Stability Guarantees for Feature Attributions with Multiplicative Smoothing

Anton Xue **Rajeev Alur** **Eric Wong**
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104
{antonxue,alur,exwong}@seas.upenn.edu

Abstract

Explanation methods for machine learning models tend not to provide any formal guarantees and may not reflect the underlying decision-making process. In this work, we analyze stability as a property for reliable feature attribution methods. We prove that relaxed variants of stability are guaranteed if the model is sufficiently Lipschitz with respect to the masking of features. We develop a smoothing method called Multiplicative Smoothing (MuS) to achieve such a model. We show that MuS overcomes the theoretical limitations of standard smoothing techniques and can be integrated with any classifier and feature attribution method. We evaluate MuS on vision and language models with various feature attribution methods, such as LIME and SHAP, and demonstrate that MuS endows feature attributions with non-trivial stability guarantees.

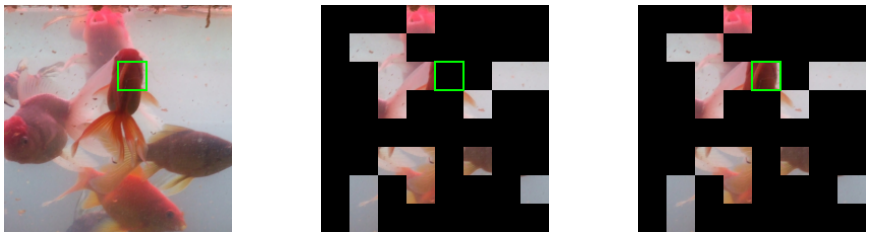
1 Introduction

Modern machine learning models are incredibly powerful at challenging prediction tasks but notoriously black-box in their decision-making. One can therefore achieve impressive performance without fully understanding *why*. In settings like medical diagnosis [1, 2] and legal analysis [3, 4] where accurate and well-justified decisions are important, however, such power without proof is insufficient. In order to fully wield the power of such models while ensuring reliability and trust, a user needs accurate and insightful *explanations* of model behavior.

One popular family of explanation methods is *feature attributions* [5, 6, 7, 8]. Given a model and input, a feature attribution method generates a score for each input feature that denotes its importance to the overall prediction. For instance, consider Figure 1, in which the Vision Transformer [9] classifier predicts the full image (left) as "Goldfish". We then use a feature attribution method like SHAP [7] to score each feature and select the top-25%, for which the masked image (middle) is consistently predicted as "Goldfish". However, including a single patch of features (right) alters the prediction confidence so much that it now yields "Axolotl". This suggests that the explanation is brittle [10], as small changes easily cause it to induce some other class. In this paper, we study how to overcome such behavior by analyzing the *stability* of an explanation: an explanation is *stable* if adding more features does not change the prediction once the explanatory features are included.

Stability implies that the selected features are enough to explain the prediction [11, 12, 13] and that this selection maintains strong explanatory power even in the presence of additional information [10, 14]. Similar properties are studied in literature and identified as useful for interpretability [15], and we emphasize that our main focus is on analyzing and achieving provable guarantees. Stability guarantees, in particular, are useful as they allow one to accurately predict how model behavior varies with the explanation. Given a stable explanation, one can include more features, i.e., adding context and information, while maintaining confidence in the consistency of the underlying explanatory

XAI in Action: Past, Present, and Future Applications @ NeurIPS 2023.



Goldfish (98.72%)
Cockroach (0.0218%)

Goldfish (95.86%)
Axolotl (1.62%)

Axolotl (55.14%)
Goldfish (30.24%)

Figure 1: Classification by VisionTransformer [9] on an attribution generated by SHAP [7] with top-25% selection. A single 28×28 pixel patch of difference between the two attributions (marked **green**) significantly affects prediction confidence and results in a classification flip.

power. Crucially, we observe that such guarantees only make sense when jointly considering the model and explanation method: the explanation method necessarily depends on the model to yield an explanation, and stability is then evaluated with respect to the model.

Thus far, existing work on feature attributions with formal guarantees faces computational tractability and explanatory utility challenges. While some methods take an axiomatic approach [8, 16], others use metrics that appear reasonable but may not reliably reflect useful model behavior, a common and known limitation [17]. Such explanations have been criticized as, at best, a plausible guess and, at worst, completely misleading [18].

In this paper, we study how to construct explainable models with provable stability guarantees. We jointly consider the classification model and explanation method and present a formalization for studying such properties that we call *explainable models*. We focus on *binary feature attributions* [19] wherein each feature is either marked as explanatory (1) or not explanatory (0). We present a method to solve this problem, which is inspired by techniques from adversarial robustness, in particular randomized smoothing [20, 21]. Our method can take *any* off-the-shelf classifier and feature attribution method to efficiently yield an explainable model that satisfies provable stability guarantees. In summary, our contributions are as follows:

- We formalize stability as a key property for binary feature attributions and study this in the framework of explainable models. We prove that relaxed variants of stability are guaranteed if the model is sufficiently Lipschitz with respect to the masking of features.
- To achieve the sufficient Lipschitz condition, we develop a smoothing method called Multiplicative Smoothing (MuS). We show that MuS achieves strong smoothness conditions, overcomes key theoretical and practical limitations of standard smoothing techniques, and can be integrated with any classifier and feature attribution method.
- We evaluate MuS on vision and language models along with different feature attribution methods. We demonstrate that MuS-smoothed explainable models achieve strong stability guarantees at a small cost to accuracy.

2 Overview

We observe that formal guarantees for explanations must consider both the model and explanation method. For this, we present in Section 2.1 a pairing that we call *explainable models*. This formulation allows us to describe the desired stability properties in Section 2.2. We show in Section 2.3 that classifiers with sufficient Lipschitz smoothness with respect to feature masking allow us to yield provable stability guarantees.

2.1 Explainable Models

We first present explainable models as a formalism for rigorously studying explanations. Let $\mathcal{X} = \mathbb{R}^n$ be the space of inputs, a classifier $f : \mathcal{X} \to [0, 1]^m$ maps inputs $x \in \mathcal{X}$ to m class probability values that sum to 1, where the class of $f(x) \in [0, 1]^m$ is taken to be the largest coordinate. Similarly,

an explanation method $\varphi : \mathcal{X} \to \{0,1\}^n$ maps an input $x \in \mathcal{X}$ to an explanation $\varphi(x) \in \{0,1\}^n$ that indicates which features are considered explanatory for the prediction $f(x)$. In particular, we may pick and adapt φ from among a selection of existing feature attribution methods like LIME [6], SHAP [7], and many others [5, 8, 22, 23, 24], wherein φ may be thought of as a top-k feature selector. Note that the selection of input features necessarily depends on the explanation method executing or analyzing the model, and so it makes sense to jointly study the model and explanation method: given a classifier f and explanation method φ, we call the pairing $\langle f, \varphi \rangle$ an *explainable model*. Given some $x \in \mathcal{X}$, the explainable model $\langle f, \varphi \rangle$ maps x to both a prediction and explanation. We show this in Figure 2, where $\langle f, \varphi \rangle(x) \in [0,1]^m \times \{0,1\}^n$ pairs the class probabilities and the feature attribution.

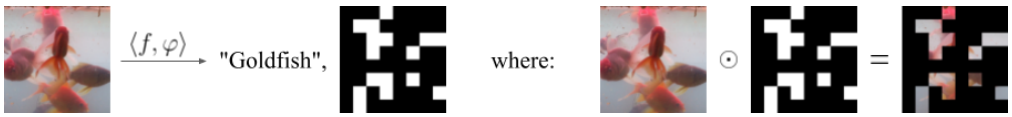


Figure 2: An explainable model $\langle f, \varphi \rangle$ outputs both a classification and a feature attribution. The feature attribution is a binary-valued mask (white 1, black 0) that can be applied over the original input. Here f is Vision Transformer [9] and φ is SHAP [7] with top-25% feature selection.

For an input $x \in \mathcal{X}$, we will evaluate the quality of the binary feature attribution $\varphi(x)$ through its masking on x. That is, we will study the behavior of f on the masked input $x \odot \varphi(x) \in \mathcal{X}$, where $\odot$ is the element-wise vector product. To do this, we define a notion of *prediction equivalence*: for two $x, x' \in \mathcal{X}$, we write $f(x) \cong f(x')$ to mean that $f(x)$ and $f(x')$ yield the same class. This allows us to formalize the intuition that an explanation $\varphi(x)$ should recover the prediction of x under f.

Definition 2.1. *The explainable model $\langle f, \varphi \rangle$ is consistent at x if $f(x) \cong f(x \odot \varphi(x))$.*

Evaluating f on $x \odot \varphi(x)$ this way lets us apply the model as-is and therefore avoids the challenge of constructing a surrogate model that is accurate to the original [25]. Moreover, this approach is popular in domains like vision — where one intuitively expects that a masked image retaining only the important features should induce the intended prediction. Indeed, architectures like Vision Transformer [9] can maintain high accuracy with only a fraction of the image present [26].

Particularly, we would like for $\langle f, \varphi \rangle$ to generate explanations that are stable and concise (i.e. sparse). The former is our central guarantee and is ensured through smoothing. The latter implies that $\varphi(x)$ has few ones entries, and is desirable since a good explanation should not contain too much redundant information. However, sparsity is more difficult to enforce, as this is contingent on the model having high accuracy with respect to heavily masked inputs.

2.2 Stability Properties of Explainable Models

Given an explainable model $\langle f, \varphi \rangle$ and some $x \in \mathcal{X}$, stability means that the prediction does not change even if one adds more explanatory features to $\varphi(x)$. For instance, the model-explanation pair in Figure 1 is *not* stable, as the inclusion of a single feature group (patch) changes the prediction. To formalize this notion of stability, we first introduce a partial ordering: for $\alpha, \alpha' \in \{0,1\}^n$, we write $\alpha \succeq \alpha'$ iff $\alpha_i \geq \alpha'_i$ for all $i = 1, \ldots, n$. That is, $\alpha \succeq \alpha'$ iff α includes all the features selected by α'.

Definition 2.2. *The explainable model $\langle f, \varphi \rangle$ is stable at x if $f(x \odot \alpha) \cong f(x \odot \varphi(x))$ for all $\alpha \succeq \varphi(x)$.*

Note that the constant explanation $\varphi(x) = \mathbf{1}$, the vector of ones, makes $\langle f, \varphi \rangle$ trivially stable at every $x \in \mathcal{X}$, though this is not a concise explanation. Additionally, stability at x implies consistency at x.

Unfortunately, stability is a difficult property to enforce in general, as it requires that f satisfy a monotone-like behavior with respect to feature inclusion — which is especially challenging for complex models like neural networks. Checking stability without additional assumptions on f is also hard: if $k = \|\varphi(x)\|_1$ is the number of ones in $\varphi(x)$, then there are 2^{n-k} possible $\alpha \succeq \varphi(x)$ to check. This large space of possible $\alpha \succeq \varphi(x)$ motivates us to examine instead *relaxations* of stability. We introduce lower and upper relaxations of stability below.

Definition 2.3. *The explainable model $\langle f, \varphi \rangle$ is incrementally stable at x with radius r if $f(x \odot \alpha) \cong f(x \odot \varphi(x))$ for all $\alpha \succeq \varphi(x)$ where $\|\alpha - \varphi(x)\|_1 \leq r$.*

Incremental stability is the lower relaxation since it considers the case where the mask α has only a few features more than $\varphi(x)$. For instance, if one can provably add up to r features to a masked $x \odot \varphi(x)$ without altering the prediction, then $\langle f, \varphi \rangle$ would be incrementally stable at x with radius r. We next introduce the upper relaxation that we call decremental stability.

Definition 2.4. *The explainable model $\langle f, \varphi \rangle$ is decrementally stable at x with radius r if $f(x \odot \alpha) \cong f(x \odot \varphi(x))$ for all $\alpha \succeq \varphi(x)$ where $\|\mathbf{1} - \alpha\|_1 \leq r$.*

Decremental stability is a subtractive form of stability in contrast to the additive nature of incremental stability. Particularly, decremental stability considers the case where α has much more features than $\varphi(x)$. If one can provably remove up to r non-explanatory features from the full x without altering the prediction, then $\langle f, \varphi \rangle$ is decrementally stable at x with radius r. Note that decremental stability necessarily entails consistency of $\langle f, \varphi \rangle$, but for simplicity of definitions, we do not enforce this for incremental stability. Finally, note that for sufficiently large radius of $r = \lceil (n - \|\varphi(x)\|_1)/2 \rceil$, incremental and decremental stability together imply stability.

Remark 2.5. *Similar notions to the above have been proposed in the literature, and we refer to [15] for an extensive survey. In particular, for [15], consistency is akin to* preservation*, and stability is similar to* continuity*, except we are concerned with adding features. Also, incremental stability is most similar to* incremental addition *and decremental stability to* incremental deletion*.*

2.3 Lipschitz Smoothness Entails Stability Guarantees

If $f : \mathcal{X} \to [0,1]^m$ is Lipschitz with respect to the masking of features, then we can guarantee relaxed stability properties for the explainable model $\langle f, \varphi \rangle$. In particular, we require for all $x \in \mathcal{X}$ that $f(x \odot \alpha)$ is Lipschitz with respect to the mask $\alpha \in \{0,1\}^n$. This allows us to present our main results in smoothness and stability, which we formalize in Section 3.1. A sketch of the stability result is first given below in Remark 2.6.

Remark 2.6 (Sketch of main result)**.** *Consider an explainable model $\langle f, \varphi \rangle$ where for all $x \in \mathcal{X}$ the function $g(x, \alpha) = f(x \odot \alpha)$ is λ-Lipschitz in $\alpha \in \{0,1\}^n$ with respect to the ℓ^1 norm. Then at any x, the radius of incremental stability r_{inc} and radius of decremental stability r_{dec} are respectively:*

$$r_{\text{inc}} = \big[g_A(x, \varphi(x)) - g_B(x, \varphi(x))\big]/(2\lambda), \qquad r_{\text{dec}} = \big[g_A(x, \mathbf{1}) - g_B(x, \mathbf{1})\big]/(2\lambda),$$

where $g_A - g_B$ is called the confidence gap, with g_A, g_B the top-two class probabilities:

$$g_A(x, \alpha) = g_{k^\star}(x, \alpha), \qquad g_B(x, \alpha) = \max_{i \neq k^\star} g_i(x, \alpha), \qquad k^\star = \operatorname*{argmax}_{1 \leq k \leq m} g_k(x, \alpha) \tag{1}$$

Observe that Lipschitz smoothness is a stronger assumption than necessary, as besides $\alpha \succeq \varphi(x)$, it also imposes guarantees on $\alpha \preceq \varphi(x)$. Nevertheless, Lipschitz smoothness is one of the few properties that can be guaranteed and analyzed at scale on arbitrary models [21, 27]. Importantly, we may apriori pick the Lipschitz constant λ for our smoothed classifier, allowing us to establish known guarantees before test time. The details for establishing the Lipschitz constant through our randomized smoothing method are described in Theorem 3.1.

3 Multiplicative Smoothing for Lipschitz Constants

In this section we present our main technical contribution in Multiplicative Smoothing (MuS). The goal is to transform an arbitrary base classifier $h : \mathcal{X} \to [0,1]^m$ into a smoothed classifier $f : \mathcal{X} \to [0,1]^m$ that is Lipschitz with respect to the masking of features. This then allows one to appropriately couple an explanation method φ with f to form an explainable model $\langle f, \varphi \rangle$ with provable stability guarantees. Appendix A gives an extended discussion of results.

3.1 Technical Overview of MuS

Our key insight is that randomly dropping (i.e., zeroing) features attains the desired smoothness. In particular, we uniformly drop features with probability $1 - \lambda$ by sampling binary masks $s \in \{0,1\}^n$ from some distribution $\mathcal{D}$ where each coordinate is distributed as $\Pr[s_i = 1] = \lambda$. Then define:

$$f(x) = \mathop{\mathbb{E}}_{s \sim \mathcal{D}} h(x \odot s), \qquad \text{such that } s_i \sim \mathcal{B}(\lambda) \text{ for } i = 1, \ldots, n, \tag{2}$$

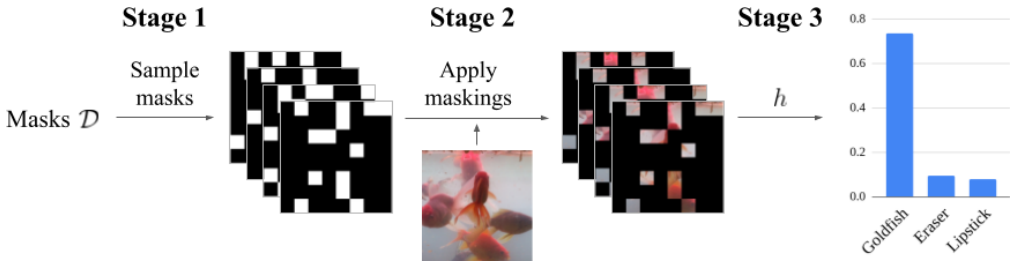


Figure 3: Evaluating $f(x)$ is done in three stages. **(Stage 1)** Generate N samples of binary masks $s^{(1)}, \ldots, s^{(N)} \in \{0,1\}^n$, where each coordinate is Bernoulli with parameter λ (here $\lambda = 1/4$). **(Stage 2)** Apply each mask on the input to yield $x \odot s^{(i)}$ for $i = 1, \ldots, N$. **(Stage 3)** Average over $h(x \odot s^{(i)})$ to compute $f(x)$, and note that the predicted class is given by a weighted average.

where $\mathcal{B}(\lambda)$ is the Beronulli distribution with parameter $\lambda \in [0,1]$. We give an overview of evaluating $f(x)$ in Figure 3. Importantly, our main results of smoothness (Theorem 3.1) and stability (Theorem 3.2) hold provided each coordinate of $\mathcal{D}$ is marginally Bernoulli with parameter λ, and so we avoid fixing a particular choice for now. However, it will be easy to intuit the exposition with $\mathcal{D} = \mathcal{B}^n(\lambda)$, the coordinate-wise i.i.d. Bernoulli distribution with parameter λ.

We can equivalently parametrize f using the mapping $g(x, \alpha) = f(x \odot \alpha)$, where it follows that:

$$g(x, \alpha) = \mathop{\mathbb{E}}_{s \sim \mathcal{D}} h(x \odot \tilde{\alpha}), \qquad \tilde{\alpha} = \alpha \odot s. \tag{3}$$

Note that one could have alternatively first defined g and then f due to the identity $g(x, \mathbf{1}) = f(x)$. We require that the relationship between f and g follows an identity that we call *masking equivalence*: $g(x \odot \alpha, \mathbf{1}) = f(x \odot \alpha) = g(x, \alpha)$ for all $x \in \mathcal{X}$ and $\alpha \in \{0,1\}^n$. This follows by the definition of g, and the relevance to stability is this: if masking equivalence holds, then we can rewrite stability properties involving f in terms of g's second parameter as follows:

$$f(x \odot \alpha) = g(x, \alpha) \cong g(x, \varphi(x)) = f(x \odot \varphi(x)) \qquad \text{for all } \alpha \succeq \varphi(x), \qquad \text{(c.f. Definition 2.2)}$$

where incremental and decremental stability may be analogously defined. This translation is useful, as we will prove that g is λ-Lipschitz in its second parameter (Theorem 3.1), which then allows us to establish the desired stability properties (Theorem 3.2). Importantly, we are motivated to develop MuS because standard smoothing techniques, namely additive smoothing [20, 21], may fail to satisfy masking equivalence. This is further explained in Section A.1.

3.2 Certifying Stability with Lipschitz Classifiers

Our core technical result shows that f as defined in (2) is Lipschitz to the masking of features. We present MuS in terms of g, where it is parametric with respect to the distribution $\mathcal{D}$. In particular, $\mathcal{D}$ is usable with MuS so long as it satisfies a coordinate-wise Bernoulli condition.

Theorem 3.1 (MuS). *Let $\mathcal{D}$ be any distribution on $\{0,1\}^n$ where each coordinates of $s \sim \mathcal{D}$ is distributed as $s_i \sim \mathcal{B}(\lambda)$. Consider any $h : \mathcal{X} \to [0,1]$ and define $g : \mathcal{X} \times \{0,1\}^n \to [0,1]$ as*

$$g(x, \alpha) = \mathop{\mathbb{E}}_{s \sim \mathcal{D}} h(x \odot \tilde{\alpha}), \qquad \tilde{\alpha} = \alpha \odot s.$$

Then the function $g(x, \cdot) : \{0,1\}^n \to [0,1]$ is λ-Lipschitz in the ℓ^1 norm for all $x \in \mathcal{X}$.

The strength of this result is in its weak assumptions. First, the theorem applies to any model h and input $x \in \mathcal{X}$. It further suffices that each coordinate is distributed as $s_i \sim \mathcal{B}(\lambda)$, and we emphasize that statistical independence between different s_i, s_j is *not assumed*. This allows us to construct $\mathcal{D}$ with structured dependence in Section A.2, such that we may exactly and efficiently evaluate $g(x, \alpha)$ at a sample complexity of $N \ll 2^n$. A low sample complexity is important for the practicality of MuS, as otherwise, one must settle for the expected value subject to probabilistic guarantees. For instance, simpler distributions like $\mathcal{B}^n(\lambda)$ do satisfy the requirements of Theorem 3.1 — but may cost 2^n samples because of coordinate-wise independence. Whatever choice of valid $\mathcal{D}$, one can guarantee stability so long as g is Lipschitz in its second argument.

Theorem 3.2 (Stability). *Consider any $h : \mathcal{X} \to [0,1]^m$ with coordinates $h_1, \ldots, h_m$. Fix $\lambda \in [0,1]$ and let $g_1, \ldots, g_m$ be the respectively smoothed coordinates as in Theorem 3.1, using which we analogously define $g : \mathcal{X} \times \{0,1\}^n \to [0,1]^m$. Also define $f(x) = g(x, \mathbf{1})$. Then for any explanation method φ and input $x \in \mathcal{X}$, the explainable model $\langle f, \varphi \rangle$ is incrementally stable with radius r_{inc} and decrementally stable with radius r_{dec}:*

$$r_{\text{inc}} = \big[g_A(x, \varphi(x)) - g_B(x, \varphi(x))\big]/(2\lambda), \qquad r_{\text{dec}} = \big[g_A(x, \mathbf{1}) - g_B(x, \mathbf{1})\big]/(2\lambda),$$

where g_A, g_B are the first and second largest class probability values as in (1).

Note that non-trivial stability guarantees exist only in the case where the radius ≥ 1. As each g_k has range $[0,1]$, one needs $\lambda \leq 1/2$ for non-trivial guarantees.

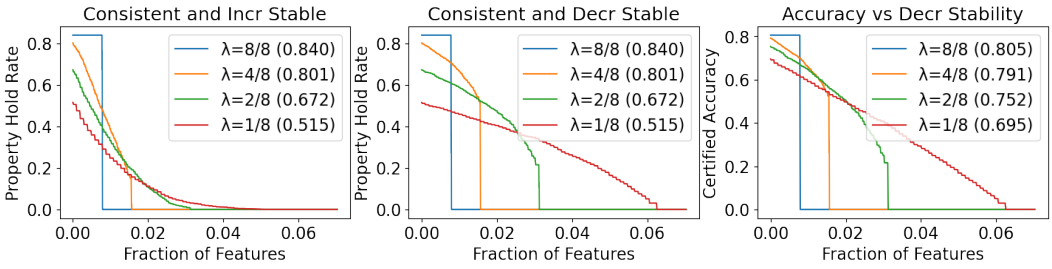


Figure 4: **(Left; Middle)** Consistency and incremental (resp. decremental) stability. *Property hold rate* is the fraction of images that are consistent and stable up to radius r when using a mask from SHAP-top25%. Recall that every $\langle f, \varphi \rangle$ is trivially stable at radius $r = 0$. **(Right)** Overall accuracy vs the radius of decremental stability. *Certified accuracy* is the fraction of images for which f predicts the true label on the entire unmasked x while achieving decremental stability at radius r.

4 Empirical Evaluations

(Experimental Setup) We highlight a subset of our results and refer to our extended manuscript [28] for comprehensive experiments. In this section, we show results with Vision Transformer [9] and ImageNet1K [29]. We group features on the $3 \times 224 \times 224$ dimensional input into $n = 64$ superpixels, and report stability radii r as a fraction of the features, i.e. r/n. For methods, we use SHAP [7] with top-25% feature selection. All experiments here use ImageNet1K with a sample size of $N = 2000$.

4.1 (E1) How Good are the Stability Guarantees?

We measure the radius to which one can certify incremental and incremental stability. To do this, we measure the rate at which stability and consistency holds at some radius r (expressed as r/n). We show our results in Figure 4, where we show consistent and incremental stability (left) and consistent and decremental stability (middle).

4.2 (E2) What is the Cost of Smoothing?

To increase the radius of provable stability, we decrease λ. However, this λ decrease means that fewer features are seen in the smoothing process. To study the stability-accuracy trade-off, we plotted the accuracy attained by the smoothed classifier vs. the radius of decremental stability and show the results in Figure 4 (right), where as expected the clean accuracy (in parentheses) decreases with λ. For Vision Transformer we see that the accuracy remains high even under non-trivial noise.

5 Conclusion

We study provable stability guarantees for binary feature attribution methods through the framework of explainable models. A selection of features is stable if the additional inclusion of other features does not alter its explanatory power. We show that if the classifier is Lipschitz with respect to the masking of features, then one can guarantee relaxed variants of stability. To achieve this Lipschitz condition we develop a smoothing method called Multiplicative Smoothing (MuS). We show that MuS yields strong stability guarantees at only a small cost to accuracy.

References

[1] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3):e190043, 2020.

[2] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE transactions on neural networks and learning systems*, 32(11):4793–4813, 2020.

[3] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[4] Adrien Bibal, Michael Lognoul, Alexandre De Streel, and Benoît Frénay. Legal requirements on explainability in machine learning. *Artificial Intelligence and Law*, 29:149–169, 2021.

[5] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[6] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[7] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[8] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019.

[11] Gavin Brown. A new perspective for information theoretic feature selection. In *Artificial intelligence and statistics*, pages 49–56. PMLR, 2009.

[12] Jianbo Chen, Le Song, Martin Wainwright, and Michael Jordan. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892. PMLR, 2018.

[13] Xiaoping Li, Yadi Wang, and Rubén Ruiz. A survey on sparse learning models for feature selection. *IEEE transactions on cybernetics*, 52(3):1642–1660, 2020.

[14] Akhilan Boopathy, Sijia Liu, Gaoyuan Zhang, Cynthia Liu, Pin-Yu Chen, Shiyu Chang, and Luca Daniel. Proper network interpretability helps adversarial robustness in classification. In *International Conference on Machine Learning*, pages 1014–1023. PMLR, 2020.

[15] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *arXiv preprint arXiv:2201.08164*, 2022.

[16] L.S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, pages 307–317, 1953.

[17] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9623–9633, 2022.

[18] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, 2020.

[19] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.

[20] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[21] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR, 2020.

[22] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[23] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *International conference on machine learning*, pages 9269–9278. PMLR, 2020.

[24] Yongchan Kwon and James Zou. Weightedshap: analyzing and improving shapley based feature attributions. *arXiv preprint arXiv:2209.13429*, 2022.

[25] Reza Alizadeh, Janet K Allen, and Farrokh Mistree. Managing computational complexity using surrogate models: a critical review. *Research in Engineering Design*, 31:275–298, 2020.

[26] Hadi Salman, Saachi Jain, Eric Wong, and Aleksander Madry. Certified patch robustness via smoothed vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15137–15147, 2022.

[27] Alexander J Levine and Soheil Feizi. Improved, deterministic smoothing for l_1 certified robustness. In *International Conference on Machine Learning*, pages 6254–6264. PMLR, 2021.

[28] Anton Xue, Rajeev Alur, and Eric Wong. Stability guarantees for feature attributions with multiplicative smoothing. *arXiv preprint arXiv:2307.05902*, 2023.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

A Extended Results for Section 3

We give an extended discussion of content from Section 3.

A.1 Standard Smoothing Does Not Satisfy Masking Equivalence

We are motivated to develop MuS because standard smoothing techniques, namely additive smoothing [20, 21], may fail to satisfy masking equivalence. Additive smoothing is by far the most popular smoothing technique, and differs from our scheme (3) in how noise is applied, where let $\mathcal{D}_{\text{add}}$ and $\mathcal{D}_{\text{mult}}$ be any two distributions on $\mathbb{R}^n$:

$$g(x, \alpha) = \mathop{\mathbb{E}}_{s \sim \mathcal{D}} h(x \odot \tilde{\alpha}), \qquad \tilde{\alpha} = \begin{cases} \alpha + s, & s \sim \mathcal{D}_{\text{add}} \\ \alpha \odot s, & s \sim \mathcal{D}_{\text{mult}} \end{cases}$$

where $\mathcal{D}_{\text{add}}$ denotes additive smoothing, and $\mathcal{D}_{\text{mult}}$ denotes multiplicative smoothing. Particularly, additive smoothing has counterexamples to masking equivalence.

Proposition A.1. *There exists $h : \mathcal{X} \to [0, 1]$ and distribution $\mathcal{D}$, where for*

$$g^+(x, \alpha) = \mathop{\mathbb{E}}_{s \sim \mathcal{D}} h(x \odot \tilde{\alpha}), \qquad \tilde{\alpha} = \alpha + s,$$

we have $g^+(x, \alpha) \neq g^+(x \odot \alpha, \mathbf{1})$ for some $x \in \mathcal{X}$ and $\alpha \in \{0, 1\}^n$.

Proof. Observe that it suffices to have h, x, α such that $h(x \odot (\alpha + s)) > h((x \odot \alpha) \odot (\mathbf{1} + s))$ for a non-empty set of $s \in \mathbb{R}^n$. Let $\mathcal{D}$ be a distribution on these s, then:

$$g^+(x, \alpha) = \mathop{\mathbb{E}}_{s \sim \mathcal{D}} h(x \odot (\alpha + s)) > \mathop{\mathbb{E}}_{s \sim \mathcal{D}} h((x \odot \alpha) \odot (\mathbf{1} + s)) = g^+(x \odot \alpha, \mathbf{1})$$

□

Intuitively, this occurs because additive smoothing primarily applies noise by perturbing feature values, rather than completely masking them. As such, there might be "information leakage" when non-explanatory bits of α are changed into non-zero values. This then causes each sample of $h(x \odot \tilde{\alpha})$ within $g(x, \alpha)$ to observe more features of x than it would have been able to otherwise.

A.2 Exploiting Structured Dependency

We now present $\mathcal{L}_{qv}(\lambda)$, a distribution on $\{0, 1\}^n$ that allows for efficient and exact evaluation of a MuS-smoothed classifier. Our construction is an adaption of [27] from uniform to Bernoulli noise, where the primary insight is that one can parametrize n-dimensional noise using a single dimension via structured coordinate-wise dependence. In particular, we use a *seed vector* v, where with an integer *quantization parameter* $q > 1$ there will only exist q distinct choices of $s \sim \mathcal{L}_{qv}(\lambda)$. All the while, we still enforce that any such s is coordinate-wise Bernoulli with $s_i \sim \mathcal{B}(\lambda)$. Thus for a sufficiently small quantization parameter (i.e. $q \ll 2^n$) we may tractably enumerate through all q possible choices of s and thereby evaluate a MuS-smoothed model with only q samples.

Proposition A.2. *Fix integer $q > 1$ and consider any vector $v \in \{0, 1/q, \ldots, (q-1)/q\}^n$ and scalar $\lambda \in \{1/q, \ldots, q/q\}$. Define $s \sim \mathcal{L}_{qv}(\lambda)$ to be a random vector in $\{0, 1\}^n$ with coordinates given by*

$$s_i = \mathbb{I}[t_i \leq \lambda], \qquad t_i = v_i + s_{\text{base}} \mod 1,$$

where $s_{\text{base}} \sim \mathcal{U}(\{1/q, \ldots, q/q\}) - 1/(2q)$. Then there are q distinct values of s and each coordinate is distributed as $s_i \sim \mathcal{B}(\lambda)$.

Proof. First, observe that each of the q distinct values of s_{base} defines a unique value of s, since we have assumed v and λ to be fixed. Next, observe that each t_i has q unique values uniformly distributed as $t_i \sim \mathcal{U}(1/q, \ldots, q/q\}) - 1/(2q)$. Because $\lambda \in \{1/q, \ldots, q/q\}$ we therefore have $\Pr[t_i \leq \lambda] = \lambda$, which implies that $s_i \sim \mathcal{B}(\lambda)$. □

The seed vector v is the source of our structured coordinate-wise dependence and the one-dimensional source of randomness s_{base} is used to generate the n-dimensional s. Such $s \sim \mathcal{L}_{qv}(\lambda)$ then satisfies the conditions for use in MuS (Theorem 3.1), and this noise allows for an exact evaluation of the smoothed classifier in q samples. We have found $q = 64$ to be sufficient in practice and values as low as $q = 16$ to also yield good performance. We remark that one drawback is that one may get an unlucky seed v, but we have not yet observed this in our experiments.

B Proofs and Extensions

Here we present the proofs of our main results, as well as some extensions to MuS.

B.1 Proof of Theorem 3.1

Proof. By linearity we have:

$$g(x,\alpha) - g(x,\alpha') = \mathop{\mathbb{E}}_{s\sim\mathcal{D}} h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}'), \qquad \tilde{\alpha} = \alpha\odot s, \qquad \tilde{\alpha}' = \alpha'\odot s,$$

so it suffices to analyze an arbitrary term by fixing some $s\sim\mathcal{D}$. Consider any $x\in\mathcal{X}$, let $\alpha,\alpha'\in\{0,1\}^n$, and define $\delta = \alpha - \alpha'$. Observe that $\tilde{\alpha}_i \neq \tilde{\alpha}'_i$ exactly when $|\delta_i| = 1$ and $s_i = 1$. Since $s_i\sim\mathcal{B}(\lambda)$, we thus have $\Pr[\tilde{\alpha}_i\neq\tilde{\alpha}'_i] = \lambda|\delta_i|$, and applying the union bound:

$$\Pr_{s\sim\mathcal{D}}[\tilde{\alpha}\neq\tilde{\alpha}'] = \Pr_{s\sim\mathcal{D}}\left[\bigcup_{i=1}^{n}\tilde{\alpha}_i\neq\tilde{\alpha}'_i\right] \leq \sum_{i=1}^{n}\lambda|\delta_i| = \lambda\|\delta\|_1,$$

and so:

$$\begin{aligned}
|g(x,\alpha) - g(x,\alpha')| &= \left|\mathop{\mathbb{E}}_{s\sim\mathcal{D}}[h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}')]\right| \\
&= \Big|\Pr_{s\sim\mathcal{D}}[\tilde{\alpha}\neq\tilde{\alpha}']\cdot\mathop{\mathbb{E}}_{s\sim\mathcal{D}}[h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}')\,|\,\tilde{\alpha}\neq\tilde{\alpha}'] \\
&\qquad - \Pr_{s\sim\mathcal{D}}[\tilde{\alpha}=\tilde{\alpha}']\cdot\mathop{\mathbb{E}}_{s\sim\mathcal{D}}[h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}')\,|\,\tilde{\alpha}=\tilde{\alpha}']\Big|.
\end{aligned}$$

Note that $\mathbb{E}[h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}')\,|\,\tilde{\alpha}=\tilde{\alpha}'] = 0$, and so

$$\begin{aligned}
|g(x,\alpha) - g(x,\alpha')| &= \Pr_{s\sim\mathcal{D}}[\tilde{\alpha}\neq\tilde{\alpha}']\cdot\underbrace{\left|\mathop{\mathbb{E}}_{s\sim\mathcal{D}}[h(x\odot\tilde{\alpha}) - h(x\odot\tilde{\alpha}')\,|\,\tilde{\alpha}\neq\tilde{\alpha}']\right|}_{\leq 1 \text{ because } h(\cdot)\in[0,1]} \\
&\leq \Pr_{s\sim\mathcal{D}}[\tilde{\alpha}\neq\tilde{\alpha}'] \leq \lambda\|\delta\|_1.
\end{aligned}$$

Thus, $g(x,\cdot)$ is λ-Lipschitz in the ℓ^1 norm. □

B.2 Proof of Theorem 3.2

Proof. We first show incremental stability. Consider any $x\in\mathcal{X}$, then by masking equivalence:

$$f(x\odot\varphi(x)) = g(x\odot\varphi(x),\mathbf{1}) = g(x,\varphi(x)),$$

and let g_A, g_B be the top two class probabilities of g as defined in (1). By Theorem 3.1, both g_A, g_B are Lipschitz in their second parameter, and so for all $\alpha\in\{0,1\}^n$:

$$\begin{aligned}
\|g_A(x,\varphi(x)) - g_A(x,\alpha)\|_1 &\leq \lambda\|\varphi(x)-\alpha\|_1 \\
\|g_B(x,\varphi(x)) - g_B(x,\alpha)\|_1 &\leq \lambda\|\varphi(x)-\alpha\|_1
\end{aligned}$$

Observe that if α is sufficiently close to $\varphi(x)$, i.e.:

$$2\lambda\|\varphi(x)-\alpha\|_1 \leq g_A(x,\varphi(x)) - g_B(x,\varphi(x)),$$

then the top class probability index of $g(x,\varphi(x)$ and $g(x,\alpha)$ are the same. This means that $g(x,\varphi(x)) \cong g(x,\alpha)$ and thus $f(x\odot\varphi(x)) \cong f(x\odot\alpha)$, thus proving incremental stability with radius $d(x,\varphi(x))/(2\lambda)$.

The decremental stability case is similar, except we replace $\varphi(x)$ with $\mathbf{1}$. □

B.3 Feature Grouping

We have so far assumed that $\mathcal{X} = \mathbb{R}^n$, but sometimes it may be desirable to group features together, e.g. color channels of the same pixel. Our results also hold for more general $\mathcal{X} = \mathbb{R}^{d_1}\times\cdots\times\mathbb{R}^{d_n}$, where for such $x\in\mathcal{X}$ and $\alpha\in\mathbb{R}^n$ we lift $\odot$ as:

$$\odot : \mathcal{X}\times\mathbb{R}^n\to\mathcal{X}, \qquad (x\odot\alpha)_i = x_i\cdot\mathbb{I}[\alpha_i = 1]\in\mathbb{R}^{d_i}.$$

All of our proofs are identical under this construction, with the exception of the dimensionalities of terms like $(x\odot\alpha)$. An example of feature grouping is given in Figure 1.