Robust Explanation for Free or At the Cost of Faithfulness

Zeren Tan [1] **Yang Tian** [2 3]

Abstract

Devoted to interpreting the explicit behaviors of machine learning models, explanation methods can identify implicit characteristics of models to improve trustworthiness. However, explanation methods are shown as vulnerable to adversarial perturbations, implying security concerns in high-stakes domains. In this paper, we investigate when robust explanations are necessary and what they cost. We prove that the robustness of explanations is determined by the robustness of the model to be explained. Therefore, we can have robust explanations for *free* for a robust model. To have robust explanations for a non-robust model, composing the original model with a kernel is proved as an effective way that returns strictly more robust explanations. Nevertheless, we argue that this also incurs a *robustness-faithfulness trade-off*, i.e., contrary to common expectations, an explanation method may also become less faithful when it becomes more robust. This argument holds for any model. We are the first to introduce this trade-off and theoretically prove its existence for SmoothGrad. Theoretical findings are verified by empirical evidence on six state-of-the-art explanation methods and four backbones.

1. Introduction

Through the years, many explanation methods have been designed to interpret the behaviors of black-box machine learning (ML) models. For instance, a commonly-used method is to explain model behaviors by attributing an importance score to each feature (Ribeiro et al., 2016; Lundberg & Lee, 2017; Baehrens et al., 2010; Simonyan et al., 2013; Montavon et al., 2017; Smilkov et al., 2017; Sundararajan et al., 2017). These importance scores can help users identify the most influential features for the model they use.

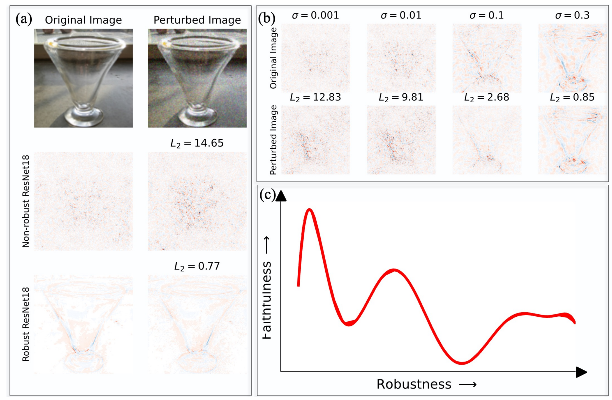


Figure 1. **Illustration of our results.** (a) Gradient explanations of robust and non-robust ResNet18 on an image $\mathbf{x}$ and a perturbed image $\mathbf{x} + \delta$ by adding Gaussian noise with magnitude $\delta \sim \mathcal{N}(\mathbf{0}, 0.1\mathbf{I})$. The numbers in the second column are L_2 distance between Gradient map on $\mathbf{x}$ and $\mathbf{x} + \delta$. Explanations of robust ResNet18 are more robust. (b) SmoothGrad explanations of non-robust ResNet18 with different noise level σ. Explanation becomes more robust when σ increases. (c) As the explanation becomes more robust, its faithfulness w.r.t. non-robust ResNet18 first increases and then decreases.

However, these explanation methods have been shown to be vulnerable to adversarial perturbations by recent works (Dombrowski et al., 2019; Ghorbani et al., 2019; Heo et al., 2019; Lakkaraju & Bastani, 2020; Le Merrer & Trédan, 2020; Slack et al., 2020; 2021). The fragility of explanation may mislead users make wrong decisions and thus cause security concerns in high-stakes domains such as finance, healthcare, and criminal justice (Ghorbani et al., 2019; Agarwal et al., 2021; Wang et al., 2020). For example, when a doctor prescribes medicines and diagnoses based on attribution maps on patients' chest imaging, it would cause misdiagnosis if explanations are different for two almost visually indistinguishable images.

Therefore, many efforts have been devoted to investigating robust attributions (Alvarez Melis & Jaakkola, 2018; Dombrowski et al., 2019; Yang et al., 2020; Rieger & Hansen,

[*]Equal contribution [1]Institute for Interdisciplinary Information Science, Tsinghua University, 100084, Beijing, China [2]Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, 100084, Beijing, China [3]Department of Psychology, Tsinghua University, 100084, Beijing, China. Correspondence to: Yang Tian <tiany20@mails.tsinghua.edu.cn & tyanyang04@gmail.com>, Zeren Tan <thutzr1019@gmail.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2020; Lakkaraju & Bastani, 2020; Wang et al., 2020; Chen et al., 2019; Boopathy et al., 2020; Anders et al., 2020; Lakkaraju et al., 2020; Upadhyay et al., 2021; Bykov et al., 2022). Many of them achieve this objective by retraining the model with an extra regularization term and show empirically that robust explanations can be obtained by training a robust model.

Nevertheless, retraining a model is time-consuming. Meanwhile, the retrained model may differ dramatically with the original model. These limitations raise three questions: *(1) Do we really need to pay extra efforts to make our explanations robust? (2) Can we achieve robust explanations without retraining? (3) Are robust explanations really better than their non-robust counterparts?*

In this paper, we theoretically show that robust models are guaranteed to have more robust explanations than their non-robust counterparts (see Theorem 4.1), which provides an answer to question (1). As shown in Figure 1 (a), attribution maps on two visually similar images are computed for both robust and non-robust ResNet18. The L_2 distance between two attribution maps for non-robust ResNet18 is much larger than that for robust ResNet18. The intuition behind this result is that a robust model behaves similarly on similar inputs and, therefore, explanations should also be similar. This result also sheds light on how adding regularization and retraining can achieve robust attributions as shown in (Wang et al., 2020; Chen et al., 2019; Boopathy et al., 2020). Specifically, regularization can lead to a locally more robust model, which in turn produces more robust explanations.

To attribute robustly without retraining, we propose to smooth the model by composing it with a kernel, i.e., for a classifier f, use $\hat{f} = \mathbb{E}_{\epsilon\sim\mu}[f(\mathbf{x}+\epsilon)]$ for attribution where μ is a probability distribution. With an appropriate μ, we prove that explanations are strictly more robust which shows a positive answer to question (2). SmoothGrad (Smilkov et al., 2017) just computes Gradient explanation by smoothing with Gaussian kernel. In Figure 1 (b), SmoothGrad explanations with different noise level σ are computed on non-robust ResNet18. As σ increases, explanations become more robust since the L_2 distance between explanations on two images becomes smaller. The derived theoretical result not only provides evidence for why SmoothGrad and UniGrad are effective but also has broader implications. Specifically, our finding suggests that other types of kernels may also be able to achieve similar results. This insight has important practical implications for developing more effective and efficient attribution methods.

Although robust attribution can be achieved in diverse ways, we argue that there exists a trade-off between robustness and faithfulness. Specifically, we prove that the faithfulness of SmoothGrad may decrease during the increase of σ in SmoothGrad, which shows a negative answer to question (3). As shown in Figure 1 (c), more robust explanations are not necessarily more faithful. SmoothGrad is Gradient explanation for a smoothed model $\hat{f}$, and the behaviors of $\hat{f}$ and f become increasingly different when σ becomes larger. Explanations for $\hat{f}$ should not be expected as faithful to f. Empirically, this trade-off also exists for explanations across different explanation methods (see Figure 6).

The most similar work to ours is (Yeh et al., 2019). Our definitions of robustness and faithfulness look similar to sensitivity and infidelity in (Yeh et al., 2019) but are in fact different. Yeh et al. (2019) prove under certain conditions, smoothing with kernel improves robustness and faithfulness at the same time while we show both theoretically and empirically that there exists a robustness-faithfulness trade-off. This trade-off also occurs in Figure 6 of (Yeh et al., 2019), but further analysis is not provided in (Yeh et al., 2019).

To summarize, our contributions are as follows:

- We prove that the robustness of an explanation is determined by the robustness of the ML model, i.e., *for a robust model, we can obtain robust explanations for free*. We validate this finding with experiments on six explanation methods and four backbones.
- Without a robust model, we argue that by composing the ML model with an appropriate kernel, an explanation method can become strictly more robust. SmoothGrad, as an example, is the Gradient explanation of a model composed with the Gaussian kernel.
- Although robust explanations can be achieved in many ways, we prove the existence of a robustness-faithfulness trade-off for SmoothGrad. When σ becomes larger, explanations become more robust while they become less faithful at the same time. Empirical evidence suggests that this trade-off also exists across methods with different robustness.

2. Related Work

Due to the space limitation, we primarily name some of the most related works in this section. Please refer to Appendix A for a more complete review.

Fragility of Explanation. A line of research investigates the fragility of explanation. Dombrowski et al. (2019); Ghorbani et al. (2019) find that explanation can be manipulated by adversarial perturbations. Slack et al. (2021) show similar result for counterfactual explanations. Besides adversarial perturbation, Heo et al. (2019); Lakkaraju & Bastani (2020) show how to find a model that preserves accuracy but has different explanations with the original model at the same time. Slack et al. (2020); Le Merrer & Trédan (2020) propose that explanations cannot be easily trusted as they present an attack that can generate explanations to hide the use of an arbitrary set of features by a classifier.

Robust Attribution. Another line of research aims to design robust and stable explanations. Rieger & Hansen (2020) propose that the explanation averaged across different methods is more robust than a single explanation method. For any classifier, Anders et al. (2020) prove that there exists another classifier which behaves the same on the data but has attributions arbitrarily close to target attributions. They demonstrate that projecting attributions to a low-dimensional submanifold helps improve robustness. Lakkaraju et al. (2020) formulate a minimax objective to find explanations robust to input distribution shift. Many works propose to add an extra regularization term and retrain the model (Wang et al., 2020; Chen et al., 2019; Boopathy et al., 2020). They theoretically and empirically demonstrate that these methods return robust attributions. However, as we have suggested, retraining a model is time-consuming and this model may differ dramatically with the original model.

Connection between Model Robustness and Interpretability. There also exist many works investigating the connection between model robustness and interpretability. Etmann et al. (2019) prove that an increase in robustness may induce an increase in the alignment between an input image and its respective saliency map for linear models. Ignatiev et al. (2019) relate adversarial example and explanations by hitting set duality and propose an algorithm to compute adversarial examples from explanations or *vice versa*. Chalasani et al. (2020) theoretically find that adversarial training using an l_∞-bounded adversary produces models with sparse attribution vectors. Meanwhile, the natural training that encourages stable explanations is equivalent to adversarial training for 1-layer networks. Agarwal et al. (2022b) show the first analysis on the behavior of various state-of-the-art GNN explanation methods with respect to faithfulness, stability, and fairness preservation.

3. Preliminary

In this section, we introduce the notations used in our paper and the six explanation methods that we focus on.

3.1. Notation

We denote $\mathcal{X}$ and $\mathcal{Y}$ as the input and output space of a model $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} \subset \mathbb{R}^d, \mathcal{Y} \subset \mathbb{R}$. For example, model f is a ResNet50 classifier that takes a CIFAR10 image as an input and generate the probability of the most likely class as its output. Mapping $\phi : \mathcal{X} \to \mathcal{E}$ is an explanation function that interprets the behaviors of f, where $\mathcal{E}$ is the explanation space. For example, given an image $\mathbf{x} \in \mathcal{X}$ in CIFAR10 and the associate model output $f(\mathbf{x})$, function $\phi(\mathbf{x}) = \nabla f$ outputs a scalar value for each pixel in $\mathbf{x}$. We denote $\odot$, $\|\cdot\|_2$, $\|\cdot\|_F$, and $\mathcal{O}$ as element-wise product, L_2 norm, Frobenius norm, and the big O notation, respectively.

Note that we assume $\mathcal{X}$ and f as bounded throughout this paper, i.e., $\exists \beta, R > 0, \forall \mathbf{x} \in \mathcal{X}, \|\mathbf{x}\|_2 \leq \beta, |f(\mathbf{x})| \leq R$.

3.2. Post-hoc Explanation

Post-hoc explanation aims at interpreting model behaviors without access to model details. One of its subclass, feature importance explanation, assigns each feature a score of the importance to model output, i.e., a feature with a higher score can influence the output more. This paper focuses on six widely-used feature importance explanation methods:

Gradient(Grad): It returns $\phi(\mathbf{x}) = \nabla f$ to measure the influence of each feature under infinitesimal perturbation (Baehrens et al., 2010; Simonyan et al., 2013).

Gradient×Input(GI): $\phi(\mathbf{x}) = \mathbf{x} \odot \nabla f$ for this method (Montavon et al., 2017). It masks input by its corresponding gradient.

SmoothGrad(SG): Vanilla gradient explanations are shown to be noisy. SmoothGrad proposes to smooth out noise by averaging gradients at local neighborhood (Smilkov et al., 2017). The feature importance is thus $\phi(\mathbf{x}) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2\mathbf{I})}[\nabla f(\mathbf{x} + \epsilon)]$. In case dependence on σ should be shown explicitly, we use $\phi_\sigma(\mathbf{x})$.

Integrated Gradient(IG): This method is designed to satisfy several axioms (Sundararajan et al., 2017). It computes the path integral from a baseline $\mathbf{x}_0$ to $\mathbf{x}$, $\phi(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_0) \odot \int_0^1 \nabla f(\mathbf{x}_0 + \alpha(\mathbf{x} - \mathbf{x}_0))\mathrm{d}\alpha$.

LIME: LIME obtains samples in the local neighborhood of $\mathbf{x}$ by adding perturbations and then approximates f locally by an interpretable model (Ribeiro et al., 2016). In specific, $\phi(\mathbf{x}) = \arg\min_{g \in \mathcal{G}} \sum_{\mathbf{z}} \pi_{\mathbf{x}}(\mathbf{z})(f(\mathbf{x} \odot \mathbf{z}) - g(\mathbf{z}))^2 + \Omega(g)$, where $\mathcal{G}$ is a class of interpretable models, $\pi_{\mathbf{x}}(\mathbf{z})$ weights perturbed samples and Ω measures the complexity of g. $\mathbf{z} \in \{0, 1\}^d$ in LIME is a interpretable representation(e.g., superpixels in an image) that represents the inclusion and exclusion of features. The perturbed samples are obtained by uniformly sampling elements in $\mathbf{z}$ and then setting features not present with a baseline value. In this paper, we consider $\mathcal{G}, \pi_{\mathbf{x}}(\mathbf{z}), \Omega(g)$ as the following:

$$\begin{aligned}\mathcal{G} =& \{g(\mathbf{z}) | g(\mathbf{z}) = \mathbf{w}^\top \mathbf{z}, \mathbf{w} \in \mathbb{R}^d\}, \\ \pi_{\mathbf{x}}(\mathbf{z}) =& \exp\left(- \frac{\|\mathbf{x} - \mathbf{x} \odot \mathbf{z}\|}{\sigma^2} \right), \Omega(g) = \lambda \|\mathbf{w}\|^2.\end{aligned}$$

SHAP: SHAP is an additive feature attribution method that unifies several explanation methods (Lundberg & Lee, 2017). The feature importance of the i^{th} feature provided by SHAP is $\phi_i(\mathbf{x}) = \sum_{S \subset [d] \setminus \{i\}} \frac{|S|!(d-|S|-1)!}{d!}[f(\mathbf{x}_{S \cup \{i\}}) - f(\mathbf{x}_S)]$ where $\hat{\mathbf{x}} = \mathbf{x}_S$ is such that $\hat{\mathbf{x}}_j = \mathbf{x}_j, \forall j \in S$ and $\hat{\mathbf{x}}_j$ equals to a reference value for $j \notin S$.

3.3. Definitions

Robustness. We define the robustness of ϕ as its local Lipschitz.

Definition 3.1 (Explanation Robustness (Alvarez-Melis & Jaakkola, 2018; Wang et al., 2020))**.** An explanation function ϕ is said to be (δ, L)-Lipschitz if we have

$$\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2 \leq L\|\mathbf{x} - \mathbf{x}'\|_2$$

for $\forall \mathbf{x}, \mathbf{x}'$ that satisfy $\|\mathbf{x} - \mathbf{x}'\|_2 \leq \delta$.

For the machine-learning model f itself, we introduce two robustness measures: L-Lipschitz and H-smoothness.

Definition 3.2 ((δ, L)-Lipschitz)**.** A model f is (δ, L)-Lipschitz if $\forall x, x'$, $\|\mathbf{x} - \mathbf{x}'\|_2 \leq \delta$, we have

$$\|f(\mathbf{x}) - f(\mathbf{x}')\| \leq L\|\mathbf{x} - \mathbf{x}'\|_2$$

Definition 3.3 ((δ, H)-Smoothness)**.** A model f is (δ, H)-smooth if $\forall \mathbf{x}, \mathbf{x}'$, $\|\mathbf{x} - \mathbf{x}'\|_2 \leq \delta$, we have

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \leq H\|\mathbf{x} - \mathbf{x}'\|_2$$

Faithfulness. We use the similarity between feature importance and the marginal contribution of each feature as the faithfulness measure.

Definition 3.4 (Faithfulness)**.** The faithfulness of an explanation method is defined as follows:

$$\mathcal{F}(\phi(\mathbf{x})) = \mathrm{Sim}(\phi(\mathbf{x}), p(\mathbf{x})), \tag{1}$$

$$p(\mathbf{x}) = [p_1(\mathbf{x}), \cdots, p_d(\mathbf{x})]^\top, p_i(\mathbf{x}) = \frac{f(\mathbf{x}) - f(\bar{\mathbf{x}}_{-i})}{x_i - r_i}$$

where $\mathrm{Sim}(\cdot, \cdot)$ is a similarity metric and $\bar{\mathbf{x}}_{-i}$ equals $\mathbf{x}$ in each dimension except setting the value of dimension i to a reference value r_i, i.e., $(\bar{\mathbf{x}}_{-i})_j = x_j, j \neq i, (\bar{\mathbf{x}}_{-i})_i = r_i$.

A similar faithfulness definition has been introduced in (Liu et al., 2021a) where they choose Sim to be Pearson correlation and $f(\bar{\mathbf{x}}_{-i})$ to be the expected output of removing feature i. Sim could be any similarity measure, for example, the reciprocal of L_2 distance between $\phi(\mathbf{x})$ and $p(\mathbf{x})$.

4. Robust Explanation for Free

In this section, we aim to answer question (1): Do we really need to pay extra effort to make our explanation robust? We theoretically show that a robust model has robust explanations, i.e., *a robust model requires no extra payment.*

4.1. Robust Explanation for Robust Model

An explanation method is designed to reveal the underlying reasoning process of the model. Explanations are expected to be similar if the underlying inferences of f for two similar inputs $\mathbf{x}, \mathbf{x}'$ are similar. On the other hand, if the underlying inferences of f for two similar inputs $\mathbf{x}, \mathbf{x}'$ are different, the explanations provided are expected to be different. Hence, it is intuitive to argue that explanations for robust models are more robust than those for non-robust models as robust models produce similar outputs on similar inputs.

Formally, we can prove that explanation robustness is determined by model robustness:

Theorem 4.1. *Let $f : \mathcal{X} \to \mathcal{Y}$ be a (δ, L)-Lipschitz function, then we have SmoothGrad, LIME, SHAP are $(\delta, \mathcal{L})$-Lipschitz with corresponding $\mathcal{L}$ as the following:*

$$\mathcal{L}_{SG} = \mathcal{O}(L/\sigma)$$
$$\mathcal{L}_{LIME} = \mathcal{O}(\frac{\sqrt{d}L}{\lambda} + \frac{\beta R(\lambda + d)\sqrt{d}}{\lambda^2\sigma^2}\exp(\frac{2\beta}{\sigma^2}))$$
$$\mathcal{L}_{SHAP} = \mathcal{O}(\sqrt{d}L)$$

If f is also H-smooth, then we have Gradient, Gradient$\times$Input and Integrated Gradient(IG) are $(\delta, \mathcal{L})$-Lipschitz with corresponding $\mathcal{L}$ as the following:

$$\mathcal{L}_{Grad} = \mathcal{O}(H)$$
$$\mathcal{L}_{GI} = \mathcal{O}(\beta H + L)$$
$$\mathcal{L}_{IG} = \mathcal{O}(\beta H + L)$$

Note that from Theorem 4.1, the robustness of LIME and SHAP depends on the input dimension while SmoothGrad is independent of the input dimension. For sufficiently large σ, the local Lipschitz of SmoothGrad explanation is smaller than that of the classifier itself. Since Gradient, Gradient$\times$Input, and Integrated Gradient rely on gradient information, we need to bound the gradient change in the neighborhood, which needs the smoothness condition. Therefore, the local Lipschitz of these three methods depends on H. The reason SmoothGrad does not rely on smoothness condition is that we can get rid of the gradient by Stein's Lemma (see (Lin et al., 2019)).

Connection with robust attribution by regularization. Many efforts have been dedicated to developing robust attribution methods by adding a regularization term to the loss function and retraining the model. For example, Dombrowski et al. (2022); Wang et al. (2020) regularize the Hessian. Theorem 4.1 explains the mechanisms underlying their success: the explanation can have a smaller local Lipschitz because the regularization makes the trained model have a small local Lipschitz or be locally more smooth.

4.2. Robust Explanation for Any Model: Smoothing

From the above discussion and Theorem 4.1, we see that if f is robust, i.e., L, H are small, explanations are robust with lower local Lipschitz. However, one may ask if it is possible to have a lower Lipschitz without adjusting the explanation method when f is not robust. We discuss this question in this section to provide an answer for question (2).

In randomized smoothing literature, it has been shown that $f_\sigma(\mathbf{x}) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2\mathbf{I})}[f(\mathbf{x}+\epsilon)]$ has certified robustness. Actually, we can prove the following Proposition.

Proposition 4.2. *Let $f : \mathcal{X} \to \mathcal{Y}$ be a bounded above (δ, L)-Lipschitz function, i.e., $\exists R > 0, |f(\mathbf{x})| \le R, \forall \mathbf{x} \in \mathcal{X}$. The smoothed function $f_\sigma(\mathbf{x}) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)}[f(\mathbf{x}+\epsilon)]$ is $(\delta, \mathcal{L})$-Lipschitz where $\mathcal{L} = \min\left(L, \frac{R}{2\sigma}\right)$. If $\sigma > R/(2L)$, then $\mathcal{L} < L$, i.e., f_σ has strictly smaller local Lipschitz than f.*

It can be implied from the above proposition that the smoothed version of f has lower Lipschitz (by choosing $\sigma > \beta L/2$) than the original version.

Since f_σ has lower Lipschitz, the explanations computed on f_σ also have provably lower Lipschitz than the explanations computed on the original f, implying that *we can have robust explanations by smoothing f*. This result is similar but not equivalent to the observation in (Wang et al., 2020) because their bound on σ for SmoothGrad and Gradient depends on δ while our bound does not have this dependence.

Although Gaussian kernel is widely adopted, we suggest that other appropriate kernels can achieve the same results.

Theorem 4.3. *Let $f : \mathcal{X} \to \mathcal{Y}$ be a bounded above (δ, L)-Lipschitz function. For any measure μ and variable $\mathbf{z} \sim \mu$, we denote $\mu_\mathbf{x}$ as the probability measure w.r.t. $\mathbf{z}+\mathbf{x}$. If the total variation distance between $\mu_\mathbf{x}$ and $\mu_{\mathbf{x}'}$ is bounded by*

$$d_{TV}(\mu_\mathbf{x}, \mu_{\mathbf{x}'}) \le \gamma \|\mathbf{x} - \mathbf{x}'\|_2$$

for any $\|\mathbf{x} - \mathbf{x}'\| \le \delta$, then $f_\mu(\mathbf{x}) = \mathbb{E}_{\mathbf{z}\sim\mu}[f(\mathbf{x}+\mathbf{z})]$ is $(\delta, \mathcal{L})$-Lipschitz with $\mathcal{L} = R\gamma$. If $\gamma < L/R$, then $\mathcal{L} < L$, i.e, f_μ has a lower local Lipschitz than f.

The above theorem states that smoothing f with a kernel can return a more robust function than f if this kernel is robust in the sense that the total variation between the original distribution and the distribution after shifting is small.

Connection with UniGrad and SmoothGrad. Uniform Gradient proposed by Wang et al. (2020) and SmoothGrad are just gradient with uniform kernel $\mu = \mathbb{I}[\mathbf{z} \in [-r, r]^d]$ and Gaussian kernel $\mu = \mathcal{N}(0, \sigma^2\mathbf{I})$, respectively. For SmoothGrad, we have $\gamma = \frac{1}{2\sigma}$, and thus $\mathcal{L} = R\gamma = \frac{R}{2\sigma}$ (see Appendix C.2 for details), which is consistent with Proposition 4.2. In Appendix C.2 we prove that $\gamma = \frac{\sqrt{d}}{r}$ for UniGrad. Therefore, Theorem 4.3 actually unifies and generalizes UniGrad and SmoothGrad. We can obtain a model f_Φ by composing f with any kernel Φ that satisfies Theorem 4.3. Then we can compute explanations for f on f_Φ. These explanations are provably more robust than explanations computed on f. UniGrad and SmoothGrad are Gradient explanations on f_Φ. However, any explanation method ϕ can also be applied on f_Φ. The explanations computed are more robust that those computed by applying ϕ on f.

Connection with β-smoothing. Dombrowski et al. (2019) prove that replacing ReLU non-linearity with softplus improves the robustness of gradient-based explanation methods. Their method is equivalent to SmoothGrad for a one-layer neural network and leads to visually similar maps for deep networks. Thus, their work bridges between softplus and Gaussian kernel. Theorem 4.3 suggests that the equivalence between other non-linearities and kernels could be further investigated. If any connection could be built, we could potentially replace ReLU with the non-linearity and robust explanations could be achieved by computing explanations on the modified network.

5. Robustness-Faithfulness Trade-off

In this section, we attempt to explain why the trade-off exists by illustrating it with a toy example and providing theoretical results for general cases.

5.1. Why Trade-off Exists

In Section 4, we analyze the relationship between model robustness and explanation robustness. We show that a robust model has robust explanations for free. To improve the robustness of a given explanation, we develop a smoothing technique that generalizes previous methods (Smilkov et al., 2017; Wang et al., 2020) and returns provably more robust explanations by choosing an appropriate smoothing kernel.

However, by applying explanation methods after adding a smoothing kernel to f, we are actually explaining another model which may be dramatically different from the original one. Therefore, new explanations do not necessarily explain f as desired. Smoothing may not only smooth out the noise contained in the explanation Smilkov et al. (2017) but also smooth out useful information. Therefore, the faithfulness of explanations may be hurt.

Another way to see why robustness does not necessarily imply faithfulness is to consider what explanations should be. Explanations are computed with the hope that they reflect the underlying reasoning process of a model. If the model to be explained is not robust, i.e., behave differently for similar inputs, a faithful explanation method is expected to output different explanations as the model is actually reasoning differently. On the contrary, if explanations are similar on two inputs while the model behaves differently on them, the explanations are not faithful. If the explanation method incurs extra instability, it is not faithful.

In summary, the intuition behind why trade-off exists is

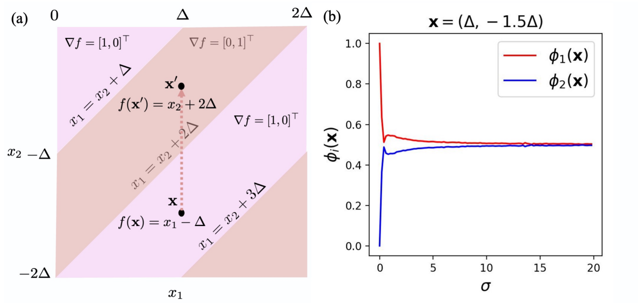


Figure 2. **A Toy Example.** (a) Gradient of f defined in Equation 2. Increasing x_2 by Δ also increase f (b) SmoothGrad attributions of x_1, x_2 for different σ. As σ increases, the attribution of each feature tends to be 0.5 which is not faithful for $\mathbf{x} = [\Delta, -1.5\Delta]^\top$

that while over-smoothing costs useful information to be smoothed out, under-smoothing incurs noise and cost information loss. Robustness of the most faithful explanation method should *match* the robustness of the model it explained. In the example shown in Section 5.2, SmoothGrad is over-smoothed when σ is very large and it outputs almost constant explanation which is uninformative.

With the above consideration, we analyze this robustness-faithfulness trade-off in this section. We will first illustrate the intuition why the robustness-faithfulness trade-off may occur through a toy example. And then we will show theoretically that when SmoothGrad explanations become more robust (σ becomes larger), the faithfulness of these explanations decreases at some point.

5.2. An Illustrative Example

Consider a function

$$f(x_1, x_2) = \begin{cases} x_1 - n\Delta & \text{if } \lfloor \frac{|x_1 - x_2|}{\Delta} \rfloor = 2n \\ x_2 + n\Delta & \text{if } \lfloor \frac{|x_1 - x_2|}{\Delta} \rfloor = 2n - 1 \end{cases} \tag{2}$$

where $\Delta > 0$ is a small constant. Then, it is easy to see that f is continuous and differentiable a.e. and

$$\nabla f(x_1, x_2) = \begin{cases} [1, 0]^\top & \text{if } \lfloor \frac{|x_1 - x_2|}{\Delta} \rfloor = 2n \\ [0, 1]^\top & \text{if } \lfloor \frac{|x_1 - x_2|}{\Delta} \rfloor = 2n - 1 \end{cases} \tag{3}$$

For $\mathbf{x} = (x_1, x_2), x_2 + 2\Delta \leq x_1 < x_2 + 3\Delta$, we have $f(\mathbf{x}) = x_1 - \Delta, \nabla f(\mathbf{x}) = [1, 0]^\top$ as shown in Figure 2. Explanation methods that only use information at $\mathbf{x}$(e.g., Gradient, Gradient×Input) would attribute 0 to x_2 indicating that x_2 has no contribution to $f(\mathbf{x})$ which is not faithful since increasing x_2 by Δ changes the value of f: $f(x_1, x_2 + \Delta) = x_2 + 2\Delta > x_1 - \Delta = f(\mathbf{x})$.

Gradient and Gradient×Input attribute nothing to x_2 because $\nabla f(\mathbf{x})$ aggregates information from an infinitesimal neighborhood of $\mathbf{x}$ while f remains constant w.r.t x_2 with infinitesimal perturbations. When information from a larger neighborhood is aggregated, attributions change as the perturbations on x_2 can change the value of f. However, if information from points far away is aggregated, attribution can be incorrect. For example, as noise level σ in SmoothGrad tends to infinity, the attribution output will tend to $[0.5, 0.5]^\top$ (see Figure 2 (b)), which indicates that x_1, x_2 are equally important but f depends more on x_1 at $\mathbf{x}$.

Therefore, we hypothesize that for an explanation method to be faithful, it should use information from a local neighborhood that is not very small, as the importance of a feature cannot be revealed in a very small neighborhood, nor very large, as much irrelevant information is included. Our hypothesis is validated by our theoretical analysis of SmoothGrad in Section 5.3 and empirical evidence in Section 6.3.

Remark 5.1. The hypothesis above recommends to add a relatively large perturbation to the original input when evaluating the faithfulness of explanations. The perturbation should be large enough for function f to have notably different outputs. For image data, pixel-wise perturbation is often too small, and grouping pixels into superpixels may be a more appropriate way to perturb the image on a larger scale. As for text data, if the perturbation is on the token level, the perturbation scale should be carefully chosen. Word-level perturbation may be a better way to perturb the input text on a larger scale. Finally, for tabular data, changing the value for categorical features may be enough. For continuous features, the perturbation can be chosen by splitting the value range into bins and selecting a value in a bin that is different from the bin where the original input value lies.

5.3. Theoretical Analysis

In this section, we analyze the relationship between the robustness and faithfulness of SmoothGrad. For SmoothGrad, we use ϕ_σ to denote its dependence on σ. We choose $\mathrm{Sim}(\mathbf{u}, \mathbf{v})$ to be a decreasing function w.r.t. $\|\mathbf{u} - \mathbf{v}\|_2$. We prove that ϕ_σ first increases and then decreases w.r.t. σ:

Theorem 5.2. *$f : \mathcal{X} \to \mathcal{Y}$ is a continuously differentiable function that is (δ, L)-Lipschitzbounded above, i.e., $\exists \beta > 0, |f(\mathbf{x})| \leq \beta, \forall \mathbf{x} \in \mathcal{X}$. If the following two conditions*

1. *$\exists \alpha > 0$, for any $0 < \sigma < \infty$, $\langle \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})}[\nabla f(\mathbf{x} + \epsilon)], p(\mathbf{x}) \rangle > \alpha \|\mathbb{E}_{\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I})}[\nabla f(\mathbf{x} + \epsilon)]\| \|p(\mathbf{x})\|$,*
2. *$\exists \tau, \nu > 0$ s.t. $\nu < \|p(\mathbf{x})\| / \|\nabla f(\mathbf{x})\| \leq \tau$ and $2\tau - 1 - \alpha^2 \nu^2 < 0$*

hold, there exists $0 < \sigma^* < \infty$ such that $\sigma^* = \arg\max_\sigma \mathcal{F}(\phi_\sigma(\mathbf{x}))$, i.e, the faithfulness of ϕ_σ first increases and then decreases as σ increases from 0 to $+\infty$.

The first assumption of the above theorem roughly states that the angle between the explanation returned by SmoothGrad and $p(\mathbf{x})$ is acute. This assumption is mild since $\phi_\sigma(\mathbf{x})$ and $p(\mathbf{x})$ are both close to $\nabla f(\mathbf{x})$. $\phi_\sigma(\mathbf{x})$ is close to $\nabla f(\mathbf{x})$ because ϵ is zero mean. In practice, on image data for example, one use Gaussian blur or average value of $\mathbf{x}$ as a reference so $\bar{\mathbf{x}}_{-i} \approx \mathbf{x}, \forall i$. Consequently, $p(\mathbf{x})$ is close to $\nabla f(\mathbf{x})$. This also corroborates the mildness of our second assumption, where the norm of $p(\mathbf{x})$ is bounded by $\nabla f(\mathbf{x})$.

This finding shows neither the most robust nor the most non-robust explanation is most faithful. In addition, it also shows given a non-robust model, we cannot expect robust explanations for free as it may lead to a loss of faithfulness. Although we only prove the robustness-faithfulness trade-off for SmoothGrad, we empirically observe this trade-off across different methods in Figure 6. Therefore, we suggest practitioners to be aware of this robustness-faithfulness trade-off and choose the best explanation method by examining their robustness and faithfulness.

6. Experiments

6.1. Experimental Setup

Datasets and Models. We perform our experiments on 1000 randomly selected images from CIFAR10. We use *robustness* (Engstrom et al., 2019) library to train both robust and non-robust versions of GoogLeNet (Szegedy et al., 2015), VGG16 (Simonyan & Zisserman, 2014), ResNet18, ResNet50 (He et al., 2016) and a tiny Swin Transformer, Swin-T (Liu et al., 2021b). See Appendix B for details.

Metrics. We evaluate the robustness and faithfulness of explanations from Gradient, Gradient×Input, SmoothGrad, Integrated Gradient, LIME, and SHAP. We compute the local Lipschitz to evaluate robustness. For each image $\mathbf{x}$, we compute its explanation and explanations for n images $\mathbf{x}_1, \cdots, \mathbf{x}_n$ sampled from $\mathcal{N}(\mathbf{x}, \sigma^2\mathbf{I})$. Then, we approximate the local Lipschitz by the following:

$$L(\mathbf{x}) = \max_i \frac{\|\phi(\mathbf{x}) - \phi(\mathbf{x}_i)\|}{\|\mathbf{x} - \mathbf{x}_i\|} \tag{4}$$

We average this value on 1000 images to obtain the final robustness measure. For faithfulness, we choose

$$\mathcal{F}(\phi(\mathbf{x})) = \frac{1}{\|\phi(\mathbf{x}) - p(\mathbf{x})\|}. \tag{5}$$

Note that the same results hold for any $\mathrm{Sim}(\phi(\mathbf{x}), p(\mathbf{x}))$ that is a decreasing function w.r.t. $\|\phi(\mathbf{x}) - p(\mathbf{x})\|$ and some other similarity measures (e.g., Pearson correlation and Spearman rank correlation, see Appendix B for more results).

6.2. Explanation for Robust Model is Robust

Explanations are more robust for more robust models. We compute the robustness of the explanations for six methods on eight models in Table 1. We list explanation robustness for each backbone model on its robust and non-robust versions. Rows are in descending order according to the robustness of corresponding robust models. From the results, we can draw several conclusions:

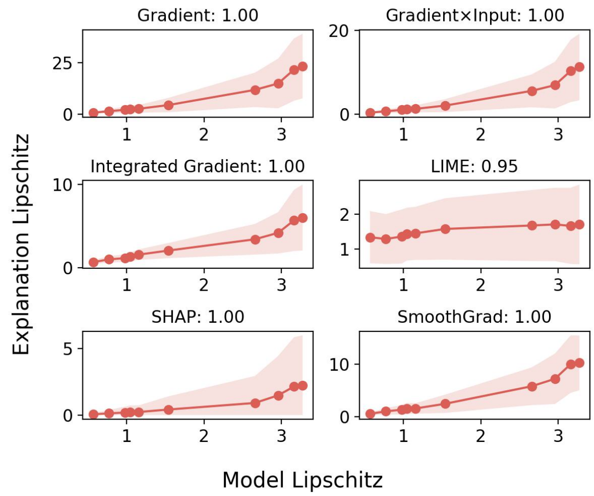


Figure 3. **The robustness of six explanation methods for ResNet18 with different robustness.** The line with a marker in the middle shows the mean local Lipschitz while the shadow is area within one standard deviation. The number on the top of each subfigure is the rank correlation between classifier robustness and corresponding explanation robustness.

- Explanations of robust models are more robust than their non-robust counterparts, which corroborates the theoretical results of Theorem 4.1.
- By composing a smoothing kernel with the original model, the robustness of its Gradient explanation increases. This confirms our arguments in Section 4.2 and Theorem 4.3.
- As shown by the last row of Table 1, a more robust model has more robust explanations. From VGG16 to GoogLeNet, the robustness of the classifier decreases while the robustness of explanations also decreases, which supports our results in Theorem 4.1.

To further explore how local Lipschitz of explanation changes w.r.t. local Lipschitz of the model, we use robust training to train 10 models with different robustness on the ResNet18 backbone. We show local Lipschitz of these models and local Lipschitz of their explanations in Figure 3. In the title of each subfigure, we show the rank correlation between them. It is clear that on each explanation method, the local Lipschitz of explanation is almost perfectly correlated with the local Lipschitz of the underlying model.

	Grad		GI		IG		SG		LIME		SHAP	
	NR	Rob.	NR	Rob.	NR	Rob.	NR	Rob.	NR	Rob.	NR	Rob.
Swin-T	6.556	0.084	3.163	0.044	1.140	0.107	4.590	0.087	1.558	0.566	0.240	0.012
VGG16	17.318	0.414	8.171	0.207	6.463	0.459	9.937	0.302	7.883	1.215	0.832	0.304
ResNet18	23.835	0.724	11.394	0.357	6.723	0.699	10.894	0.535	8.401	1.341	0.571	0.334
ResNet50	29.859	0.860	14.043	0.423	7.211	0.982	12.525	0.637	6.695	1.536	0.490	0.382
GoogLeNet	34.415	1.184	16.315	0.575	7.442	1.021	13.134	0.859	6.540	1.381	0.599	0.315
Rank Corr.	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.1	0.9	0.3	0.7

Table 1. **Lipschitz of six explanation methods on five different backbone models.** For each backbone, we show the results on both non-robust(NR) version and robust(Rob.) versions. The robustness of 5 robust models decreases from the top to the bottom row. The last row shows the Spearman rank correlation between classifier robustness and explanation robustness. It can be observed that for all six methods, explanation robustness correlates almost perfectly to classifier robustness.

Smoothing Improves Explanation Robustness of Any Model. Figure 4 shows how explanation robustness varies w.r.t σ in SmoothGrad. A larger σ leads to better explanation robustness across four different models. Given the same σ, models with smaller local Lipschitz generally have more robust explanations. This result validates the local Lipschitz bound for SmoothGrad in Theorem 4.1.

Additionally, by comparing these results with the local Lipschitz of Gradient explanation listed in Table 1 which corresponds to $\sigma = 0$ in SmoothGrad, it can be concluded that smoothing improves the explanation robustness of any model which is argued in Section 4.2 and Theorem 4.3.

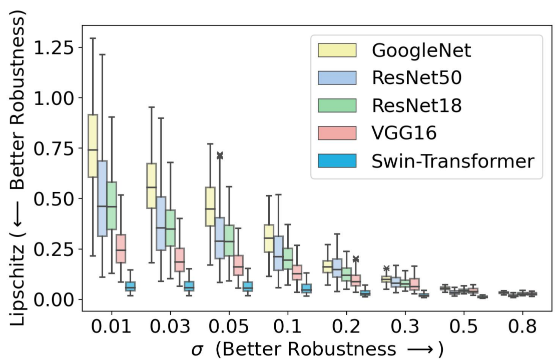


Figure 4. **Sensitivity of the robustness of SmoothGrad on smoothing noise σ.** The line shows the change of the mean of each box w.r.t σ. This shows that as σ increases, the robustness of SmoothGrad also increases. In addition, for the same σ, SmoothGrad is more robust for more robust classifiers.

6.3. Robustness-Faithfulness Trade-off

Trade-off on σ in SmoothGrad. For different σ, explanations and their robustness and faithfulness are computed on five models. Since the faithfulness of different models is measured on different scales, we normalize the faithfulness of each model by dividing it by its maximum value. Results are shown in Figure 5. With increasing σ, explanation robustness increases, as validated both theoretically and empirically in Theorem 4.1 and Section 6.2, while faithfulness decreases at some point. This phenomenon emerges for all five models. This shows that choosing a suitable σ may be tricky and a very large σ may not be a good choice as it may exchange faithfulness for robustness. If SmoothGrad is used to provide explanations, σ should be tuned according to the user's utility of robustness and faithfulness.

Trade-off on different methods. For six explanation methods, their local Lipschitz and faithfulness on five models are presented in Figure 6. The above observation leads us to hypothesize that such trade-offs may also exist across different methods. For each model, we normalize the local Lipschitz on six methods by the maximum of them so that the largest local Lipschitz is 1 for each model. At least two conclusions can be drawn: (1). On different methods, there also exists a robustness-faithfulness trade-off, and the trend is similar to that shown in Figure 5. (2). The order of explanation robustness and their faithfulness stays almost the same across the five models. For example, Gradient has the largest local Lipschitz, and LIME and SHAP are the most robust and unfaithful methods.

With the above evidence, the arguments in Section 5 are supported. Therefore, in practice, users should be aware of this trade-off and do not trust explanation systems that are claimed to be robust unconditionally. Users should choose a system and corresponding parameters under this trade-off in their applications.

7. Conclusion

By the intuition that a robust model should have robust explanations, we prove that the local Lipschitz of explanation is determined by the local Lipschitz of the model to be explained. By composing an appropriate smoothing kernel with a model, the local Lipschitz is proved to be reduced so

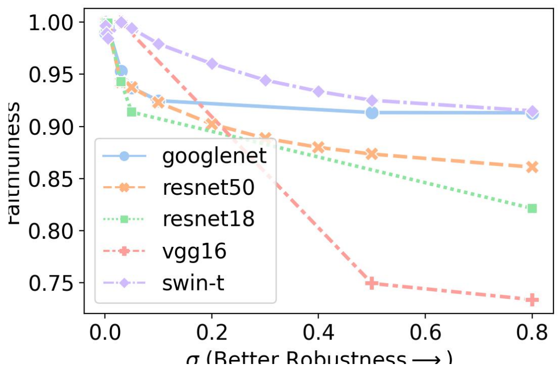


Figure 5. **Sensitivity of faithfulness of SmoothGrad w.r.t σ.** For all models, faithfulness of SmoothGrad first increases when σ is very small and then decreases when σ becomes large.

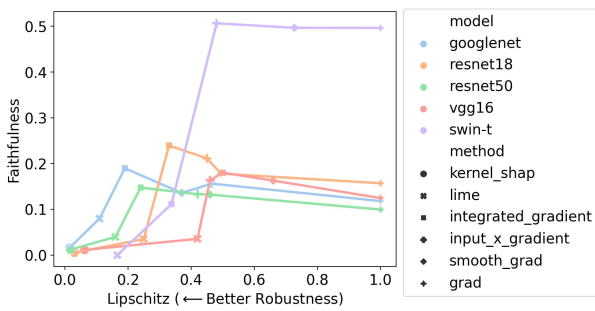


Figure 6. **Robustness-faithfulness trade-off on 6 explanation methods for 5 non-robust models trained on CIFAR10.** Each line represents a model while each marker stands for an explanation method. For each model, the Lipschitz constants for 6 methods are divided by the maximum of them. Therefore, the maximum Lipschitz shown for each model is 1. We can see that as explanations become robust starting from being non-robust, their faithfulness first increases and drops quickly afterward.

that its explanation becomes robust in the meantime. We show both theoretically and empirically that there exists a robustness-faithfulness trade-off. The idea is that the most faithful explanation is one "matches" the robustness of the underlying model. Thus, neither explanation that is extremely non-robust nor extremely robust is not faithful. We are the first to introduce the robustness-faithfulness trade-off. It would be interesting to explore this trade-off further both theoretically and empirically. For example, we only prove this trade-off for SmoothGrad and show results on different methods by experiments. It would be exciting to provide a more general theoretical justification.

References

Agarwal, C., Saxena, E., Krishna, S., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., and Lakkaraju, H. Openxai: Towards a transparent evaluation of model explanations. *arXiv preprint arXiv:2206.11104*, 2022a.

Agarwal, C., Zitnik, M., and Lakkaraju, H. Probing gnn explainers: A rigorous theoretical and empirical analysis of gnn explanation methods. In *International Conference on Artificial Intelligence and Statistics*, pp. 8969–8996. PMLR, 2022b.

Agarwal, S., Jabbari, S., Agarwal, C., Upadhyay, S., Wu, S., and Lakkaraju, H. Towards the unification and robustness of perturbation and gradient based explanations. In *International Conference on Machine Learning*, pp. 110–119. PMLR, 2021.

Alvarez Melis, D. and Jaakkola, T. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018.

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Anders, C., Pasliev, P., Dombrowski, A.-K., Müller, K.-R., and Kessel, P. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pp. 314–323. PMLR, 2020.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

Bansal, N., Agarwal, C., and Nguyen, A. Sam: The sensitivity of attribution methods to hyperparameters. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 8673–8683, 2020.

Bhatt, U., Weller, A., and Moura, J. M. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

Boopathy, A., Liu, S., Zhang, G., Liu, C., Chen, P.-Y., Chang, S., and Daniel, L. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning*, pp. 1014–1023. PMLR, 2020.

Bykov, K., Hedström, A., Nakajima, S., and Höhne, M. M.-C. Noisegrad—enhancing explanations by introducing stochasticity to model weights. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6132–6140, 2022.

Chalasani, 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. PMLR, 2020.

Chen, J., Wu, X., Rastogi, V., Liang, Y., and Jha, S. Robust attribution regularization. *Advances in Neural Information Processing Systems*, 32, 2019.

Dai, J., Upadhyay, S., Aivodji, U., Bach, S. H., and Lakkaraju, H. Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. *arXiv preprint arXiv:2205.07277*, 2022.

DeYoung, J., Jain, S., Rajani, N. F., Lehman, E., Xiong, C., Socher, R., and Wallace, B. C. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*, 2019.

Dombrowski, A.-K., Alber, M., Anders, C., Ackermann, M., Müller, K.-R., and Kessel, P. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32, 2019.

Dombrowski, A.-K., Anders, C. J., Müller, K.-R., and Kessel, P. Towards robust explanations for deep neural networks. *Pattern Recognition*, 121:108194, 2022.

Engstrom, L., Ilyas, A., Salman, H., Santurkar, S., and Tsipras, D. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

Etmann, C., Lunz, S., Maass, P., and Schönlieb, C.-B. On the connection between adversarial robustness and saliency map interpretability. *arXiv preprint arXiv:1905.04172*, 2019.

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.

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

Heo, J., Joo, S., and Moon, T. Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems*, 32, 2019.

Huai, M., Liu, J., Miao, C., Yao, L., and Zhang, A. Towards automating model explanations with certified robustness guarantees. 2022.

Ignatiev, A., Narodytska, N., and Marques-Silva, J. On relating explanations and adversarial examples. *Advances in neural information processing systems*, 32, 2019.

Lakkaraju, H. and Bastani, O. " how do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 79–85, 2020.

Lakkaraju, H., Arsov, N., and Bastani, O. Robust and stable black box explanations. In *International Conference on Machine Learning*, pp. 5628–5638. PMLR, 2020.

Le Merrer, E. and Trédan, G. Remote explainability faces the bouncer problem. *Nature Machine Intelligence*, 2(9): 529–539, 2020.

Levine, A., Singla, S., and Feizi, S. Certifiably robust interpretation in deep learning. *arXiv preprint arXiv:1905.12105*, 2019.

Lin, W., Khan, M. E., and Schmidt, M. Stein's lemma for the reparameterization trick with exponential family mixtures. *arXiv preprint arXiv:1910.13398*, 2019.

Liu, Y., Khandagale, S., White, C., and Neiswanger, W. Synthetic benchmarks for scientific research in explainable machine learning. *arXiv preprint arXiv:2106.12543*, 2021a.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

Paulavičius, R. and Žilinskas, J. Analysis of different norms and corresponding lipschitz constants for global optimization. *Technological and Economic Development of Economy*, 12(4):301–306, 2006.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Rieger, L. and Hansen, L. K. A simple defense against adversarial attacks on heatmap explanations. *arXiv preprint arXiv:2007.06381*, 2020.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S., and Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Slack, D., Hilgard, S., Jia, E., Singh, S., and Lakkaraju, H. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186, 2020.

Slack, D., Hilgard, A., Lakkaraju, H., and Singh, S. Counterfactual explanations can be manipulated. *Advances in Neural Information Processing Systems*, 34:62–75, 2021.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Upadhyay, S., Joshi, S., and Lakkaraju, H. Towards robust and reliable algorithmic recourse. *Advances in Neural Information Processing Systems*, 34:16926–16937, 2021.

Wang, Z., Wang, H., Ramkumar, S., Mardziel, P., Fredrikson, M., and Datta, A. Smoothed geometry for robust attribution. *Advances in Neural Information Processing Systems*, 33:13623–13634, 2020.

Yang, W., Lorch, L., Graule, M., Lakkaraju, H., and Doshi-Velez, F. Incorporating interpretable output constraints in bayesian neural networks. *Advances in Neural Information Processing Systems*, 33:12721–12731, 2020.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Yu, M., Chang, S., Zhang, Y., and Jaakkola, T. S. Rethinking cooperative rationalization: Introspective extraction and complement control. *arXiv preprint arXiv:1910.13294*, 2019.

A. Related Work

Fragility of Explanation. A line of research investigates the fragility of explanation. Dombrowski et al. (2019) find that explanations can be manipulated by adding unperceivable noise to inputs. They show theoretically that it is due to the large curvature of the underlying manifold and propose to replace ReLU non-linearity with SoftPlus to make explanations robust. Ghorbani et al. (2019) investigate adversarial perturbation to neural network interpretation and develop an algorithm to find target perturbations for two classes of interpretation methods. Heo et al. (2019) demonstrate that neural network explanation methods are easily fooled by a model fine-tuning step that alters the explanations while preserving accuracy. Lakkaraju & Bastani (2020) show the existence of a high-fidelity explanation that does not accurately reflect the biases in the black box model which may mislead users into trusting a problematic model. Le Merrer & Trédan (2020) propose that explanations cannot be easily trusted as they present an attack that can generate explanations to hide the use of an arbitrary set of features by a classifier. Slack et al. (2021) show that counterfactual explanations can be manipulated by adding small changes to the input so that the optimization algorithm can find a lower cost recourse.

Robust Attribution. Another line of research aims to design robust and stable explanations. Rieger & Hansen (2020) propose to average explanation from different methods and show that it is more robust than the single explanation method. Lakkaraju et al. (2020) formulate a minimax objective to find explanations robust to input distribution shift. Wang et al. (2020) propose Smooth Surface Regularization(SSR) that regularizes the maximum eigenvalue of the Hessian matrix of the original loss and propose UniGrad that is similar to SmoothGrad but with the uniform kernel. They show that both SSR and UniGrad are able to output robust explanations. Chen et al. (2019) add the attribution change computed by IntegratedGradient as a regularization term so that attribution is robust locally. Boopathy et al. (2020) prove that interpretation discrepancy is lower bounded by classification margin and propose interpretability-aware robust training which adds the maximum interpretation discrepancy in a δ-neighborhood as a regularization term. Anders et al. (2020) prove that for any classifier there exists another classifier that behaves the same on the data while having attribution arbitrarily close to target attribution and demonstrate that projecting attribution to low-dimensional submanifold helps improve robustness.

Evaluation of Faithfulness and Robustness. *(1) Robustness.* Alvarez-Melis & Jaakkola (2018) define local Lipschitz as a measure of explanation robustness and calculate the robustness of several widely-used local explanation methods. Yeh et al. (2019) propose to evaluate explanations with infidelity and sensitivity which is defined in their paper and proved that smoothing attribution can reduce infidelity and sensitivity so that explanation becomes more faithful and robust. Dai et al. (2022) use the expected L_1 distance between explanations of original input and perturbed input with Gaussian noise to measure the stability of explanations. Levine et al. (2019); Huai et al. (2022) use top-K overlap to measure explanation robustness. *(2) Faithfulness.* Faithfulness measures how accurately an explanation method reflects the true reasoning process of the model. Samek et al. (2016) evaluate heatmap by iteratively removing the most important features and use the area over the perturbation curve as the final metric. Yu et al. (2019); DeYoung et al. (2019) propose two faithfulness metrics: comprehensiveness and sufficiency which measure the degree by which the model is influenced by the removal and inclusion of the highest-ranked features, respectively. Bhatt et al. (2020) measure faithfulness of an explanation by subsampling feature subsets and calculate the correlation between total attribution scores of features in the subset and prediction change after setting features in the subset to a reference value. Dai et al. (2022); Agarwal et al. (2022a) use Prediction Gap Fidelity which computes expected prediction change while adding random noise to unimportant features recognized by attribution. Liu et al. (2021a) define faithfulness as the Pearson correlation coefficient between the feature importance and the approximate marginal contribution for each feature.

Connection between Model Robustness and Interpretability. There are also many works investigating the connection between model robustness and interpretability. Etmann et al. (2019) observe that robust network gives more clearer indication of what the classifier deems to be discriminative features. They prove it for linear model that an increase in robustness may induce an increase in the alignment between an input image and its respective saliency map. Ignatiev et al. (2019) relate adversarial example and explanations by hitting set duality and propose an algorithm that computes adversarial examples from explanations and vice-versa. Chalasani et al. (2020) theoretically find that adversarial training using an l_∞-bounded adversary produces models with sparse attribution vectors, while natural training that encouraging stable explanations is equivalent to adversarial training for 1-layer networks. Agarwal et al. (2022b) show the first analysis on the behavior of various state-of-the-art GNN explanation methods with respect to faithfulness, stability and fairness preservation.

B. Details on Experiments

B.1. Training Models

We use *robustness* library to train all of our models. The parameters we specify are listed in Table 2. The parameters used for training non-robust models are the same as those for training robust models except that we use adversarial training to train robust models. We adopt the implementation of Swin Transformer in *vision-transformers-cifar10*[1]. For Swin-T, we use

Model	eps	constraint	total_epochs	attack_lr
Swin-T	0.031372	inf	400	0.00784313
ResNet18	0.031372	inf	150	0.00784313
ResNet50	0.031372	inf	200	0.00784313
Vgg16	0.031372	inf	200	0.00784313
GoogLeNet	0.031372	inf	200	0.00784313

Table 2. Training Parameters.

Adam as the training optimizer while SGD is used for other models. We choose the patch size to be 4 because the height and width of images in CIFAR10 is 32×32.

B.2. Implementation of Explanation Methods

For Gradient, Gradient$\times$Input, Integrated Gradient, SmoothGrad, we adopt code from (Bansal et al., 2020)[2] while for LIME and SHAP, we use the implementation from *captum*[3].

B.3. Robustness Computation

We select 1000 image from CIFAR10, for each image $\mathbf{x}$ we randomly sample 50 points from $\mathcal{N}(\mathbf{x}, \sigma^2\mathbf{I})$ with $\sigma = 0.03$. For each sample $\mathbf{x}'$, we compute $\|\phi_{\mathbf{x}'} - \phi_{\mathbf{x}}\|_2 / \|\mathbf{x} - \mathbf{x}'\|$. Then, we take the maximum of these 50 values as the local Lipschitz of $\mathbf{x}$. The parameters we use for each explanation method are as follows:

Integrated Gradient: We use zero baseline and 10 intermediate points to compute the integral.

SmoothGrad: We use $\sigma = 0.03$ as the default value. The number of samples n is determined by σ. For $\sigma < 0.01$, $n = 10$. For $\sigma \geq 0.01, n = \lfloor \sigma/0.01 \rfloor \cdot 10$.

LIME: Quickshift segmentation algorithm is used to segment images to superpixels. kennel_size is set to 1, max_dist is set to 200, and ratio is set to 0.1. We choose num_samples as 100 and α in Ridge regressor as 1.

SHAP: We choose num_samples to be 100.

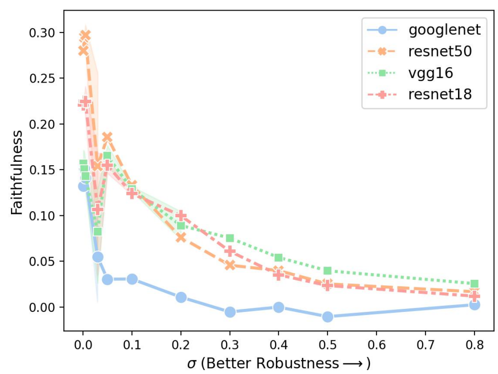
The random seed is fixed in our experiments. For images, we fix the random seed to its index before computing its local Lipschitz.

B.4. Faithfulness Computation

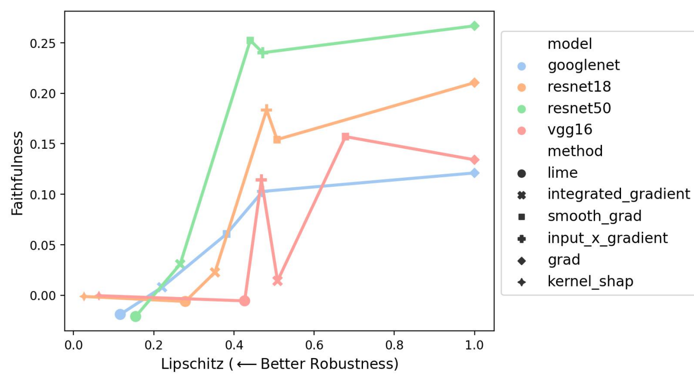
We split pixels into $\sim$100 groups and regard pixels in a group as a feature. Then we use Equation 1 to compute faithfulness for $\mathbf{x}$. The reference value is zero. The final faithfulness of an explanation method is the average of 1000 values computed.

B.5. More Experiment Results

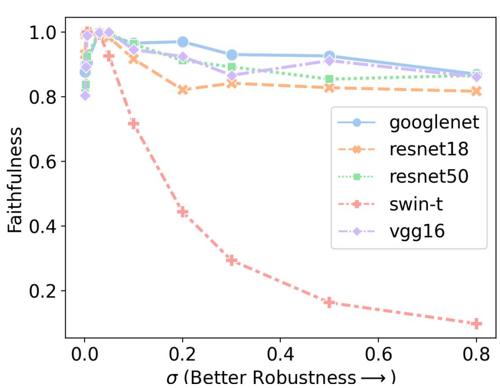
[1]https://github.com/kentaroy47/vision-transformers-cifar10
[2]https://github.com/anguyen8/sam
[3]https://github.com/pytorch/captum



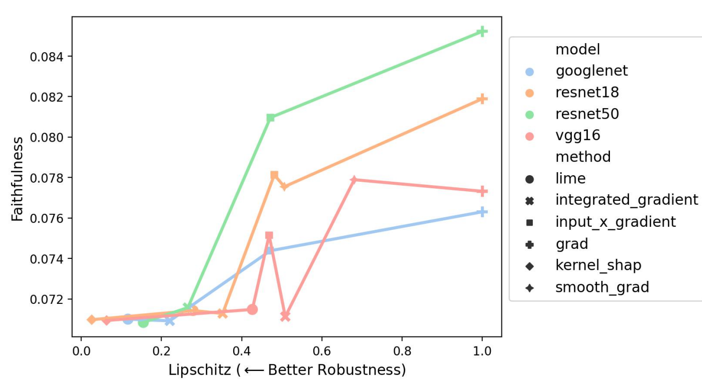
(a) Robustness-Faithfulness Tradeoff on σ (Pearson correlation).



(b) Robustness-Faithfulness Tradeoff on different methods (Pearson correlation).



(a) Robustness-Faithfulness Tradeoff on σ (Spearman rank correlation).



(b) Robustness-Faithfulness Tradeoff on different methods (Spearman rank correlation).

C. Proofs

C.1. Proof of Theorem 4.2

Proposition C.1. *Let $f : \mathcal{X} \to \mathcal{Y}$ be a (δ, L)-Lipschitz function that is bounded above, i.e., $\exists \beta > 0, |f(\mathbf{x})| \leq R, \forall \mathbf{x} \in \mathcal{X}$. Then the smoothed function $f_\sigma(\mathbf{x}) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)}[f(\mathbf{x}+\epsilon)$ is $(\delta, \mathcal{L})$-Lipschitz where $\mathcal{L} = \min\left(L, \frac{R}{2\sigma}\right)$. If $\sigma > R/(2L)$, then $\mathcal{L} < L$, i.e., f_σ has strictly smaller local Lipschitz than f.*

Proof. We only need to show that for $\mathbf{x}, \mathbf{x}'$, $\|\mathbf{x} - \mathbf{x}'\| \leq \delta$, we have

$$|f_\sigma(\mathbf{x}) - f_\sigma(\mathbf{x}')| \leq L\|\mathbf{x} - \mathbf{x}'\|$$

and

$$|f_\sigma(\mathbf{x}) - f_\sigma(\mathbf{x}')| \leq \frac{R}{2\sigma}\|\mathbf{x} - \mathbf{x}'\|$$

Since $f_\sigma(\mathbf{x}) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)}[f(\mathbf{x}+\epsilon)$, we have

$$\begin{aligned}
|f_\sigma(\mathbf{x}) - f_\sigma(\mathbf{x}')| =& |\mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)}[f(\mathbf{x}+\epsilon) - f(\mathbf{x}'+\epsilon)]| \\
\leq& \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)}[|f(\mathbf{x}+\epsilon) - f(\mathbf{x}'+\epsilon)|] \\
\leq& \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)}[L\|\mathbf{x} - \mathbf{x}'\|] \\
=& L\|\mathbf{x} - \mathbf{x}'\|
\end{aligned}$$

We can also bound the difference by bounding the difference between two distributions.

$$\begin{aligned}
|f_\sigma(\mathbf{x}) - f_\sigma(\mathbf{x}')| =& |\mathbb{E}_{\mu \sim \mathcal{N}(\mathbf{x},\sigma^2 I)}[f(\mu)] - \mathbb{E}_{\nu \sim \mathcal{N}(\mathbf{x}',\sigma^2 I)}[f(\nu)]| \\
=& |\int_{R^d} f(\mathbf{z})P_\mu(\mathbf{z}) - P_\nu(\mathbf{z})\mathrm{d}\mathbf{z}| \\
\leq& \int_{R^d} |f(\mathbf{z})||P_\mu(\mathbf{z}) - P_\nu(\mathbf{z})|\mathrm{d}\mathbf{z} \\
\leq& R\int_{R^d} |P_\mu(\mathbf{z}) - P_\nu(\mathbf{z})|\mathrm{d}\mathbf{z} \\
=& R d_{TV}(P_\mu, P_\nu)
\end{aligned}$$

where P_μ is the induced probability measure of $\mu \sim \mathcal{N}(\mathbf{x}, \sigma^2 I)$ and $d_{TV}(P_\mu, P_\nu)$ is the total variation distacne between probability measure P_μ, P_ν.

By Pinsker's inequality,

$$d_{TV}(P_\mu, P_\nu) \leq \sqrt{KL(P_\mu || P_\nu)/2} = \frac{\|\mathbf{x} - \mathbf{x}'\|}{2\sigma}.$$

Therefore, we have

$$|f_\sigma(\mathbf{x}) - f_\sigma(\mathbf{x}')| \leq \frac{R}{2\sigma}\|\mathbf{x} - \mathbf{x}'\|.$$

Combining two bounds together, we have

$$|f_\sigma(\mathbf{x}) - f_\sigma(\mathbf{x}')| \leq \mathcal{L}\|\mathbf{x} - \mathbf{x}'\|.$$

where $\mathcal{L} = \min\left(L, \frac{R}{2\sigma}\right)$. And it follows directly that, when $\sigma > R/(2L)$, $\mathcal{L} < L$, i.e., f_σ has strictly smaller local Lipschitz than f. □

C.2. Proof of Theorem 4.3

Theorem C.2. *Let $f : \mathcal{X} \to \mathcal{Y}$ be a (δ, L)-Lipschitz function that is bounded above, i.e., $\exists R > 0, |f(\mathbf{x})| \leq R, \forall \mathbf{x} \in \mathcal{X}$. For any probability measure μ and random variable $z \sim \mu$, denote $\mu_{\mathbf{x}}$ as the probability measure w.r.t. $z + \mathbf{x}$. If for $\|\mathbf{x} - \mathbf{x}'\| \leq \delta$, the total variation distance bewtween $\mu_{\mathbf{x}}, \mu_{\mathbf{x}'}$ is bounded by*

$$d_{TV}(\mu_{\mathbf{x}}, \mu_{\mathbf{x}'}) \leq \gamma\|\mathbf{x} - \mathbf{x}'\|_2,$$

then $f_\mu(\mathbf{x}) = \mathbb{E}_{z \sim \mu}[f(\mathbf{x} + z)]$ is $(\delta, \mathcal{L})$-Lipschitz with $\mathcal{L} = R\gamma$. If $\gamma < L/R$, then $\mathcal{L} < L$, that is f_μ has a lower local Lipschitz than f.

Proof. For $f_\mu(\mathbf{x}) = \mathbb{E}_{z\sim\mu}[f(\mathbf{x}+z)]$ and $\mathbf{x}, \mathbf{x}', \|\mathbf{x}-\mathbf{x}'\| \leq \delta$,

$$\begin{aligned}|f_\mu(\mathbf{x}) - f_\mu(\mathbf{x}')| =& |\mathbb{E}_{\mathbf{z}\sim\mu_{\mathbf{x}}}[f(\mathbf{z})] - \mathbb{E}_{\mathbf{z}\sim\mu_{\mathbf{x}'}}[f(\mathbf{z})]| \\ =& |\int_{\mathbb{R}^d} f(\mathbf{z})(P_{\mu_{\mathbf{x}}}(\mathbf{z}) - P_{\mu_{\mathbf{x}'}}(\mathbf{z}))\mathrm{d}\mathbf{z}| \\ \leq& \int_{\mathbb{R}^d} |f(\mathbf{z})||P_{\mu_{\mathbf{x}}}(\mathbf{z}) - P_{\mu_{\mathbf{x}'}}(\mathbf{z})|\mathrm{d}\mathbf{z} \\ \leq& R d_{TV}(\mu_{\mathbf{x}}, \mu_{\mathbf{x}'})\end{aligned}$$

Thus, if $d_{TV}(\mu_{\mathbf{x}}, \mu_{\mathbf{x}'}) \leq \gamma\|\mathbf{x}-\mathbf{x}'\|_2$, we have

$$|f_\mu(\mathbf{x}) - f_\mu(\mathbf{x}')| \leq \mathcal{L}\|\mathbf{x}-\mathbf{x}'\|_2, \mathcal{L} = R\gamma$$

If $\gamma < L/R$, then $\mathcal{L} < L$, that is f_μ has a lower local Lipschitz than f.

For UniGrad, $\mu_{\mathbf{x}} = U(\mathbf{x} + [-r, r]^d)$ which is uniform distribution centered at $\mathbf{x}$ with radius r. we can prove that

$$d_{TV}(\mu_{\mathbf{x}}, \mu_{\mathbf{x}'}) \leq \frac{\sqrt{d}}{r}\|\mathbf{x}-\mathbf{x}'\|.$$

Therefore, UniGrad is $(\delta, R\sqrt{d}/r)$-Lipschitz. We prove the above inequality holds in the following.

Denote $\rho = \mathbf{x}-\mathbf{x}' = (\rho_1, \cdots, \rho_d)$, $\sqrt{\sum_i \rho_i^2} = \|\rho\|$. The total variation distance between $\mu_{\mathbf{x}}, \mu_{\mathbf{x}'}$ is then the volume of two hypercubes $\mathbf{x} + [-r, r]^d$ and $\mathbf{x}' + [-r, r]^d$ minus the volume of their intersection divided by $(2r)^d$, i.e.,

$$d_{TV}(\mu_{\mathbf{x}}, \mu_{\mathbf{x}'}) = \frac{1}{(2r)^d}\left(\mathrm{Vol}(\mathbf{x}+[-r,r]^d) + \mathrm{Vol}(\mathbf{x}'+[-r,r]^d) - 2\mathrm{Vol}((\mathbf{x}+[-r,r]^d)\cap(\mathbf{x}'+[-r,r]^d))\right)$$

It is easy to see that $\mathrm{Vol}(\mathbf{x}+[-r,r]^d) = \mathrm{Vol}(\mathbf{x}'+[-r,r]^d) = (2r)^d$. The volume of intersection is $\prod_i(2r-\rho_i)$. Therefore, $d_{TV}(\mu_{\mathbf{x}}, \mu_{\mathbf{x}'})$ equals to

$$2 - 2\prod(1-\frac{\rho_i}{2r}) \leq 2 - 2(1-\sum_i\frac{\rho_i}{2r}) = 2\frac{\sum_i\rho_i}{2r} \leq 2\frac{\sqrt{d}}{2r}\|\rho\| = \frac{\sqrt{d}}{r}\|\mathbf{x}-\mathbf{x}'\|$$

The first inequality follows from Weierstrass inequality and the second one follows from Cauchy-Schwarz inequality. □

C.3. Proof of Theorem 5.2

For notation simplicity, we denote $f_\sigma : \mathbf{x} \mapsto \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[f(\mathbf{x}+\epsilon)]$ and ϕ_σ as explanations on f_σ.

Lemma C.3. *For $f : \mathcal{X} \to \mathcal{Y}$ is C^1, i.e., its gradient is continuous. Suppose f is bounded above, i.e., $\exists R > 0, |f(\mathbf{x})| \leq R, \forall\mathbf{x} \in \mathcal{X}$. Then $\lim_{\sigma\to\infty} f_\sigma(\mathbf{x}) = c, \forall\mathbf{x} \in \mathcal{X}$, i.e., f_σ is constant function when $\sigma \to \infty$. In addition, for any of six considered explanation method ϕ, we have $\phi_\sigma(\mathbf{x}) = \mathbf{0}, \forall x \in \mathcal{X}$.*

Proof. Intuitively, f_σ is the average of f in a neighborhood. When $\sigma \to \infty$, $f_\sigma(\mathbf{x})$ is the average of $f(\mathcal{X})$. More formally,

$$\nabla f_\sigma(\mathbf{x}) = \nabla\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[f(\mathbf{x}+\epsilon)] = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[\nabla f(\mathbf{x}+\epsilon)]$$

By Stein's lemma, this also equals to

$$\nabla f_\sigma(\mathbf{x}) = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[\sigma^{-2}\epsilon f(\mathbf{x}+\epsilon)] = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}[\frac{\epsilon}{\sigma}f(\mathbf{x}+\sigma\epsilon)]$$

Since f is bounded above, we have

$$\|\nabla f_\sigma(\mathbf{x})\| = \|\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}[\frac{\epsilon}{\sigma}f(\mathbf{x}+\sigma\epsilon)]\| \leq \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\mathbf{I})}[\|\frac{\epsilon}{\sigma}\|R] \to 0, \sigma \to \infty$$

This holds for all $\mathbf{x} \in \mathcal{X}$, which means that

$$\lim_{\sigma\to\infty} f_\sigma(\mathbf{x}) = c, \forall\mathbf{x} \in \mathcal{X}$$

Since the gradient of f_σ is 0 when $\sigma \to \infty$, for any gradient based explanation method ϕ, the corresponding $\phi_\sigma = \mathbf{0}$. For SHAP, $f_\sigma(S \cup \{i\}) = f_\sigma(S)$, and therefore $(\phi_\sigma)_i = 0, i = 1, \cdots, d$. For LIME, it tends to a constant times c from the expression of $\mathbf{x}$ in Lemma C.8. In the official implementation of LIME[4], f_σ is pre-processed to be zero-mean, and thus $c = 0$ which means $\phi_\sigma(\mathbf{x}) = 0$. □

Theorem C.4. *For $f : \mathcal{X} \to \mathcal{Y}$ is C^1, i.e., its gradient is continuous. Suppose f is bounded above, i.e., $\exists R > 0, \|f(\mathbf{x})\| \leq R, \forall \mathbf{x} \in \mathcal{X}$. If the following three conditions hold*

1. *$\exists \alpha > 0$, for any $0 < \sigma < \infty$, $\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x}) \rangle > \alpha \|\phi_\sigma(\mathbf{x})\| \|p(\mathbf{x})\|$,*
2. *$\exists \tau, \nu > 0$ s.t. $\nu < \|p(\mathbf{x})\| / \|\phi(\mathbf{x})\| \leq \tau$*
3. *and if $\tau > 1/2$ then $2\tau - 1 - \alpha^2\nu^2 < 0$*

then there exists $0 < \sigma^* < \infty$, such that

$$\sigma^* = \arg\max_{\sigma} \mathcal{F}(\phi_\sigma)$$

that is, as σ increases from 0 to $+\infty$, the faithfulness of ϕ_σ has a trend that first increases and then decreases.

Proof. We define unfaithfulness $\mathcal{U}(\phi_\sigma) = \mathcal{F}^{-2}(\phi_{\sigma^*}) = \|\phi_\sigma(\mathbf{x}) - p(\mathbf{x})\|^2$.

We first derive the expressions for $\mathcal{U}(\phi_0), \mathcal{U}(\phi_\infty)$. Since f is continuous, by Lebesgue's dominated convergence theorem, we have

$$\begin{aligned}
\mathcal{U}(\phi_0) &= \lim_{\sigma \to 0} \mathcal{U}(\phi_\sigma) \\
&= \lim_{\sigma \to 0} \|\phi_\sigma(\mathbf{x}) - p(\mathbf{x})\|^2 \\
&= \|\phi(\mathbf{x}) - p(\mathbf{x})\|^2
\end{aligned}$$

By Lemma C.3

$$\begin{aligned}
\mathcal{U}(\phi_\infty) &= \lim_{\sigma \to \infty} \mathcal{U}(\phi_\sigma) \\
&= \lim_{\sigma \to \infty} \|\phi_\sigma(\mathbf{x}) - p(\mathbf{x})\|^2 \\
&= \|p(\mathbf{x})\|^2
\end{aligned}$$

Next, we prove the existence of σ^*.

$$\begin{aligned}
\mathcal{U}(\phi_\sigma) - \mathcal{U}(\phi_\infty) &= \|\phi_\sigma(\mathbf{x}) - p(\mathbf{x})\|^2 - \|p(\mathbf{x})\|^2 \\
&= \|\phi_\sigma(\mathbf{x})\|^2 - 2\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x}) \rangle
\end{aligned}$$

$$\begin{aligned}
\mathcal{U}(\phi_\sigma) - \mathcal{U}(\phi_0) &= \|\phi_\sigma(\mathbf{x}) - p(\mathbf{x})\|^2 - \|\phi(\mathbf{x}) - p(\mathbf{x})\|^2 \\
&= \|\phi_\sigma(\mathbf{x}) - \phi(\mathbf{x})\|^2 + 2\langle \phi_\sigma(\mathbf{x}) - \phi(\mathbf{x}), \phi(\mathbf{x}) - p(\mathbf{x}) \rangle
\end{aligned}$$

In order to prove the existence of σ^*, we only need to prove that there exists $0 < \sigma < \infty$, satisfies the following three conditions simultaneously.

$$\begin{aligned}
\|\phi_\sigma(\mathbf{x})\|^2 - 2\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x}) \rangle &\leq 0 \\
\|\phi_\sigma(\mathbf{x}) - \phi(\mathbf{x})\|^2 + 2\langle \phi_\sigma(\mathbf{x}) - \phi(\mathbf{x}), \phi(\mathbf{x}) - p(\mathbf{x}) \rangle &\leq 0
\end{aligned}$$

Next, we consider two cases:

[4] https://github.com/marcotcr/lime

1. If $\mathcal{U}(\phi_0) > \mathcal{U}(\phi_\infty)$, we have

$$\begin{aligned}\mathcal{U}(\phi_0) - \mathcal{U}(\phi_\infty) =& \|\phi(\mathbf{x}) - p(\mathbf{x})\|^2 - \|p(\mathbf{x})\|^2 \\ =& \|\phi(\mathbf{x})\|^2 - 2\langle \phi(\mathbf{x}), p(\mathbf{x})\rangle > 0\end{aligned}$$

$$\mathcal{U}(\phi_\sigma) - \mathcal{U}(\phi_\infty) = \|\phi_\sigma(\mathbf{x})\|^2 - 2\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x})\rangle$$

Since for $0 < \sigma < \infty$,

$$\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x})\rangle > \alpha \|\phi_\sigma(\mathbf{x})\| \|p(\mathbf{x})\|$$

and

$$\|\phi_\sigma(\mathbf{x})\| \to 0, \sigma \to \infty$$

Because $\|\phi_\sigma(\mathbf{x})\|$ is continuous w.r.t. σ, and it equals to 0 when $\sigma = \infty$ and equals to $\|\phi(\mathbf{x})\|$ when $\sigma = 0$, then for any $\Delta > 0$ that is sufficiently small, $\exists \sigma_0, s.t.$

$$\frac{\Delta}{2} \leq \|\phi_\sigma(\mathbf{x})\| \leq \Delta$$

Then

$$\|\phi_\sigma(\mathbf{x})\|^2 - 2\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x})\rangle \leq \Delta^2 - \alpha\Delta \|p(\mathbf{x})\|$$

By choosing $\Delta < \alpha \|p(\mathbf{x})\|$, we have

$$\|\phi_\sigma(\mathbf{x})\|^2 - 2\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x})\rangle < 0$$

Therefore, setting $\sigma^* = \sigma_0$, we have

$$\begin{aligned}\mathcal{U}(\phi_{\sigma^*}) - \mathcal{U}(\phi_\infty) =& \|\phi_\sigma(\mathbf{x})\|^2 - 2\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x})\rangle < 0 \\ \implies 0 < \mathcal{U}(\phi_{\sigma^*}) <& \mathcal{U}(\phi_\infty) < \mathcal{U}(\phi_0) \\ \implies \mathcal{F}(\phi_{\sigma^*}) >& \mathcal{F}(\phi_\infty) > \mathcal{F}(\phi_0)\end{aligned}$$

2. If $\mathcal{U}(\phi_0) \leq \mathcal{U}(\phi_\infty)$, we have

$$\begin{aligned}\mathcal{U}(\phi_0) - \mathcal{U}(\phi_\infty) =& \|\phi(\mathbf{x}) - p(\mathbf{x})\|^2 - \|p(\mathbf{x})\|^2 \\ =& \|\phi(\mathbf{x})\|^2 - 2\langle \phi(\mathbf{x}), p(\mathbf{x})\rangle < 0\end{aligned}$$

$$\begin{aligned}\mathcal{U}(\phi_\sigma) - \mathcal{U}(\phi_0) =& \|\phi_\sigma(\mathbf{x}) - \phi(\mathbf{x})\|^2 + 2\langle \phi_\sigma(\mathbf{x}) - \phi(\mathbf{x}), \phi(\mathbf{x}) - p(\mathbf{x})\rangle \\ =& \|\phi_\sigma(\mathbf{x})\|^2 + \|\phi(\mathbf{x})\|^2 - 2\|\phi(\mathbf{x})\|^2 \\ &+ 2\langle \phi(\mathbf{x}), p(\mathbf{x})\rangle - 2\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x})\rangle \\ =& \|\phi_\sigma(\mathbf{x})\|^2 - \|\phi(\mathbf{x})\|^2 \\ &+ 2\langle \phi(\mathbf{x}), p(\mathbf{x})\rangle - 2\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x})\rangle\end{aligned}$$

Since we assume that $\nu\|\phi(\mathbf{x})\| \leq \|p(\mathbf{x})\| \leq \tau\|\phi(\mathbf{x})\|$ then

$$\langle \phi(\mathbf{x}), p(\mathbf{x})\rangle \leq \|\phi(\mathbf{x})\| \|p(\mathbf{x})\| = \tau \|\phi(\mathbf{x})\|^2$$

$$0 < -\|\phi(\mathbf{x})\|^2 + 2\langle \phi(\mathbf{x}), p(\mathbf{x})\rangle \leq (2\tau - 1)\|\phi(\mathbf{x})\|^2$$

$$-\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x})\rangle \leq -\alpha\|\phi_\sigma(\mathbf{x})\| \|p(\mathbf{x})\| \leq -\alpha\nu \|\phi_\sigma(\mathbf{x})\| \|\phi(\mathbf{x})\|$$

Combining these inequalities, we have

$$\begin{aligned}&\|\phi_\sigma(\mathbf{x})\|^2 - \|\phi(\mathbf{x})\|^2 + 2\langle \phi(\mathbf{x}), p(\mathbf{x})\rangle - 2\langle \phi_\sigma(\mathbf{x}), p(\mathbf{x})\rangle \\ \leq& \|\phi_\sigma(\mathbf{x})\|^2 + (2\tau - 1)\|\phi(\mathbf{x})\|^2 - 2\alpha\nu\|\phi_\sigma(\mathbf{x})\| \|\phi(\mathbf{x})\| \\ =& \big(\|\phi_\sigma(\mathbf{x})\| - \alpha\nu\|\phi(\mathbf{x})\|\big)^2 + (2\tau - 1 - \alpha^2\nu^2)\|\phi(\mathbf{x})\|^2\end{aligned}$$

Because $\|\phi_\sigma(\mathbf{x})\|$ is continuous w.r.t. σ, and it equals to 0 when $\sigma = \infty$ and equals to $\|\phi(\mathbf{x})\|$ when $\sigma = 0$, thus $\exists 0 < \hat{\sigma} < \infty, s.t., \|\phi_\sigma(\mathbf{x})\| = \alpha\nu\|\phi(\mathbf{x})\|$, which implies

$$\mathcal{U}(e_{f,\hat{\sigma}}) - \mathcal{U}(\phi_0) \leq (2\tau - 1 - \alpha^2\nu^2)\|\phi(\mathbf{x})\|^2 < 0$$

Therefore, setting $\sigma^* = \hat{\sigma}$, we have

$$\begin{aligned} 0 < \mathcal{U}(\phi_{\sigma^*}) &< \mathcal{U}(\phi_0) \leq \mathcal{U}(\phi_\infty) \\ \implies \mathcal{F}(\phi_{\sigma^*}) &> \mathcal{F}(\phi_\infty) \geq \mathcal{F}(\phi_0) \end{aligned}$$

In summary, there exists $0 < \sigma^* < \infty$ that achieves maximum of $\mathcal{F}(\phi_\sigma)$.

□

C.4. Proofs of Theorem 4.1

Theorem C.5. *Let $f : \mathcal{X} \to \mathcal{Y}$ be a (δ, L)-Lipschitz function, then we have SmoothGrad, LIME, SHAP are $(\delta, \mathcal{L})$-Lipschitz with corresponding $\mathcal{L}$ as the following:*

$$\begin{aligned} &\mathcal{L}_{\mathit{SmoothGrad}} = \mathcal{O}(L/\sigma) \\ &\mathcal{L}_{\mathit{LIME}} = \mathcal{O}(\frac{\sqrt{d}L}{\lambda} + \frac{\beta R(\lambda + d)\sqrt{d}}{\lambda^2\sigma^2}\exp(\frac{2\beta}{\sigma^2})) \\ &\mathcal{L}_{SHAP} = \mathcal{O}(\sqrt{d}L) \end{aligned}$$

If f is also H-smooth, then we have Gradient, Gradient$\times$Input and Integrated Gradient(IG) are $(\delta, \mathcal{L})$-Lipschitz with corresponding $\mathcal{L}$ as the following:

$$\begin{aligned} &\mathcal{L}_{\mathit{Gradient}} = \mathcal{O}(H) \\ &\mathcal{L}_{\mathit{Gradient}\times\mathit{Input}} = \mathcal{O}(\beta H + L) \\ &\mathcal{L}_{\mathit{IG}} = \mathcal{O}(\beta H + L) \end{aligned}$$

C.4.1. SmoothGrad

Lemma C.6. *For an univariate Gaussian variable $\epsilon \sim \mathcal{N}(0, \sigma^2)$, we have $\mathbb{E}|\epsilon| = \sqrt{\frac{2}{\pi}}\sigma$*

Proof.

$$\begin{aligned} \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2)}[|\epsilon|] =& 2\int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} x\exp(-\frac{x^2}{2\sigma^2})\mathrm{d}x \\ =& \sqrt{\frac{2}{\pi}}\int_0^\infty \sigma\exp(\frac{x^2}{2\sigma^2})\mathrm{d}(\frac{x^2}{2\sigma^2}) \\ =& \sqrt{\frac{2}{\pi}}\sigma\int_0^\infty e^{-\mathbf{x}}\mathrm{d}x \\ =& \sqrt{\frac{2}{\pi}}\sigma \end{aligned}$$

□

Theorem C.7. *For SmoothGrad, if f is $(\delta, L)-$Lipschitz, then we have $\phi(\mathbf{x})$ is $(\delta, \sqrt{\frac{2}{\pi\sigma^2}}L)-$ Lipschitz.*

Proof. For x, x', $\|\mathbf{x}-\mathbf{x}'\|_2 \leq \delta$, we have

$$\begin{aligned}
\|\phi(\mathbf{x})-\phi(\mathbf{x}')\|_2 =& \|\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[\nabla f(\mathbf{x}+\epsilon)-\nabla f(\mathbf{x}'+\epsilon)]\|_2 \\
\text{(Stein's Lemma)} =& \|\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[\sigma^{-2}\epsilon(f(\mathbf{x}+\epsilon)-f(\mathbf{x}'+\epsilon))]\|_2 \\
=& \sigma^{-2}\sup_{u:\|u\|_2=1}|\langle u, \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[\epsilon(f(\mathbf{x}+\epsilon)-f(\mathbf{x}'+\epsilon))]\rangle| \\
=& \sigma^{-2}\sup_{u:\|u\|_2=1}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[|\langle u,\epsilon\rangle||f(\mathbf{x}+\epsilon)-f(\mathbf{x}'+\epsilon)|] \\
(f\text{ is }(\delta, L)\text{−Lipschitz}) \leq& \sigma^{-2}\sup_{u:\|u\|_2=1}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[|\langle u,\epsilon\rangle|L\|x-x'\|]
\end{aligned}$$

The third equality holds because the L_2 norm of a vector is the largest length of its projection on the unit L_2 ball:

$$\|v\|_2 = \sup_{u:\|u\|_2=1}|\langle u, v\rangle|$$

In order to draw our conclusion, we only need to bound

$$\sup_{u:\|u\|_2=1}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[|\langle u,\epsilon\rangle|]$$

Since $z = \langle u, \epsilon\rangle$ is a linear combination of Gaussian variables which in turn is also Gaussian. Since $\epsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I})$, it is easy to see that

$$\mathbb{E}[z] = 0, \mathbb{E}[z^2] = \mathbb{E}(\sum_i u_i\epsilon_i)^2 = \mathbb{E}\sum_i u_i^2\epsilon_i^2 = \sigma^2\|u\|_2^2 = \sigma^2$$

Therefore, $z \sim \mathcal{N}(0, \sigma^2)$. By Lemma C.6, we have

$$\sup_{u:\|u\|_2=1}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[|\langle u,\epsilon\rangle|] = \sqrt{\frac{2}{\pi}}\sigma$$

And it follows that

$$\|\phi(\mathbf{x})-\phi(\mathbf{x}')\|_2 \leq \sqrt{\frac{2}{\pi\sigma^2}}L\|x-x'\|$$

□

C.4.2. LIME

We first derive the closed form solution of $\mathbf{w}$ in LIME in terms of $\pi_{\mathbf{x}}$.

Lemma C.8. *For LIME with L_2 penalty:*

$$\phi(\mathbf{x}) = \arg\min_{\mathbf{w}}\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)(f(\mathbf{x}\odot\epsilon)-\mathbf{w}^\top\epsilon)^2] + \lambda\|\mathbf{w}\|_2^2,$$

we have the closed form of $\phi(\mathbf{x})$ is

$$\phi(\mathbf{x}) = \left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)\epsilon\epsilon^\top] + \lambda\mathbf{I}\right)^{-1}\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)f(\mathbf{x}\odot\epsilon)\epsilon]$$

If $\pi_{\mathbf{x}} = 1$, then

$$\phi(\mathbf{x}) = \frac{1}{\lambda+\frac{1}{4}}(\mathbf{I}-\frac{1}{\lambda+\frac{1}{4}+d}\mathbf{1}\mathbf{1}^\top)\mathbb{E}_{\epsilon\sim Bern(0.5)}[f(\mathbf{x}\odot\epsilon)\epsilon]$$

Proof. Let $O(\mathbf{x})$ be the objective function in $\phi(\mathbf{x})$, then the gradient of O w.r.t w is

$$\nabla_{\mathbf{w}}O = \mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)(2\epsilon\epsilon^\top\mathbf{w}-2f(\mathbf{x}\odot\epsilon)\epsilon)] + 2\lambda\mathbf{w}$$

For optimal $\mathbf{w}$, we have $\nabla_{\mathbf{w}} O = 0$, that is

$$\mathbf{w} = \left(\mathbb{E}_{\epsilon \sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon)\epsilon\epsilon^\top] + \lambda \mathbf{I}\right)^{-1} \mathbb{E}_{\epsilon \sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon) f(\mathbf{x} \odot \epsilon)\epsilon]$$

Therefore,

$$\phi(\mathbf{x}) = \mathbf{w} = \left(\mathbb{E}_{\epsilon \sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon)\epsilon\epsilon^\top] + \lambda \mathbf{I}\right)^{-1} \mathbb{E}_{\epsilon \sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon) f(\mathbf{x} \odot \epsilon)\epsilon]$$

If $\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon) = 1$, we have

$$\begin{aligned}
\mathbf{w} =& \left[\mathbb{E}_{\epsilon \sim Bern(0.5)}[\epsilon\epsilon^\top] + \lambda \mathbf{I}\right]^{-1} \mathbb{E}_{\epsilon \sim Bern(0.5)}[f(\mathbf{x} \odot \epsilon)\epsilon] \\
=& \big(\frac{1}{4}\mathbf{1}\mathbf{1}^\top + (\frac{1}{4} + \lambda)\mathbf{I}\big)^{-1} \mathbb{E}_{\epsilon \sim Bern(0.5)}[f(\mathbf{x} \odot \epsilon)\epsilon] \\
=& \frac{1}{4\lambda + 1}\big(\frac{1}{\lambda + \frac{1}{4}}\mathbf{1}\mathbf{1}^\top + \mathbf{I}\big)^{-1} \mathbb{E}_{\epsilon \sim Bern(0.5)}[f(\mathbf{x} \odot \epsilon)\epsilon] \\
\text{(Sherman-Morrison Formula)} =& \frac{4}{4\lambda + 1}\big(\mathbf{I} - \frac{1}{4\lambda + 1 + d}\mathbf{1}\mathbf{1}^\top\big) \mathbb{E}_{\epsilon \sim Bern(0.5)}[f(\mathbf{x} \odot \epsilon)\epsilon]
\end{aligned}$$

□

Before diving into the derivation of LIME with the exponential kernel, we first provide the local Lipschitz of LIME with $\pi_{\mathbf{x}} = 1$. The proof is much simpler, but the overall process is similar. Thus, readers can get an overview of how we obtain the local Lipschitz of LIME with exponential kernel.

Lemma C.9. *For LIME with L_2 penalty:*

$$\phi(\mathbf{x}) = \arg\min_{\mathbf{w}} \mathbb{E}_{\epsilon \sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon)(f(\mathbf{x} \odot \epsilon) - \mathbf{w}^\top \epsilon)^2] + \lambda \|\mathbf{w}\|_2^2,$$

where

$$\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon) = 1$$

we have $\phi(\mathbf{x})$ is $(\delta, 2\frac{\sqrt{d+1}L}{4\lambda+1})$-Lipschitz.

Proof. If $\pi_{\mathbf{x}}(\mathbf{z}) = 1, \forall \mathbf{x}, \mathbf{z}$, then we have

$$\begin{aligned}
\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2 =& \|\frac{4}{4\lambda + 1}\big(\mathbf{I} - \frac{1}{4\lambda + 1 + d}\mathbf{1}\mathbf{1}^\top\big) \mathbb{E}_{\epsilon \sim Bern(0.5)}[(f(\mathbf{x} \odot \epsilon) - f(\mathbf{x}' \odot \epsilon))\epsilon]\|_2 \\
\leq& \|\frac{4}{4\lambda + 1}\big(\mathbf{I} - \frac{1}{4\lambda + 1 + d}\mathbf{1}\mathbf{1}^\top\big)\|_2 \|\mathbb{E}_{\epsilon \sim Bern(0.5)}[(f(\mathbf{x} \odot \epsilon) - f(\mathbf{x}' \odot \epsilon))\epsilon]\|_2 \\
\leq& C_\lambda \mathbb{E}_{\epsilon \sim Bern(0.5)}[|f(\mathbf{x} \odot \epsilon) - f(\mathbf{x}' \odot \epsilon)| \|\epsilon\|_2] \\
\leq& C_\lambda \mathbb{E}_{\epsilon \sim Bern(0.5)}[L\|(\mathbf{x} - \mathbf{x}') \odot \epsilon\|_2 \|\epsilon\|_2]
\end{aligned}$$

$$
\begin{aligned}
\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\|(\mathbf{x}-\mathbf{x}')\odot\epsilon\|_2\|\epsilon\|_2]\right)^2 \leq & \mathbb{E}_{\epsilon\sim Bern(0.5)}[\|(\mathbf{x}-\mathbf{x}')\odot\epsilon\|_2^2\|\epsilon\|_2^2] \\
= & \mathbb{E}_{\epsilon\sim Bern(0.5)}\left[\Big(\sum_i(\mathbf{x}_i-\mathbf{x}'_i)^2\epsilon_i^2\Big)\Big(\sum_j\epsilon_j^2\Big)\right] \\
= & \mathbb{E}_{\epsilon\sim Bern(0.5)}\left[\sum_i(\mathbf{x}_i-\mathbf{x}'_i)^2\sum_j\epsilon_j^2\epsilon_i^2\right] \\
= & \sum_i\mathbb{E}_{\epsilon\sim Bern(0.5)}\left[(\mathbf{x}_i-\mathbf{x}'_i)^2\sum_j\epsilon_j^2\epsilon_i^2\right] \\
= & \frac{1+d}{4}\sum_i(\mathbf{x}_i-\mathbf{x}'_i)^2 = \frac{1+d}{4}\|\mathbf{x}-\mathbf{x}'\|_2^2
\end{aligned}
$$

Therefore,

$$
\mathbb{E}_{\epsilon\sim Bern(0.5)}[\|(\mathbf{x}-\mathbf{x}')\odot\epsilon\|_2\|\epsilon\|_2] \leq \sqrt{\frac{1+d}{4}}\|\mathbf{x}-\mathbf{x}'\|_2 = \frac{\sqrt{1+d}}{2}\|\mathbf{x}-\mathbf{x}'\|_2
$$

and

$$
\begin{aligned}
\|\phi(\mathbf{x})-\phi(\mathbf{x}')\|_2 \leq & C_\lambda\mathbb{E}_{\epsilon\sim Bern(0.5)}[L\|(\mathbf{x}-\mathbf{x}')\odot\epsilon\|_2\|\epsilon\|_2] \\
\leq & \frac{\sqrt{d+1}C_\lambda L}{2}\|\mathbf{x}-\mathbf{x}'\|_2
\end{aligned}
$$

where

$$
C_\lambda = \|\frac{4}{4\lambda+1}\Big(\mathbf{I}-\frac{1}{4\lambda+1+d}\mathbf{1}\mathbf{1}^\top\Big)\|_2 = \frac{4}{4\lambda+1}
$$

□

Theorem C.10. *For LIME with L_2 penalty:*

$$
\phi(\mathbf{x}) = \arg\min_{\mathbf{w}}\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)(f(\mathbf{x}\odot\epsilon)-\mathbf{w}^\top\epsilon)^2]+\lambda\|\mathbf{w}\|_2^2,
$$

where

$$
\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}\odot\epsilon\|_2^2}{\sigma^2}\right) = \exp\left(-\frac{\|\mathbf{x}\odot(1-\epsilon)\|_2^2}{\sigma^2}\right)
$$

we have $\phi(\mathbf{x})$ is $(\delta, \mathcal{O}(\frac{\sqrt{d}L}{\lambda}+\frac{\beta R(\lambda+d)\sqrt{d}}{\lambda^2\sigma^2}\exp(\frac{2\beta}{\sigma^2})))$-Lipschitz

Proof. For exponential kernel $\pi_{\mathbf{x}}(\mathbf{z}) = \exp(-\|\mathbf{x}-\mathbf{z}\|_2^2/\sigma^2)$, we have

$$
\begin{aligned}
\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2 =& \|\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)\epsilon\epsilon^\top] + \lambda\mathbf{I}\right)^{-1}\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)f(\mathbf{x}\odot\epsilon)\epsilon] \\
& - \left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)\epsilon\epsilon^\top] + \lambda\mathbf{I}\right)^{-1}\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)f(\mathbf{x}'\odot\epsilon)\epsilon]\|_2 \\
=& \| \underbrace{\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)\epsilon\epsilon^\top] + \lambda\mathbf{I}\right)^{-1}\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)f(\mathbf{x}\odot\epsilon)\epsilon]}_{(a)} \\
& \underbrace{-\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)\epsilon\epsilon^\top] + \lambda\mathbf{I}\right)^{-1}\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)f(\mathbf{x}'\odot\epsilon)\epsilon]}_{(a)} \\
& + \underbrace{\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)\epsilon\epsilon^\top] + \lambda\mathbf{I}\right)^{-1}\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)f(\mathbf{x}'\odot\epsilon)\epsilon]}_{(b)} \\
& \underbrace{-\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)\epsilon\epsilon^\top] + \lambda\mathbf{I}\right)^{-1}\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)f(\mathbf{x}'\odot\epsilon)\epsilon]\|_2}_{(b)}
\end{aligned}
$$

We bound $(a), (b)$ separately in the following.

$$
\begin{aligned}
\|(a)\|_2 =& \|\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)\epsilon\epsilon^\top] + \lambda\mathbf{I}\right)^{-1}\mathbb{E}_{\epsilon\sim Bern(0.5)}\left[(\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)f(\mathbf{x}\odot\epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)f(\mathbf{x}'\odot\epsilon))\epsilon\right]\|_2 \\
\leq& \| \underbrace{\left(\frac{1}{4}e^{-\frac{\beta^2}{\sigma^2}}\mathbf{1}\mathbf{1}^\top + (\frac{1}{4}e^{-\frac{\beta^2}{\sigma^2}} + \lambda)\mathbf{I}\right)^{-1}}_{\eta}\mathbb{E}_{\epsilon\sim Bern(0.5)}\left[(\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)f(\mathbf{x}\odot\epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)f(\mathbf{x}'\odot\epsilon))\epsilon\right]\|_2 \\
=& \|\eta\mathbb{E}_{\epsilon\sim Bern(0.5)}\left[(\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)f(\mathbf{x}\odot\epsilon) - \pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)f(\mathbf{x}'\odot\epsilon) + \pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)f(\mathbf{x}'\odot\epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)f(\mathbf{x}'\odot\epsilon))\epsilon\right]\|_2 \\
\leq& \|\eta\|_2\Big[\mathbb{E}_{\epsilon\sim Bern(0.5)}\left[\|\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)(f(\mathbf{x}\odot\epsilon) - f(\mathbf{x}'\odot\epsilon))\|_2\|\epsilon\|_2\right] \\
& + \mathbb{E}_{\epsilon\sim Bern(0.5)}\left[\|(\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon))f(\mathbf{x}'\odot\epsilon)\|_2\|\epsilon\|_2\right]\Big] \\
\leq& \|\eta\|_2\left[\mathbb{E}_{\epsilon\sim Bern(0.5)}\left[\|f(\mathbf{x}\odot\epsilon) - f(\mathbf{x}'\odot\epsilon)\|_2\|\epsilon\|_2\right] + \mathbb{E}_{\epsilon\sim Bern(0.5)}\left[\|\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)\|_2 R\|\epsilon\|_2\right]\right] \\
\leq& \|\eta\|_2\left[\mathbb{E}_{\epsilon\sim Bern(0.5)}\left[L\|(\mathbf{x}-\mathbf{x}')\odot\epsilon\|_2\|\epsilon\|_2\right] + \mathbb{E}_{\epsilon\sim Bern(0.5)}\left[R|\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)|\|\epsilon\|_2\right]\right] \\
\leq& \|\eta\|_2\left[\frac{\sqrt{1+d}L}{2}\|\mathbf{x}-\mathbf{x}'\|_2 + \mathbb{E}_{\epsilon\sim Bern(0.5)}\left[R|\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)|\|\epsilon\|_2\right]\right]
\end{aligned}
$$

If $\|\mathbf{x} \odot (1-\epsilon)\|_2 > \|\mathbf{x}' \odot (1-\epsilon)\|_2$,

$$\begin{aligned}
|\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}' \odot \epsilon)| =& | \exp(-\frac{\|\mathbf{x} \odot (1-\epsilon)\|_2^2}{\sigma^2}) - \exp(-\frac{\|\mathbf{x}' \odot (1-\epsilon)\|_2^2}{\sigma^2})| \\
=& \exp(-\frac{\|\mathbf{x} \odot (1-\epsilon)\|_2^2}{\sigma^2})\big(\exp(\frac{\|\mathbf{x} \odot (1-\epsilon)\|_2^2 - \|\mathbf{x}' \odot (1-\epsilon)\|_2^2}{\sigma^2}) - 1\big) \\
\leq& \exp(\frac{(\|\mathbf{x} \odot (1-\epsilon)\|_2 + \|\mathbf{x}' \odot (1-\epsilon)\|_2)(\|\mathbf{x} \odot (1-\epsilon)\|_2 - \|\mathbf{x}' \odot (1-\epsilon)\|_2)}{\sigma^2}) - 1 \\
\leq& \exp(\frac{2\beta(\|\mathbf{x} \odot (1-\epsilon)\|_2 - \|\mathbf{x}' \odot (1-\epsilon)\|_2)}{\sigma^2}) - 1 \\
\leq& \exp(\frac{2\beta\|(\mathbf{x} - \mathbf{x}') \odot (1-\epsilon)\|_2}{\sigma^2}) - 1
\end{aligned}$$

The last inequality follows from triangle inequality.

If on the other hand $\|\mathbf{x}' \odot (1-\epsilon)\|_2 > \|\mathbf{x} \odot (1-\epsilon)\|_2$,

$$\begin{aligned}
|\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}' \odot \epsilon)| =& | \exp(-\frac{\|\mathbf{x} \odot (1-\epsilon)\|_2^2}{\sigma^2}) - \exp(-\frac{\|\mathbf{x}' \odot (1-\epsilon)\|_2^2}{\sigma^2})| \\
=& \exp(-\frac{\|\mathbf{x}' \odot (1-\epsilon)\|_2^2}{\sigma^2})\big(\exp(\frac{\|\mathbf{x}' \odot (1-\epsilon)\|_2^2 - \|\mathbf{x} \odot (1-\epsilon)\|_2^2}{\sigma^2}) - 1\big) \\
\leq& \exp(\frac{(\|\mathbf{x}' \odot (1-\epsilon)\|_2 + \|\mathbf{x} \odot (1-\epsilon)\|_2)(\|\mathbf{x}' \odot (1-\epsilon)\|_2 - \|\mathbf{x} \odot (1-\epsilon)\|_2)}{\sigma^2}) - 1 \\
\leq& \exp(\frac{2\beta(\|\mathbf{x}' \odot (1-\epsilon)\|_2 - \|\mathbf{x} \odot (1-\epsilon)\|_2)}{\sigma^2}) - 1 \\
\leq& \exp(\frac{2\beta\|(\mathbf{x} - \mathbf{x}') \odot (1-\epsilon)\|_2}{\sigma^2}) - 1
\end{aligned}$$

If $\|\mathbf{x}' \odot (1-\epsilon)\|_2 = \|\mathbf{x} \odot (1-\epsilon)\|_2$, the bound we derive in the following also holds.

Consider function $q(z) = \exp(\frac{2\beta z}{\sigma^2}) - 1 - \exp(\frac{2\beta}{\sigma^2})\frac{2\beta z}{\sigma^2}$,

$$q'(z) = \frac{2\beta}{\sigma^2}\exp(\frac{2\beta z}{\sigma^2}) - \exp(\frac{2\beta}{\sigma^2})\frac{2\beta}{\sigma^2}$$

Let $q'(z) = 0$, we have $z = 1$. Then for $z \leq 1$, we have $q'(z) \leq 0$. Overall, we have $q(z) \leq q(0) = 0, \forall z \leq 1$, that is

$$\exp(\frac{2\beta z}{\sigma^2}) - 1 \leq \exp(\frac{2\beta}{\sigma^2})\frac{2\beta z}{\sigma^2}, \forall z \leq 1$$

Therefore, as $\|(\mathbf{x} - \mathbf{x}') \odot (1-\epsilon)\|_2 \leq \|\mathbf{x} - \mathbf{x}'\|_2 \leq \delta \leq 1$

$$|\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}' \odot \epsilon)| \leq \exp(\frac{2\beta\|(\mathbf{x} - \mathbf{x}') \odot (1-\epsilon)\|_2}{\sigma^2}) - 1 \leq \exp(\frac{2\beta}{\sigma^2})\frac{2\beta}{\sigma^2}\|(\mathbf{x} - \mathbf{x}') \odot (1-\epsilon)\|_2$$

Thus,

$$\begin{aligned}
\mathbb{E}_{\epsilon \sim Bern(0.5)}\big[R|\pi_{\mathbf{x}}(\mathbf{x} \odot \epsilon) - \pi_{\mathbf{x}'}(\mathbf{x}' \odot \epsilon)|\|\epsilon\|_2\big] \leq& \mathbb{E}_{\epsilon \sim Bern(0.5)}\big[R\exp(\frac{2\beta}{\sigma^2})\frac{2\beta}{\sigma^2}\|\|(\mathbf{x} - \mathbf{x}') \odot (1-\epsilon)\|_2 \epsilon\|_2\big] \\
\leq& R\frac{2\beta}{\sigma^2}\exp(\frac{2\beta}{\sigma^2})\frac{\sqrt{d-1}}{2}\|\mathbf{x} - \mathbf{x}'\|_2 \\
=& \frac{\beta R\sqrt{d-1}}{\sigma^2}\exp(\frac{2\beta}{\sigma^2})\|\mathbf{x} - \mathbf{x}'\|_2
\end{aligned}$$

The last inequality is due to the following fact

$$
\begin{aligned}
\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\|(\mathbf{x}-\mathbf{x}')\odot(1-\epsilon)\|_2\|\epsilon\|_2]\right)^2 \leq & \mathbb{E}_{\epsilon\sim Bern(0.5)}[\|(\mathbf{x}-\mathbf{x}')\odot\epsilon\|_2^2\|\epsilon\|_2^2] \\
= & \mathbb{E}_{\epsilon\sim Bern(0.5)}\left[\big(\sum_i(\mathbf{x}_i-\mathbf{x}'_i)^2(1-\epsilon_i)^2\big)\big(\sum_j\epsilon_j^2\big)\right] \\
= & \mathbb{E}_{\epsilon\sim Bern(0.5)}\left[\sum_i(\mathbf{x}_i-\mathbf{x}'_i)^2\sum_j\epsilon_j^2(1-\epsilon_i)^2\right] \\
= & \sum_i\mathbb{E}_{\epsilon\sim Bern(0.5)}\left[(\mathbf{x}_i-\mathbf{x}'_i)^2\sum_j\epsilon_j^2(1-\epsilon_i)^2\right] \\
= & \frac{d-1}{4}\sum_i(\mathbf{x}_i-\mathbf{x}'_i)^2=\frac{d-1}{4}\|\mathbf{x}-\mathbf{x}'\|_2^2
\end{aligned}
$$

So far, we have proved the following upper bound for (a),

$$
\|(a)\|_2 \leq \|\eta\|_2\left[\frac{\sqrt{1+n}L}{2}+\frac{\beta R\sqrt{d-1}}{\sigma^2}\exp(\frac{2\beta}{\sigma^2})\right]\|\mathbf{x}-\mathbf{x}'\|_2
$$

Next, we show how to upper bound (b).

$$
\begin{aligned}
\|(b)\|_2 =& \| \underbrace{\left[\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)\epsilon\epsilon^\top]+\lambda\mathbf{I}\right)^{-1}-\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)\epsilon\epsilon^\top]+\lambda\mathbf{I}\right)^{-1}\right]}_{c} \\
& \mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)f(\mathbf{x}'\odot\epsilon)\epsilon]\|_2
\end{aligned}
$$

Let

$$
\gamma(\mathbf{x})=\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)\epsilon\epsilon^\top]+\lambda\mathbf{I}\right)^{-1},\mu(\mathbf{x})=\gamma(\mathbf{x})^{-1}=\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)\epsilon\epsilon^\top]+\lambda\mathbf{I}
$$

$$
\begin{aligned}
\|c\|_2 =& \|\gamma(\mathbf{x})-\gamma(\mathbf{x}')\|_2=\|\gamma(\mathbf{x})(\mu(\mathbf{x}')-\mu(\mathbf{x}))\gamma(\mathbf{x}')\|_2 \\
\leq& \|\gamma(\mathbf{x})\|_2\|\mu(\mathbf{x}')-\mu(\mathbf{x})\|_2\|\gamma(\mathbf{x}')\|_2
\end{aligned}
$$

Since $\pi_{\mathbf{x}}>0,\epsilon>0,\lambda>0$, we have

$$
\begin{aligned}
\|\gamma(\mathbf{x})\|_2 =& \|(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)\epsilon\epsilon^\top]+\lambda\mathbf{I})^{-1}\|_2 \\
\leq& \|(\exp(-\frac{\beta^2}{\sigma^2})\mathbb{E}_{\epsilon\sim Bern(0.5)}[\epsilon\epsilon^\top]+\lambda\mathbf{I})^{-1}\|_2 \\
=& \|(\frac{1}{4}e^{-\frac{\beta^2}{\sigma^2}}\mathbf{1}\mathbf{1}^\top+(\frac{1}{4}e^{-\frac{\beta^2}{\sigma^2}}+\lambda)\mathbf{I})^{-1}\|_2 \\
=& \|\eta\|_2
\end{aligned}
$$

$$
\begin{aligned}
\|\mu(\mathbf{x})-\mu(\mathbf{x}')\|_F =& \|\mathbb{E}_{\epsilon\sim Bern(0.5)}[(\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)-\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon))\epsilon\epsilon^\top]\|_2 \\
\leq& \|\mathbb{E}_{\epsilon\sim Bern(0.5)}[|\pi_{\mathbf{x}}(\mathbf{x}\odot\epsilon)-\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)|\epsilon\epsilon^\top]\|_2 \\
\leq& \mathbb{E}_{\epsilon\sim Bern(0.5)}[\exp(\frac{2\beta}{\sigma^2})\frac{2\beta}{\sigma^2}\|(\mathbf{x}-\mathbf{x}')\odot(1-\epsilon)\|_2\|\epsilon\epsilon^\top\|_2] \\
\leq& \frac{\sqrt{2}\beta(d-1)}{\sigma^2}\exp(\frac{2\beta}{\sigma^2})\|\mathbf{x}-\mathbf{x}'\|_2
\end{aligned}
$$

$$
\begin{aligned}
(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\|(\mathbf{x}-\mathbf{x}')\odot(1-\epsilon)\|_2\|\epsilon\epsilon^\top\|])^2 \leq &\mathbb{E}_{\epsilon\sim Bern(0.5)}[\|(\mathbf{x}-\mathbf{x}')\odot(1-\epsilon)\|_2^2\|\epsilon\epsilon^\top\|^2] \\
=&\sum_i(\mathbf{x}_i-\mathbf{x}'_i)^2\mathbb{E}_{\epsilon\sim Bern(0.5)}[(1-\epsilon_i)^2\sum_j\sum_k\epsilon_j^2\epsilon_k^2] \\
=&\sum_i(\mathbf{x}_i-\mathbf{x}'_i)^2\cdot\frac{(d-1)^2}{8}=\frac{(d-1)^2}{8}\|\mathbf{x}-\mathbf{x}'\|_2^2
\end{aligned}
$$

To summarize, we have

$$
\|c\|\leq\|\gamma(\mathbf{x})\|\|\mu(\mathbf{x}')-\mu(\mathbf{x})\|\|\gamma(\mathbf{x}')\|\leq\|\eta\|_2^2\frac{\sqrt{2}\beta(d-1)}{\sigma^2}\exp(\frac{2\beta}{\sigma^2})\|\mathbf{x}-\mathbf{x}'\|_2
$$

$$
\|\mathbb{E}_{\epsilon\sim Bern(0.5)}[\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)f(\mathbf{x}'\odot\epsilon)\epsilon]\|\leq\mathbb{E}_{\epsilon\sim Bern(0.5)}[|\pi_{\mathbf{x}'}(\mathbf{x}'\odot\epsilon)||f(\mathbf{x}'\odot\epsilon)|\|\epsilon\|]\leq R\mathbb{E}_{\epsilon\sim Bern(0.5)}[\|\epsilon\|]\leq\frac{\sqrt{2d}R}{2}
$$

The last inequality is derived as follows:

$$
\mathbb{E}_{\epsilon\sim Bern(0.5)}[\|\epsilon\|]\leq\left(\mathbb{E}_{\epsilon\sim Bern(0.5)}[\|\epsilon\|^2]\right)^{-\frac{1}{2}}=\frac{\sqrt{2d}}{2}
$$

where the inequality holds by Jensen's inequality.

Therefore, summarizing the above results, we have

$$
\|(b)\|_2\leq\|\eta\|_2^2\frac{\beta R(d-1)\sqrt{d}}{\sigma^2}\exp(\frac{2\beta}{\sigma^2})\|\mathbf{x}-\mathbf{x}'\|_2
$$

In order to obtain the final result, we only need to bound $\|\eta\|_2$.

$$
\begin{aligned}
\|\eta\|_2=&\|(\frac{1}{4}e^{-\frac{\beta^2}{\sigma^2}}\mathbf{1}\mathbf{1}^\top+(\frac{1}{4}e^{-\frac{\beta^2}{\sigma^2}}+\lambda)\mathbf{I})^{-1}\|_2 \\
=&\frac{4}{e^{-\frac{\beta^2}{\sigma^2}}+4\lambda}\|(\frac{e^{-\frac{\beta^2}{\sigma^2}}}{e^{-\frac{\beta^2}{\sigma^2}}+4\lambda}\mathbf{1}\mathbf{1}^\top+\mathbf{I})^{-1}\|_2 \\
=&\frac{4}{e^{-\frac{\beta^2}{\sigma^2}}+4\lambda}\|\mathbf{I}-\frac{e^{-\frac{\beta^2}{\sigma^2}}}{e^{-\frac{\beta^2}{\sigma^2}}+4\lambda+e^{-\frac{\beta^2}{\sigma^2}}d}\mathbf{1}\mathbf{1}^\top\|_2 \\
=&\frac{4}{e^{-\frac{\beta^2}{\sigma^2}}+4\lambda}=C_{\lambda,\sigma}
\end{aligned}
$$

In summary, we have

$$
\begin{aligned}
\|\phi(\mathbf{x})-\phi(\mathbf{x}')\|_2\leq&\left[C_{\lambda,\sigma}\left[\frac{\sqrt{1+d}L}{2}+\frac{\beta R\sqrt{d-1}}{\sigma^2}\exp(\frac{2\beta}{\sigma^2})\right]+C_{\lambda,\sigma}^2\frac{\beta R(d-1)\sqrt{d}}{\sigma^2}\exp(\frac{2\beta}{\sigma^2})\right]\|\mathbf{x}-\mathbf{x}'\|_2 \\
=&\mathcal{O}(\frac{\sqrt{d}L}{\lambda}+\frac{\beta R(\lambda+d)\sqrt{d}}{\lambda^2\sigma^2}\exp(\frac{2\beta}{\sigma^2}))\|\mathbf{x}-\mathbf{x}'\|_2
\end{aligned}
$$

When $\sigma\to+\infty$, it is easy to see that the coefficient in the bracket tends to

$$
C_{\lambda,\sigma}\left[\frac{\sqrt{1+d}L}{2}+\frac{\beta R\sqrt{d-1}}{\sigma^2}\exp(\frac{2\beta}{\sigma^2})\right]+C_{\lambda,\sigma}^2\frac{\beta R(d-1)\sqrt{d}}{\sigma^2}\exp(\frac{2\beta}{\sigma^2})\to\frac{\sqrt{d+1}C_\lambda L}{2}.
$$

Since $\pi_{\mathbf{x}}\to 1$ as $\sigma\to+\infty$, LIME tends to be Uniform LIME in the limit. The above result just shows that as $\sigma\to+\infty$, the Lipschitz constant of LIME converges to the Lipschitz constant of Uniform LIME, which validates the correctness and compactness of our proof. $\square$

C.4.3. SHAP

Theorem C.11. *For SHAP, if f is $(\delta, L)-$Lipschtz, we have $\forall \mathbf{x}, \mathbf{x}', \|\mathbf{x}-\mathbf{x}'\|_2 \leq \delta$*

$$\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2 \leq 2\sqrt{d}L\|\mathbf{x}-\mathbf{x}'\|_2$$

that is, $\phi(\mathbf{x})$ is $(\delta, 2\sqrt{d}L)-$Lipschitz

Proof. Denote $\phi(\mathbf{x})_i$ as the $i-$th element of $\phi(\mathbf{x})$.

$$\begin{aligned}
|\phi(\mathbf{x})_i - \phi(\mathbf{x}')_i| =& \Big|\sum_{S\subset[d]:i\in S} \frac{(|S|-1)!(d-|S|)!}{d!}\big[(f(\mathbf{x}\odot m_S) - f(\mathbf{x}\odot m_{S\setminus\{i\}})) \\
&\qquad - (f(\mathbf{x}'\odot m_S) - f(\mathbf{x}'\odot m_{S\setminus\{i\}}))\big]\Big| \\
=& \Big|\sum_{S\subset[d]:i\in S} \frac{(|S|-1)!(d-|S|)!}{d!}\big[(f(\mathbf{x}\odot m_S) - f(\mathbf{x}'\odot m_S)) \\
&\qquad - (f(\mathbf{x}\odot m_{S\setminus\{i\}}) - f(\mathbf{x}'\odot m_{S\setminus\{i\}}))\big]\Big| \\
\text{(Triangle Inequality)} \leq& \sum_{S\subset[d]:i\in S} \frac{(|S|-1)!(d-|S|)!}{d!}\big[|f(\mathbf{x}\odot m_S) - f(\mathbf{x}'\odot m_S)| \\
&\qquad + |f(\mathbf{x}\odot m_{S\setminus\{i\}}) - f(\mathbf{x}'\odot m_{S\setminus\{i\}})|\big]\Big| \\
(f \text{ is } (\delta, L-\text{Lipschitz}) \leq& \sum_{S\subset[d]:i\in S} \frac{(|S|-1)!(d-|S|)!}{d!}\big[L\|(\mathbf{x}-\mathbf{x}')\odot m_S\|_2 \\
&\qquad + L\|(\mathbf{x}-\mathbf{x}')\odot m_{S\setminus\{i\}}\|_2\big] \\
\leq& \sum_{S\subset[d]:i\in S} \frac{(|S|-1)!(d-|S|)!}{d!}\big[L\|\mathbf{x}-\mathbf{x}'\|_2 + L\|\mathbf{x}-\mathbf{x}'\|_2\big] \\
=& 2L\|\mathbf{x}-\mathbf{x}'\|_2
\end{aligned}$$

The last inequality is due to the fact that $m_S \in \{0,1\}^n$ and that

$$\|\mathbf{x}\odot m_s\|_2^2 = \sum_{i:i\in S} x_i^2 \leq \sum_{i=1}^d x_i^2 = \|\mathbf{x}\|_2^2.$$

The last equality holds by the following derivation:

$$\sum_{S\subset[d]:i\in S} \frac{(|S|-1)!(d-|S|)!}{d!} = \sum_{k=1}^d \sum_{S:|S\setminus\{i\}|=k} \frac{(k-1)!(d-k)!}{d!} = \sum_{k=1}^d \binom{d-1}{k-1}\frac{(k-1)!(d-k)!}{d!} = \sum_{k=1}^d \frac{1}{d} = 1$$

With bound on $|\phi(\mathbf{x})_i - \phi(\mathbf{x}')_i|, \forall i$, we can easily bound $\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2$

$$\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2 \leq \sqrt{d}\max_i |\phi(\mathbf{x})_i - \phi(\mathbf{x}')_i| = 2\sqrt{d}L\|\mathbf{x}-\mathbf{x}'\|_2$$

□

C.4.4. INTEGRATEDGRADIENT

Lemma C.12. *For two vectors $a, b \in \mathbb{R}^n$, we have*

$$\|a\odot b\|_2 \leq \|a\|_2\|b\|_2$$

Proof.

$$
\begin{aligned}
\|a \odot b\|_2^2 &= \sum_{i=1}^{n} a_i^2 b_i^2 \\
L_2 \text{ norm is bounded by } L_1 \text{ norm } &\leq \Big(\sum_{i=1}^{n} |a_i b_i|\Big)^2 \\
\text{Holder's Inequality } &\leq \Big(\sum_{i=1}^{n} a_i^2\Big)\Big(\sum_{i=1}^{n} b_i^2\Big) \\
&= \|a\|_2^2 \|b\|_2^2
\end{aligned}
$$

□

Lemma C.13 (Paulavičius & Žilinskas (2006) Theorem 1)**.** *If f is $(\delta, L)-$Lipschitz and f is differentiable, then we have*

$$\|\nabla f\|_2 \leq L$$

Theorem C.14. *For IntegratedGradient, assume that $|f(\mathbf{x})| \leq R, \forall x \in \mathcal{X}$, and that f is $(\delta, H)-\text{smooth}$:*

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\|_2 \leq H\|\mathbf{x} - \mathbf{x}'\|_2, \|\mathbf{x} - \mathbf{x}'\|_2 \leq \delta.$$

We have if f is $(\delta, L)-$Lipschitz, then $\phi(\mathbf{x})$ is $(\delta, \frac{\beta H + 2L}{2})-$ Lipschitz.

Proof.

$$
\begin{aligned}
\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\|_2 &= \|\mathbb{E}_{\epsilon \sim U(0,1)}[x \odot \nabla f(\epsilon x) - x' \odot \nabla f(\epsilon x')]\|_2 \\
&= \|\mathbb{E}_{\epsilon \sim U(0,1)}[x \odot \nabla f(\epsilon x) - x \odot \nabla f(\epsilon x') + x \odot \nabla f(\epsilon x') - x' \odot \nabla f(\epsilon x')]\|_2 \\
\text{(Minkowski Inequality) } &\leq \|\mathbb{E}_{\epsilon \sim U(0,1)}[x \odot \nabla f(\epsilon x) - x \odot \nabla f(\epsilon x')]\|_2 \\
&\quad + \|\mathbb{E}_{\epsilon \sim U(0,1)}[x \odot \nabla f(\epsilon x') - x' \odot \nabla f(\epsilon x')]\|_2 \\
\text{(Jensen's Inequality) } &\leq \mathbb{E}_{\epsilon \sim U(0,1)}[\|\mathbf{x} \odot \nabla f(\epsilon x) - x \odot \nabla f(\epsilon x')\|_2] \\
&\quad + \mathbb{E}_{\epsilon \sim U(0,1)}[\|\mathbf{x} \odot \nabla f(\epsilon x') - x' \odot \nabla f(\epsilon x')\|_2] \\
\text{(Lemma C.12) } &\leq \mathbb{E}_{\epsilon \sim U(0,1)}[\|\mathbf{x}\|_2 \|\nabla f(\epsilon x) - \nabla f(\epsilon x')\|_2] \\
&\quad + \mathbb{E}_{\epsilon \sim U(0,1)}[\|\mathbf{x} - \mathbf{x}'\|_2 \|\nabla f(\epsilon x')\|_2] \\
(f \text{ is } (\delta, H)-\text{smooth and Lemma C.13}) &\leq \beta \cdot H \mathbb{E}_{\epsilon \sim U(0,1)}[\epsilon \|\mathbf{x} - \mathbf{x}'\|_2] + L\|\mathbf{x} - \mathbf{x}'\|_2 \\
&= \frac{\beta H + 2L}{2} \|\mathbf{x} - \mathbf{x}'\|_2
\end{aligned}
$$

□

C.5. Gradient

Theorem C.15. *For Gradient, if f is (δ, H)-smooth, then we have $\phi(\mathbf{x})$ is (δ, H)-Lipschitz.*

Proof. Because f is (δ, H)-smooth, we have

$$\|\phi(\mathbf{x}) - \phi(\mathbf{x}')\| = \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{x}')\| \leq H\|\mathbf{x} - \mathbf{x}'\|$$

which means $\phi(\mathbf{x})$ is (δ, H)-Lipschitz. □

C.6. Gradient×Input

Theorem C.16. *For Gradient×Input, if f is (δ, L)-Lipschitz and (δ, H)-smooth, then we have $\phi(\mathbf{x})$ is $(\delta, \beta H + L)$-Lipschitz.*

Proof.

$$\begin{aligned}
\|\phi(\mathbf{x})-\phi(\mathbf{x}')\| =&\|\nabla f(\mathbf{x})\odot\mathbf{x}-\nabla f(\mathbf{x}')\odot\mathbf{x}'\| \\
=&\|\nabla f(\mathbf{x})\odot\mathbf{x}-|\nabla f(\mathbf{x})\odot\mathbf{x}'+|\nabla f(\mathbf{x})\odot\mathbf{x}-\nabla f(\mathbf{x}')\odot\mathbf{x}'\| \\
\leq&\|\nabla f(\mathbf{x})\odot\mathbf{x}-|\nabla f(\mathbf{x})\odot\mathbf{x}'\|+\|\nabla f(\mathbf{x})\odot\mathbf{x}'-\nabla f(\mathbf{x}')\odot\mathbf{x}'\| \\
=&\|\nabla f(\mathbf{x})\|\|\mathbf{x}-\mathbf{x}'\|+\|\nabla f(\mathbf{x})-\nabla f(\mathbf{x}')\|\|\mathbf{x}'\| \\
\leq&\|\nabla f(\mathbf{x})\|\|\mathbf{x}-\mathbf{x}'\|+\beta H\|\mathbf{x}-\mathbf{x}'\|
\end{aligned}$$

By Lemma C.13

$$\|\nabla f(\mathbf{x})\|\leq L$$

therefore, we have

$$\begin{aligned}
\|\phi(\mathbf{x})-\phi(\mathbf{x}')\| \leq&\|\nabla f(\mathbf{x})\|\|\mathbf{x}-\mathbf{x}'\|+\beta H\|\mathbf{x}-\mathbf{x}'\| \\
\leq&(\beta H+L)\|\mathbf{x}-\mathbf{x}'\|
\end{aligned}$$

□