C. H., Globerson, A., Roweis, S. T., and Smola, A. J. Convex learning with invariances. In *NIPS*, pp. 1489–1496, 2007.

Tibshirani, R. The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395, 1997.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7167–7176, 2017.

Zhang, X., Solar-Lezama, A., and Singh, R. Interpreting neural network judgments via minimal, stable, and symbolic corrections. In *Advances in Neural Information Processing Systems*, pp. 4874–4885, 2018.