

---

# Towards the Unification and Robustness of Perturbation and Gradient Based Explanations

---

Sushant Agarwal<sup>1</sup> Shahin Jabbari<sup>2</sup> Chirag Agarwal<sup>2\*</sup> Sohini Upadhyay<sup>2\*</sup>  
Zhiwei Steven Wu<sup>3</sup> Himabindu Lakkaraju<sup>2</sup>

## Abstract

As machine learning black boxes are increasingly being deployed in critical domains such as healthcare and criminal justice, there has been a growing emphasis on developing techniques for explaining these black boxes in a post hoc manner. In this work, we analyze two popular post hoc interpretation techniques: SmoothGrad which is a gradient based method, and a variant of LIME which is a perturbation based method. More specifically, we derive explicit closed form expressions for the explanations output by these two methods and show that they both converge to the same explanation in expectation, i.e., when the number of perturbed samples used by these methods is large. We then leverage this connection to establish other desirable properties, such as robustness, for these techniques. We also derive finite sample complexity bounds for the number of perturbations required for these methods to converge to their expected explanation. Finally, we empirically validate our theory using extensive experimentation on both synthetic and real world datasets.<sup>1</sup>

## 1. Introduction

Over the past decade, predictive models are increasingly being considered for deployment in high-stakes domains such as healthcare and criminal justice. However, the successful adoption of predictive models in these settings depends heavily on how well decision makers (e.g., doctors, judges)

can understand and consequently trust their functionality. Only if decision makers have a clear picture of the behavior of these models can they assess when and how much to rely on these models, detect potential biases in them, and develop strategies for improving them (Doshi-Velez & Kim, 2017). However, the increasing complexity as well as the proprietary nature of predictive models is making it challenging to understand these complex black boxes, thus motivating the need for tools and techniques that can explain them in a faithful and human interpretable manner.

Several techniques have been recently proposed to construct *post hoc explanations* of complex predictive models. While these techniques differ in a variety of ways, they can be broadly categorized into *perturbation vs. gradient based techniques*, based on the approaches they employ to generate explanations. For instance, LIME and SHAP (Ribeiro et al., 2016; Lundberg & Lee, 2017) are called *perturbation based* methods because they leverage perturbations of individual instances to construct interpretable local approximations (e.g., linear models), which in turn serve as explanations of individual predictions of black box models. On the other hand, SmoothGrad, Integrated Gradients and GradCAM (Simonyan et al., 2014; Sundararajan et al., 2017; Selvaraju et al., 2017; Smilkov et al., 2017) are referred to as *gradient based* methods since they leverage gradients computed at individual instances to explain predictions of complex models. Recent research has focused on empirically analyzing the behavior of perturbation and gradient based post hoc explanations. For instance, several works (Ghorbani et al., 2019; Slack et al., 2020a; Dombrowski et al., 2019; Adebayo et al., 2018; Alvarez-Melis & Jaakkola, 2018) demonstrated that explanations generated using perturbation based techniques such as LIME and SHAP may not be robust, i.e., the resulting explanations may change drastically with very small changes to the instances. Furthermore, Adebayo et al. (2018) showed that gradient based methods such as SmoothGrad and GradCAM may not generate interpretations that are faithful to the underlying models.

While several perturbation and gradient based explanation techniques have been proposed in literature and the aforementioned works have empirically examined their behavior,

---

\*Equal contribution <sup>1</sup>David R. Cheriton School of Computer Science, University of Waterloo, Waterloo, ON, Canada <sup>2</sup>Department of Computer Science, Harvard University, Cambridge, MA, USA <sup>3</sup>School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: Sushant Agarwal <sushant.agarwal@uwaterloo.ca>, Shahin Jabbari <jabbari@seas.harvard.edu>.

there is very little work that focuses on developing a rigorous theoretical understanding of these techniques and systematically exploring the connections between them. Recently, [Levine et al. \(2019\)](#) theoretically and empirically analyzed the robustness of a sparsified version of SmoothGrad but their analysis requires several key modifications to the original SmoothGrad, whereas we study SmoothGrad in its original form. Even more recently, [Garreau & von Luxburg \(2020\)](#) provided closed form solutions for and theoretically analysed Tabular LIME (LIME restricted to tabular data). We study a simpler, non-discretized variant of LIME that benefits by exhibiting several desirable properties, such as being provably robust, unlike the setting in [Garreau & von Luxburg \(2020\)](#). In addition, these works do not explore deeper connections between the two classes of techniques.

In this work, we initiate a study to unify perturbation and gradient based post hoc explanation techniques. To the best of our knowledge, this work makes the first attempt at establishing connections between these two popular classes of explanation techniques. More specifically, we make the following key contributions:

- We analyze two popular post hoc explanation methods – SmoothGrad (gradient based) and a variant of LIME (perturbation based) for continuous data that we refer to as Continuous LIME, or C-LIME for short. We derive explicit closed form expressions for the explanations output by these methods and demonstrate that they converge to the same output (explanation) in expectation, i.e., when the number of perturbed samples used by these methods is large.
- We then leverage this equivalence result to establish other desirable properties of these methods. More specifically, we prove that SmoothGrad and C-LIME satisfy Lipschitz continuity and are therefore robust to small changes in the input when the number of perturbed samples is large. This work is the first to demonstrate that a variant of LIME is provably robust.
- We also derive finite sample complexity bounds for the number of perturbed samples required for SmoothGrad and C-LIME to converge to their expected output.
- Finally, we prove that both SmoothGrad and C-LIME satisfy other interesting properties such as linearity.

We carry out extensive experimentation with synthetic and real world datasets from diverse domains such as online shopping and finance to analyze the behavior of SmoothGrad and C-LIME. Our empirical results not only validate our theoretical claims but also provide other interesting insights. We observe that both SmoothGrad and C-LIME need far fewer perturbations (than what our theory predicts) in practice to converge to their expected output and/or exhibit

robustness. SmoothGrad requires even fewer perturbations than C-LIME to be robust and also converges faster than C-LIME.

We also analyze the effects of other parameters such as the variance of the perturbed samples on the convergence as well as robustness of these methods, and find that smaller values of variance enable these methods to converge faster and exhibit robustness even with fewer perturbed samples.

## 2. Preliminaries

Let us consider a complex function  $f : X \rightarrow Y$ , where  $X \subseteq \mathbb{R}^d$  for some  $d \in \mathbb{N}$  and  $Y \subseteq \mathbb{R}$ . In this section, we provide an overview of two popular post hoc explanation techniques, namely, SmoothGrad and LIME. Both SmoothGrad and LIME are local explanation techniques i.e., they explain individual predictions  $f(x)$  of a given model  $f$ . Furthermore, both these methods fall under the broad category of feature attribution methods which determine the influence of each feature on a given prediction  $f(x)$ . Below, we describe these methods in detail. We then lay down the setting and assumptions for this work.

### 2.1. SmoothGrad

Gradient based explanations are designed to explain predictions  $f(x)$  for any  $x \in X$ , by computing the derivative of  $f(x)$  with respect to each feature of  $x$  ([Ancona et al., 2018](#)).

Vanilla gradient based explanations are often noisy, highlighting random features and ignoring important ones ([Adebayo et al., 2018](#)). A convincing explanation for the noise in gradient based saliency maps is that the derivative of  $f$  may fluctuate sharply at small scales. Hence, the gradient at any given point is less meaningful than the average of gradients at local neighboring points. This idea has led to SmoothGrad ([Smilkov et al., 2017](#)), which in practice reduces the noise in explanations compared to vanilla gradients.

More concretely, let  $S(x)$  (or simply  $S$ ) denote a set of inputs in the neighborhood of  $x$ . Using  $S$ , the (empirical) explanation of SmoothGrad for  $f$  at point  $x$  is defined to be

$$\text{SG}_S^f(x) = \frac{1}{|S|} \sum_{a \in S} \nabla f(a),$$

where  $\nabla f$  is the gradient of  $f$ . When  $S$  is drawn from a distribution  $P(x)$  (or simply  $P$ ), the expected explanation of SmoothGrad for  $f$  at input  $x$  can be defined by replacing the sample average with expectation:

$$\text{SG}_P^f(x) = \mathbb{E}_{a \sim P} [\nabla f(a)].$$

Throughout this paper, we use subscripts  $P$  and  $S$  to distinguish between expected values and empirical averages in our quantities of interest.

## 2.2. LIME

Another class of explainability models are perturbation based techniques. LIME is a popular perturbation based method that aims to explain the prediction  $f(x)$ , by learning an interpretable model that approximates  $f$  locally around  $x$  (Ribeiro et al., 2016). To obtain a local explanation, LIME creates perturbed examples in the local neighbourhood of  $x$ , observes the predictions of  $f$  for these examples, and trains an interpretable model on these labeled examples.

More concretely, let  $S$  denote a set of inputs in the neighborhood of  $x$  and  $\pi : X \times X \rightarrow \mathbb{R}^{\geq 0}$  a distance metric over  $X$ . Let  $G$  be a class of explanations (or models) and for any  $g \in G$ ,  $\Omega(g)$  denote the complexity of  $g$  e.g., the complexity of a linear explanation can be measured as the number of non-zero weights. The (empirical) explanation of LIME can be written as

$$\text{LIME}_S^f(x) = \arg \min_{g \in G} \{L_x(f, g, S, \pi) + \Omega(g)\},$$

where the loss function  $L$  is defined as

$$L_x(f, g, S, \pi) = \frac{1}{|S|} \sum_{a \in S} \pi(x, a) [f(a) - g(a)]^2.$$

When  $S$  is drawn from a distribution  $P$ , the expected explanation of LIME for  $f$  at input  $x$  can be written by replacing the sample average in the loss function with expectation. We call this quantity  $\text{LIME}_P^f(x)$ .

**Remark.** *The default implementation of LIME has an additional discretization step for the features before optimization. Our definition of LIME here ignores this discretization.*

## 2.3. Our Setting and Assumptions

In our setting, we assume  $X = \mathbb{R}^d$ , i.e., we assume the features to be continuous. Different choices of  $Y$  in our setting lead to different learning settings.  $Y = \mathbb{R}$  leads to regression.  $Y = [0, 1]$  corresponds to (binary or multi-class) classification when  $f(x)$  is interpreted as the probability of  $f$  belonging to a specific class.

For any point  $x$ , in both SmoothGrad and our variant of LIME (discussed below), we assume the sample  $S$  in the neighborhood of  $x$  is drawn from  $\mathcal{N}(x, \Sigma)$ , where  $\Sigma = \sigma^2 \mathbf{I}$  for some  $\sigma^2 > 0$ . This is a standard choice in practice (Garreau & von Luxburg, 2020; Smilkov et al., 2017).<sup>2</sup>

**C-LIME.** We use a variant of LIME for continuous features which we refer to as Continuous-LIME (or simply C-LIME). For any given function  $f$ , input point  $x$  and a sample  $S$  of inputs in the neighborhood of  $x$ , the (empirical)

<sup>2</sup>Many of our results hold for arbitrary  $\Sigma$ . We point these out explicitly when we discuss our results.

explanation of C-LIME can be written as

$$\text{C-LIME}_S^f(x) = \arg \min_{g \in G} \frac{1}{|S|} \sum_{a \in S} [f(a) - g(a)]^2,$$

where  $G$  is the class of linear models.

We now highlight the main differences between LIME and C-LIME: (i) C-LIME assumes that the distance metric  $\pi$  is a constant function that always outputs 1. Since C-LIME operates on continuous features and uses a Gaussian distribution centered at  $x$  to sample perturbations (unlike LIME which samples perturbations uniformly at random), the resulting perturbations are more likely to be closer to  $x$  and do not need to be weighted when fitting a local linear model. (ii) While LIME allows for a general class of simple explanations (or models)  $G$ , we restrict ourselves only to linear models for C-LIME since it focuses on continuous features. (iii) Lastly, we exclude the regularizer  $\Omega$  from C-LIME i.e., we set  $\Omega(g) = 0$  for all  $g \in G$ . Note that the paper that proposes LIME also advocates for enforcing sparsity by first carrying out a feature selection procedure to determine the top  $K$  features and then learning the corresponding weights via least squares (Ribeiro et al., 2016). See the full version for a discussion of the regularised version of LIME. Finally, for ease of exposition, throughout we assume the output of C-LIME is simply the weights on each feature, and ignore the intercept term. This can be done without loss of generality by centering. Moreover, we are only interested in the learned weights for the features, and not the intercept. For completeness, all of our proofs are written for the case that the intercept is present.

When clear from context, we refer to the expected output of SmoothGrad and C-LIME for explaining a function  $f$  at point  $x$  using a Gaussian distribution with mean  $x$  and covariance matrix  $\Sigma$  as  $\text{SG}_\Sigma^f$  and  $\text{C-LIME}_\Sigma^f$ , respectively. Moreover, when it is clear from the context we replace the subscript  $S$  to  $n$  in all of our quantities of interest to simply emphasize that the size of sample  $S$  is  $n$ .

## 3. Equivalence and Robustness

As our first contribution, in Section 3.1, we show that SmoothGrad and C-LIME provide identical explanations in expectation. This establishes a novel connection between gradient based and perturbation based explanation methods, which are often studied independently. Using this connection, in Section 3.2, we prove that both SmoothGrad and C-LIME are robust, i.e., the explanations provided by these methods for nearby points do not vary significantly.

### 3.1. Equivalence

As our first result, in Theorem 1, we show that the expected output of SmoothGrad and C-LIME are the same for any

function at any given input provided that SmoothGrad and C-LIME use the same Gaussian distribution for gradient computation and perturbations, respectively.

**Theorem 1.** *Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be a function. Then, for any  $x \in X$  and any invertible covariance matrix  $\Sigma \in \mathbb{R}^d \times \mathbb{R}^d$*

$$SG_{\Sigma}^f(x) = \text{LIME}_{\Sigma}^f(x) = \Sigma^{-1} \text{cov}(a, f(a)),$$

where  $a$  is a random input drawn from  $\mathcal{N}(x, \Sigma)$ , and  $\text{cov}(a, f(a))$  is a vector with the  $i$ 'th entry corresponding to the covariance of  $f(a)$  and  $i$ 'th feature of  $a$ .

*Proof sketch.* We separately derive closed forms for SmoothGrad and C-LIME. For SmoothGrad we apply a multivariate version of Stein's Lemma (Landsman & Nešlehová, 2008; Liu, 1994). The proof for C-LIME uses calculus, and recovers the explanation of C-LIME by differentiation and solving for the solution where the gradient is 0. See the full version for more details.  $\square$

We point out that Theorem 1 holds for any covariance matrix and does not require the covariance matrix to be diagonal. Furthermore, we note that the closed forms for both SmoothGrad and C-LIME have a nice structure. For a diagonal  $\Sigma$ , the  $i$ 'th coefficient of  $\text{SmoothGrad}^f(x)$  and  $\text{C-LIME}^f(x)$  depends only on the covariance of  $f$  and the  $i$ 'th feature. In particular, when  $\Sigma = \sigma^2 \mathbf{I}$ , then the  $i$ 'th coefficient is simply  $\text{cov}(f(a), a_i) \sigma^{-2}$ . This term captures the dependence of  $f$  on the  $i$ 'th feature of the input.

### 3.2. Robustness

Many interpretability methods come with the drawback that they are very sensitive to the choice of the point where the prediction of the function is going to be explained (Alvarez-Melis & Jaakkola, 2018; Ghorbani et al., 2019). It is hence desirable to have robust explainability methods where two nearby points with similar labels have similar explanations.

In this section we show that both SmoothGrad and C-LIME are robust. The notion of robustness we use is Lipschitz continuity which is formally defined as follows.

**Definition 1.** *A function  $h : \mathbb{R}^{d_1} \rightarrow \mathbb{R}^{d_2}$  for  $d_1, d_2 \in \mathbb{N}$  is  $L$ -Lipschitz if there exists a universal constant  $L \in \mathbb{R}^{>0}$ , such that  $\|h(x) - h(x')\|_2 \leq L \|x - x'\|_2$  for all  $x, x' \in \mathbb{R}^{d_1}$ .*

We now formally state our robustness result.

**Theorem 2.** *Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be a function whose gradient is bounded by  $\nabla f_{\max}$  and suppose  $\Sigma = \sigma^2 \mathbf{I}$ . Then  $SG_{\Sigma}^f$  and  $\text{C-LIME}_{\Sigma}^f$  are both  $L$ -Lipschitz with  $L = \nabla f_{\max} / (2\sigma)$ .*

*Proof sketch.* We first prove the Lipschitzness of SmoothGrad using the Pinkser and data processing inequalities (van Erven & Harremoës, 2014). Theorem 1 then implies that C-LIME is also Lipschitz. See the full version for details.  $\square$

Theorem 2 shows that both SmoothGrad and C-LIME become less robust (i.e., the Lipschitz constants grows) when explaining functions with larger magnitude of gradients, or when the variance parameter  $\sigma^2$  used in gradient computation or perturbations decreases. However, the Lipschitz constant is independent of the input dimension  $d$ .

## 4. Convergence Analysis

The results in Section 3 prove the equivalence of SmoothGrad and C-LIME and also robustness of these techniques in *expectation* which corresponds to *large sample limits* in practice. Any useful implementation of these techniques is based on finite number of gradient computations or sample perturbations. In this section, we derive sample complexity bounds to examine how fast the empirical estimates for the outputs of SmoothGrad and C-LIME at any given point will converge to their expected value. This extends the implications of the results in Section 3 to practical implementations of SmoothGrad and C-LIME.

We start by examining how fast the output of SmoothGrad will converge to its expectation.

**Proposition 1.** *Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be a function whose gradient is bounded by  $\nabla f_{\max}$ . Fix  $x \in X$ ,  $\epsilon > 0$  and  $\delta > 0$ . Let  $n \geq C(\nabla f_{\max}/\epsilon)^2 \ln(d/\delta)$  for some absolute constant  $C$ . Then with probability of at least  $1 - \delta$ , over a sample  $S$  of size  $n$  from  $\mathcal{N}(x, \Sigma)$ , for any  $\Sigma \in \mathbb{R}^d$ , we have that  $\|SG_{\Sigma}^f(x) - SG_n^f(x)\|_2 \leq \epsilon$ .*

We next examine how fast the output of C-LIME will converge to its expectation.

**Theorem 3.** *Let  $f : \mathbb{R}^d \rightarrow [-1, 1]$  be a function. Fix  $x \in X$ ,  $\epsilon > 0$  and  $\delta > 0$ . Let  $S$  denote a sample of size  $n$  from  $\mathcal{N}(x, \Sigma)$  for  $\Sigma = \sigma^2 \mathbf{I}$  where*

$$n \geq C \frac{d \ln\left(\frac{d}{\delta}\right)}{\min(\epsilon\sigma^2, \epsilon\sigma^3/\|x\|_2, \|x\|_2, 1/\sigma^2)^2},$$

for some absolute constant  $C$ . Then with probability of at least  $1 - \delta$ ,  $\|\text{C-LIME}_{\Sigma}^f(x) - \text{C-LIME}_n^f(x)\|_2 \leq \epsilon$ .

*Proof sketch.* First observe that we can write the output of C-LIME both in expectation and in finite sample using the closed-form solution of ordinary least square as follows

$$\left\| \text{C-LIME}_{\Sigma}^f(x) - \text{C-LIME}_n^f(x) \right\|_2 = \left\| \mathbb{E}[(aa^{\top})]^{-1} \mathbb{E}[af(a)] - \left(\frac{1}{n} \sum_{b \in S} bb^{\top}\right)^{-1} \left(\frac{1}{n} \sum_{b \in S} bf(b)\right) \right\|_2,$$

where the expectations are with respect to  $a \sim \mathcal{N}(x, \Sigma)$  and we use  $b$  to index a sampled data point in a sample  $S$  of size  $n$ . By algebraic manipulation and applying

Cauchy-Schwartz and triangle inequalities, the term above is bounded by

$$\begin{aligned} & \left\| \mathbb{E} [aa^\top]^{-1} \right\|_2 \left\| \mathbb{E} [af(a)] - \frac{1}{n} \sum_{b \in S} bf(b) \right\|_2 + \\ & \left\| \mathbb{E} [aa^\top]^{-1} - \frac{1}{n} \left( \sum_{b \in S} bb^\top \right)^{-1} \right\|_2 \left\| \frac{1}{n} \sum_{b \in S} bf(b) \right\|_2. \end{aligned}$$

Therefore, it suffices to bound each of the 4 terms of the above equation separately. We show that the first term is bounded by  $1/\sigma^2$  using Weyl’s inequality. We then show that, with high probability, the second term is bounded by  $\epsilon/(2\sigma^2)$  using Union bound, Sub-Gaussian and Chernoff concentration inequalities. By applying the Weyl’s Cauchy-Schwartz inequalities, Bernstein inequality in the sub-exponential case for matrices (Tropp, 2012) and covariance estimation techniques (Koltchinskii & Lounici, 2017), we show that, with high probability, the third term is bounded by  $\epsilon/(4\|x\|_2)$ . Finally, we show that the last term is, with high probability, bounded by  $2\|x\|_2$ , by using Union, Chernoff and Sub-Gaussian concentration bounds as well as Cauchy-Schwartz and triangle inequality. Multiplying the 4 bounds and applying a Union bound, we witness the theorem’s claim. See the full version for details.  $\square$

Fixing  $\sigma^2$  and  $x$ , the bound in Theorem 3 has the standard  $1/\epsilon^2$  dependency on the error parameter  $\epsilon$  and  $\ln(1/\delta)$  dependency on the probability of failure  $\delta$ . Fixing other parameters, the sample complexity increases as either  $\sigma^2$  or  $\|x\|_2$  approach 0 or grow larger and larger. In the large regime, the growth in the sample complexity is in line with the intuition that accurate estimates under higher variance scenarios require more samples. In the small regime, in our analysis, the bound on the norm of the inverse of the product matrices will grow with a rate that is proportional to  $\sigma^2$  or  $1/\|x\|_2$  causing the growth in the sample complexity. We empirically study this dependency in Section 6.

## 5. Additional Properties

In this section we study additional properties that are satisfied by both SmoothGrad and C-LIME. We defer all the omitted proofs of this section to the full version.

The first property that we study is linearity.

**Proposition 2** (Linearity). *Fix a covariance matrix  $\Sigma \in \mathbb{R}^d \times \mathbb{R}^d$ . For all  $f, g : \mathbb{R}^d \rightarrow \mathbb{R}$ ,  $d \in \mathbb{N}$  and  $\alpha, \beta \in \mathbb{R}$*

$$\begin{aligned} SG_\Sigma^{\alpha f + \beta g} &= \alpha SG_\Sigma^f + \beta SG_\Sigma^g, \text{ and,} \\ C-LIME_\Sigma^{\alpha f + \beta g} &= \alpha C-LIME_\Sigma^f + \beta C-LIME_\Sigma^g. \end{aligned}$$

Linearity implies that the explanation of a more complex function that can be written as a linear combination of two

simpler functions is simply the linear combination of the explanations of each of the simpler functions. This is useful e.g., in situations where computing explanations are computationally expensive and new explanations for linear compositions of functions can be simply derived by linear composition of the previously computed explanations.

The next property we study is proportionality.

**Proposition 3** (Proportionality). *Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be a linear function of the form  $f(x) = \theta^\top x + b$  for  $\theta \in \mathbb{R}^d$  and  $b \in \mathbb{R}$ . For any  $x \in \mathbb{R}^d$  and  $\Sigma \in \mathbb{R}^d \times \mathbb{R}^d$*

$$SG_\Sigma^f(x) = C-LIME_\Sigma^f(x) = k(x)\theta,$$

for some function  $k : \mathbb{R}^d \rightarrow \mathbb{R}$ .

Proportionality implies that when the underlying function is linear both SmoothGrad and LIME provide explanations that are proportional to the weights of the underlying function. Although explaining the weights of a linear function with another set of weights might appear unnecessary, proportionality can be interpreted as a sanity check for explainability methods. Garreau & von Luxburg (2020) prove a weaker version of proportionality for C-LIME, where the multiplier  $k$  might be different for each feature.

An immediate consequence of proportionality is that, in general, SmoothGrad and C-LIME do not provide sparse explanations (for e.g., when the underlying function  $f$  is linear and non-sparse). In practice, sparsity can be promoted by adding a regularizer (i.e., by setting  $\Omega(g)$  appropriately in our general setting). We study the regularized version of C-LIME in the full version.

## 6. Experiments

In this section we evaluate our theoretical findings empirically on synthetic and real world datasets. We analyze the equivalence and robustness of SmoothGrad and C-LIME with respect to the number of perturbations. Finally we assess the sensitivity of these results to varying the hyperparameters such as the variance  $\sigma^2$  in perturbations.

### 6.1. Experimental Setup

**Datasets:** We generate a synthetic dataset and use 2 real world classification datasets from the UCI Machine Learning Repository (Dua & Graff, 2017).

**1. Simulated.** We simulate a 1000 sample classification dataset with a 2 dimensional feature space. We fix  $y \in \{0, 1\}$  randomly for each instance and sample  $x \in \mathbb{R}^2$  from  $\mathcal{N}(\mu_y, I_2)$  where  $\mu_0 = [-1, -1]$  and  $\mu_1 = [1, 1]$ . This results in the class clusters illustrated in Figure 1.

**2. Bankruptcy.** This dataset comprises of bankruptcy prediction of Polish companies (Zikeba et al., 2016). The input

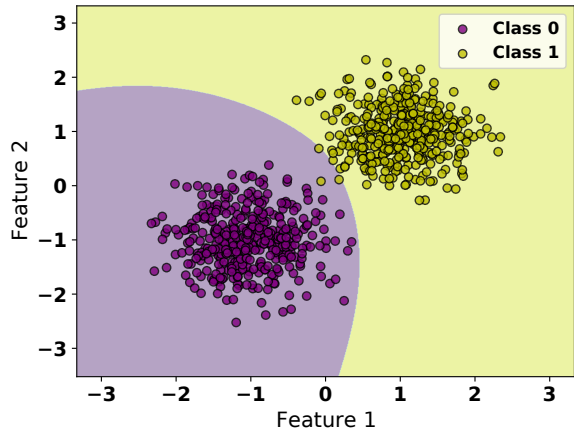


Figure 1. The decision boundary for the model trained on the simulated data. Each point corresponds to a point in our dataset. Purple and yellow denote the data points with labels 0 and 1, respectively.

attributes consist of features like net profit, sales and inventory from a pool of 10503 companies. We discard categorical features to align with our theory. As is standard practice when training neural networks, we normalize continuous features to  $\mathcal{N}(0, 1)$ . Given the resulting 15 dimensional feature set, the classification task is to predict whether the company in interest will bankrupt or not.

**3. Online Shopping.** This dataset comprises 12330 instances of online shopping interactions (Saka et al., 2019). Each sample contains 10 numerical features like the number of pages shoppers visited, time they spend on a page, metrics from Google Analytics and similar. Like with the Bankruptcy dataset, we discard categorical variables and normalize continuous variables, resulting in an 11 dimensional feature space. The target variable for classification is whether an online interaction ends in a purchase or not.

We choose the Bankruptcy and Online Shopping datasets since they contain a large number of real-valued features as assumed by our theory.

**Underlying Function:** For all our experiments, we use a two layer neural network with ELU activation function and 10 nodes per hidden layer. We follow the standard 80/20 dataset split, *i.e.*, 80% of the data was used for training the model while 20% was used for testing. These are the underlying models (functions) that we are explaining in our experiments. The models are trained using Adam optimizer using a cross-entropy loss function. Our best performing models achieve a testing accuracy of 99.50%, 96.30%, and 99.8% using 15, 60, and 100 training epochs for the Simulated, Bankruptcy, and Online Shopping datasets, respectively. We also train models using fewer than the aforementioned training epochs to assess the the impact of model accuracy on our equivalence and robustness guarantees.

**Parameters:** Consistent with our theory, for any input point  $x$ , for both C-LIME and SmoothGrad we generate perturbations from a local neighborhood of  $x$  by sampling points from  $\mathcal{N}(x, \sigma^2 \mathbf{I})$ . We study the effect of the number of perturbations and the value of  $\sigma^2$  in our experiments.

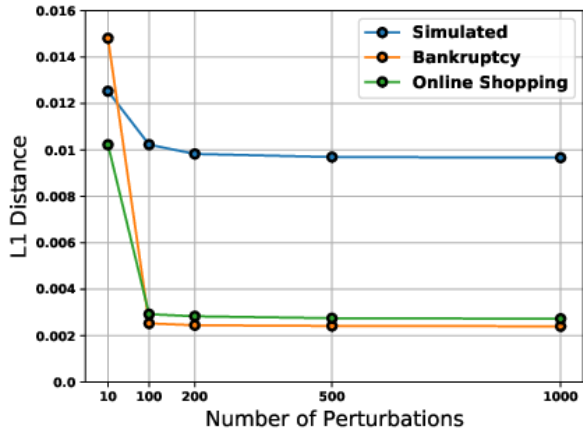


Figure 2. Equivalence plots showing that the  $L_1$  distance (Y axis) between the explanations of SmoothGrad and C-LIME decreases as we increase the number of perturbations (X axis). Each curve corresponds to a different dataset.

## 6.2. Equivalence

To evaluate the equivalence between SmoothGrad and C-LIME, we begin by generating explanations for each instance in the datasets' testing splits using  $\sigma^2 = 1$ . We probe the effect of varying  $\sigma^2$  in Section 6.4. We measure the distance between SmoothGrad and C-LIME explanations for each instance and then average these distances over the entire testing split. We repeat this process for different numbers of perturbations, plotting these average distances versus number of perturbations in Figure 2.

We observe that across all three datasets, the average  $L_1$  distance between the explanations for SmoothGrad and C-LIME decreases as we increase their respective number of perturbations, supporting equivalence. Interestingly, for all the three datasets the equivalence between the two explanation methods is achieved at as low as 100 perturbations. This is significantly lower than the finite perturbation estimates we derive in Proposition 1 and Theorem 3, suggesting that in practice, explanations approach their expected value even with a small number of perturbations.

## 6.3. Robustness

To evaluate the robustness of SmoothGrad and C-LIME, we first take each instance  $x$  in the testing splits and generate 10 nearby neighbors  $x' \sim \mathcal{N}(x, 0.01\mathbf{I})$ . We compute explanations for each original instance and its neighbors

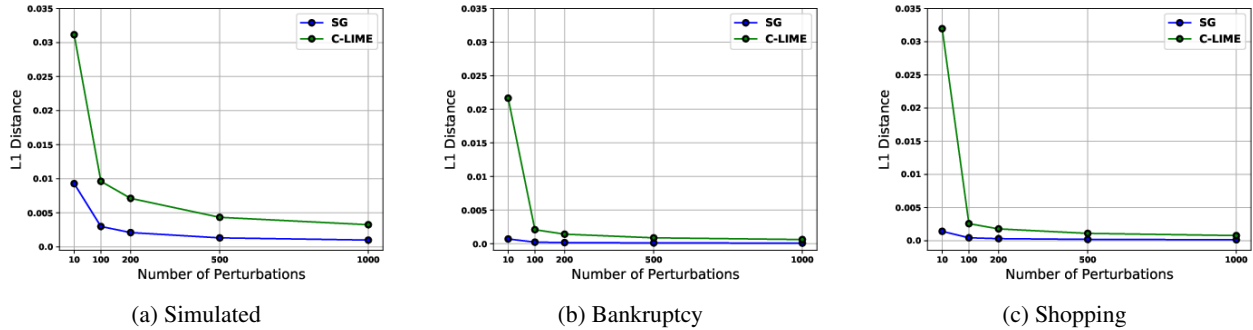


Figure 3. Robustness plots showing the maximum  $L1$  distance (Y axis) between the explanations for the original and neighboring inputs averaged over the test data points as a function of number of perturbations used for explanations (X axis). Each plot corresponds to a different dataset. In each plot there are two curves: one for SmoothGrad and another one for C-LIME.

by perturbations using  $\sigma^2 = 1$ . For each instance  $x$ , we compute the distance between the explanation for  $x$  and the explanations for each of its neighboring points  $x'$ . We take the maximum of these distances and then take the average of these maximum distances over the entire testing split. A small value for this average maximum distance suggests that explanations are robust as it implies that the difference between explanations for an instance and its nearby neighbors is small. We compute this average maximum distance for various numbers of perturbations and plot them in Figure 3.

The average maximum distance approaches zero across all three datasets, evidencing the robustness of both SmoothGrad and C-LIME. Notice that SmoothGrad appears to be more robust than C-LIME, with the average maximum distance saturating even closer to 0 than C-LIME. Furthermore, SmoothGrad saturates faster than C-LIME at perturbation numbers as small as 200. This suggests that SmoothGrad is more robust than C-LIME for fixed finite perturbations.

#### 6.4. Sensitivity Analysis

We evaluate the sensitivity of our findings to varying parameters  $\sigma^2$  and accuracy of the underlying function.

**Sensitivity to  $\sigma^2$ .** We begin by evaluating the impact of varying  $\sigma^2$  (variance on the perturbations) on our results. We choose to focus on the Bankruptcy dataset, generating the previously described equivalency and robustness plots for  $\sigma^2 = 0.01, 0.1, \text{ and } 1$ , as illustrated in Figure 4. For additional analysis for other dataset refer to the full version. Notice that SmoothGrad and C-LIME converge to equivalence faster for smaller  $\sigma^2$ . Similarly, both SmoothGrad and C-LIME appear to achieve robustness faster for smaller  $\sigma^2$ . Both of these observations are intuitive as  $\sigma^2$  controls the size of the local neighborhood used to generate perturbations. Our theory, on the other hand, predicts that the number of perturbations should increase as either  $\sigma^2$  approaches 0 or becomes very large. We suspect that this is due

to our style of analysis which requires worst-case bounds on quantities such as the inverse of sampled covariance matrix which hypothetically can grow as  $\sigma^2$  approaches 0.

**Sensitivity to Performance of Underlying Function.** Finally, we analyze whether the performance of the underlying model hinders the equivalence or robustness of SmoothGrad and C-LIME. We modulate model performance by reducing the number of training epochs. We train a model with 86% accuracy using 16 epochs and contrast it to our original model with 96% accuracy after 60 epochs. Again we choose to focus on the Bankruptcy dataset, generating the previously described equivalency and robustness plots for both models, as illustrated in Figure 5. Interestingly, the convergence rates slow down as the performance of the model becomes worse. See the full version for more details.

## 7. Related Work

Interpretability research can be categorized into learning inherently interpretable models, and constructing post hoc explanations. We provide an overview below.

**Inherently Interpretable Models.** Many approaches have been proposed to learn inherently interpretable models, for various tasks including classification and clustering. To this end, various classes of models such as decision trees, decision lists (Letham et al., 2015), decision sets (Lakkaraju et al., 2016), prototype based models (Bien & Tibshirani, 2009; Kim et al., 2014), and generalized additive models (Lou et al., 2012; Caruana et al., 2015) were proposed. However, complex models such as deep neural networks often achieve higher accuracy than simpler models (Ribeiro et al., 2016); thus, there has been a lot of interest in constructing post hoc explanations to understand their behavior.

**Post Hoc Explanations.** Several techniques have been proposed in recent literature to construct *post hoc* explana-

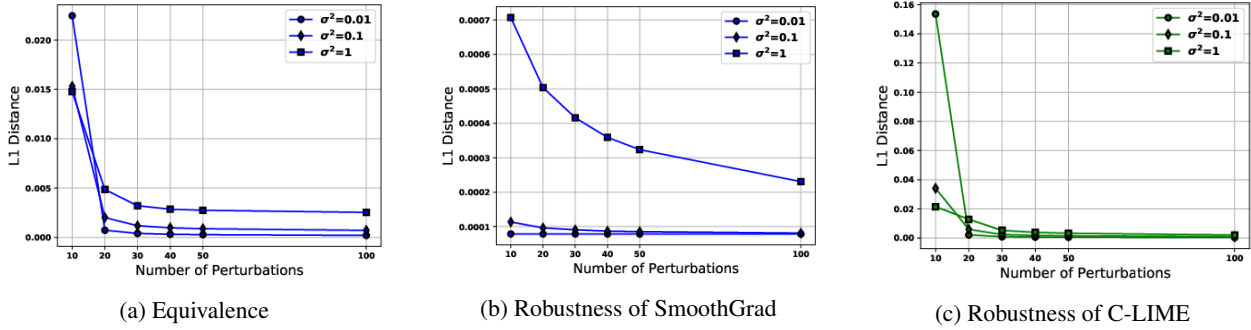


Figure 4. Equivalence (4a) and robustness plots for SmoothGrad (4b) and C-LIME (4c) for various  $\sigma^2$  on the Bankruptcy dataset. In each plot the Y axis corresponds to L1 distance and the X axis corresponds to the number of perturbations.

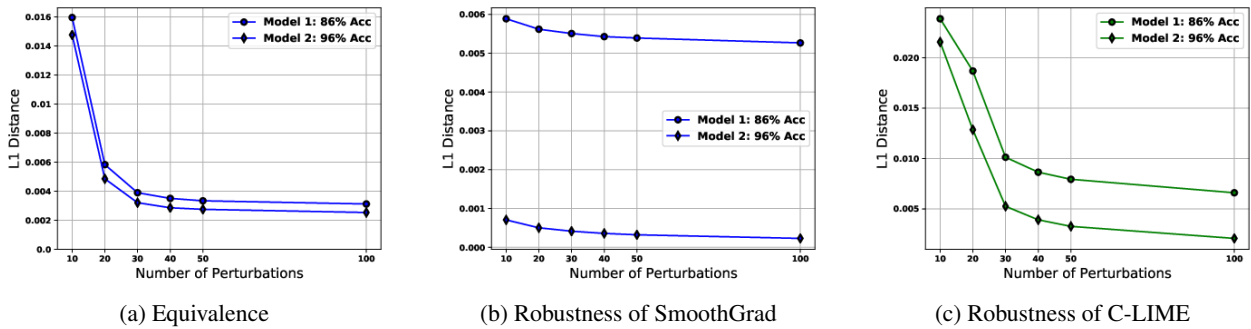


Figure 5. Equivalence (5a) and robustness plots for SmoothGrad (5b) and C-LIME (5c) for functions with various accuracy on the Bankruptcy dataset. In each plot the Y axis corresponds to L1 distance and the X axis corresponds to the number of perturbations.

tions of complex decision models. These techniques 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), and basic units of explanation (e.g., feature importance vs. rule based). In addition to LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017), there are several other *model-agnostic, local explanation* approaches that explain individual predictions of black box models such as Anchors, BayesLIME and BayesSHAP (Ribeiro et al., 2018; Slack et al., 2020b; Koh & Liang, 2017). Several of these approaches rely on input perturbations to learn interpretable local approximations.

Other local explanation methods including SmoothGrad have been proposed to compute *saliency maps* which capture local feature importance for an individual prediction by computing the gradient at that particular instance (Simonyan et al., 2014; Sundararajan et al., 2017; Selvaraju et al., 2017; Smilkov et al., 2017). There has also been recent work on constructing *counterfactual explanations* which capture what changes need to be made to a given instance in order to flip its prediction (Wachter et al., 2017; Ustun et al., 2019; Karimi et al., 2019; Poyiadzi et al., 2020; Looveren & Klaise, 2019; Barocas et al., 2020; Karimi et al., 2020a;b). Such explanations can be leveraged to provide recourse to

individuals negatively impacted by algorithmic decisions. An alternate approach is to construct *global explanations* for summarizing the complete behavior of any given black box by approximating it using interpretable models (Lakkaraju et al., 2019; Bastani et al., 2017; Kim et al., 2018).

**Analyzing Post Hoc Explanations.** Recent work has shed light on the downsides of post hoc explanation techniques. For instance, Rudin (2019) argued that post hoc explanations are not reliable, as these explanations are not necessarily faithful to the underlying models and present correlations. There has also been recent work on empirically exploring vulnerabilities of black box explanations (Adebayo et al., 2018; Slack et al., 2020a; Lakkaraju & Bastani, 2020; Rudin, 2019; Dombrowski et al., 2019)—e.g., Ghorbani et al. (2019) demonstrated that post hoc explanations may not be robust, changing drastically even with small perturbations to inputs (Alvarez-Melis & Jaakkola, 2018). In addition to the above works, there has also been some recent research that focuses on theoretically analyzing the robustness (Levine et al., 2019; Chalasani et al., 2020), and other properties (Garreau & von Luxburg, 2020) of some of the popular post hoc explanation techniques. However, these works do not attempt to explore deeper connections between different classes of these techniques.



---

## 8. Future Work

We initiate a study on the unification of perturbation and gradient based post hoc explanations, and pave the way for several promising research directions. It would be interesting to establish connections between other perturbation and gradient based explanations such as SHAP or Integrated gradients. It would also be interesting to study how perturbation and gradient based methods relate to counterfactual explanations. Furthermore, we mainly focused on the analysis of feature attribution methods. It would be exciting to analyze other kinds of explanation methods such as rule based or prototype based methods.

## Acknowledgement

We sincerely thank Amur Ghose for invaluable feedback about the proofs of Theorem 3 and Lemmas 3 and 6. We also thank Hadi Elzayn for helpful discussion about the proof of Theorem 3. Finally, we thank the anonymous ICML reviewers for their insightful feedback. This work is supported in part by the NSF award #IIS-2008461, NSF FAI award, Harvard’s Center for Research on Computation and Society, and Google. The views expressed are those of the authors and do not reflect the official policy or position of the funding agencies

## 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.
- Alvarez-Melis, D. and Jaakkola, T. On the robustness of interpretability methods. *CoRR*, abs/1806.08049, 2018.
- Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018.
- Barocas, S., Selbst, A., and Raghavan, M. The hidden assumptions behind counterfactual explanations and principal reasons. In *ACM Conference on Fairness, Accountability, and Transparency*, pp. 80–89, 2020.
- Bastani, O., Kim, C., and Bastani, H. Interpretability via model extraction. *CoRR*, abs/1706.09773, 2017.
- Bien, J. and Tibshirani, R. Classification by set cover: The prototype vector machine. *CoRR*, abs/0908.2284, 2009.
- 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 *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1721–1730, 2015.
- Chalasanani, P., Chen, J., Chowdhury, A. R., Wu, X., and Jha, S. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pp. 1383–1391, 2020.
- Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., and Kessel, P. Explanations can be manipulated and geometry is to blame. *CoRR*, abs/1906.07983, 2019.
- Doshi-Velez, F. and Kim, B. Towards a rigorous science of interpretable machine learning. *CoRR*, abs/1702.08608, 2017.
- Dua, D. and Graff, C. UCI machine learning repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- Garreau, D. and von Luxburg, U. Looking deeper into LIME. *CoRR*, abs/2008.11092, 2020.
- Ghorbani, A., Abid, A., and Zou, J. Interpretation of neural networks is fragile. In *AAAI Conference on Artificial Intelligence*, volume 33, pp. 3681–3688, 2019.
- Karimi, A.-H., Barthe, G., Balle, B., and Valera, I. Model-agnostic counterfactual explanations for consequential decisions, 2019.
- Karimi, A.-H., Schölkopf, B., and Valera, I. Algorithmic recourse: from counterfactual explanations to interventions. *CoRR*, abs/2002.06278, 2020a.
- Karimi, A.-H., von Kügelgen, J., Schölkopf, B., and Valera, I. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. *CoRR*, abs/2006.06831, 2020b.
- Kim, B., Rudin, C., and Shah, J. A. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Advances in Neural Information Processing Systems*, pp. 1952–1960, 2014.
- Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, 2018.
- Koh, P. and Liang, P. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894, 2017.
- Koltchinskii, V. and Lounici, K. Concentration inequalities and moment bounds for sample covariance operators. *Bernoulli*, 23(1):110–133, 02 2017.

- 
- Lakkaraju, H. and Bastani, O. “How do I fool you?”: Manipulating user trust via misleading black box explanations. In *AAAI Conference on Artificial Intelligence, Ethics, and Society*, pp. 79–85, 2020.
- 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*, 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*, pp. 131–138, 2019.
- Landsman, Z. and Nešlehová, J. Stein’s lemma for elliptical random vectors. *Journal of Multivariate Analysis*, 99(5): 912–927, 2008.
- Letham, B., Rudin, C., McCormick, T., and Madigan, D. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 2015.
- Levine, A., Singla, S., and Feizi, S. Certifiably robust interpretation in deep learning. *CoRR, abs/1905.12105*, 2019.
- Liu, J. Siegel’s formula via Stein’s identities. *Statistics & Probability Letters*, 21(3):247–251, 1994.
- Looveren, A. and Klaise, J. Interpretable counterfactual explanations guided by prototypes. *CoRR, abs/1907.02584*, 2019.
- Lou, Y., Caruana, R., and Gehrke, J. Intelligible models for classification and regression. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 150–158, 2012.
- Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.
- Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. FACE: Feasible and actionable counterfactual explanations. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020.
- Ribeiro, M., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *AAAI Conference on Artificial Intelligence*, pp. 1527–1535, 2018.
- Ribeiro, M. T., Singh, S., and Guestrin, C. “Why should I trust you?” Explaining the predictions of any classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.
- 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.
- Saka, O., Polat, O., Katircioglu, M., and Kastro, Y. Real-time prediction of online shoppers’ purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 31:6893–6908, 2019.
- Selvaraju, R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision*, pp. 618–626, 2017.
- Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representations*, 2014.
- 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. In *AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020a.
- Slack, D., Hilgard, S., Singh, S., and Lakkaraju, H. How much should I trust you? modeling uncertainty of black box explanations. *CoRR, abs/2008.05030*, 2020b.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: Removing noise by adding noise. *CoRR, abs/1706.03825*, 2017.
- Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328, 2017.
- Tropp, J. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.
- Ustun, B., Spangher, A., and Liu, Y. Actionable recourse in linear classification. In *ACM Conference on Fairness, Accountability, and Transparency*, pp. 10–19, 2019.
- van Erven, T. and Harremoës, P. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.
- Wachter, S., Mittelstadt, B., and Russell, C. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841, 2017.
- Zikeba, M., Tomczak, S., and Tomczak, J. Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. *Expert Systems with Applications*, 2016.