On Minimizing the Impact of Dataset Shifts on Actionable Explanations

Abstract

The Right to Explanation is an important regulatory principle which allows individuals to request actionable explanations for algorithmic decisions. However, several technical challenges arise when providing such actionable explanations in practice. For instance, models are periodically retrained to handle dataset shifts, and this may in turn invalidate some of the previously prescribed explanations thus rendering them unactionable. But, it is unclear if and when such invalidations occur, and what factors determine explanation stability i.e., if an explanation remains unchanged amidst model retraining due to dataset shifts. In this paper, we address the aforementioned gaps and provide one of the first theoretical and empirical characterizations of the factors influencing explanation stability. To this end, we conduct rigorous theoretical analysis to demonstrate that model curvature and robustness, weight decay parameters, and the magnitude of the dataset shift are key factors that determine the extent of explanation (in)stability. Extensive experimentation with real-world datasets not only validates our theoretical results, but also demonstrates that the aforementioned factors dramatically impact the stability of explanations produced by various state-of-the-art methods.

1 INTRODUCTION

Machine learning (ML) models have recently witnessed increased utility in critical real-world applications. This, in turn, led to the introduction of several regulatory principles which aim to safeguard the practice of algorithmic decision making in such settings [1, 2]. The Right to Explanation is one such important regulatory principle which allows individuals to request actionable explanations for algorithmic decisions that adversely impact them. Post hoc explanation methods such as LIME, SHAP, input gradients, and Smoothgrad [24, 30, 33, 35] have commonly been employed in practice to operationalize this principle. These methods explain complex model predictions by assigning importance scores to input features, typically via learning local linear approximations of the underlying functions [15].

Providing actionable explanations to end users is often hindered by operational challenges in practice. For instance, models are periodically retrained to handle dataset shifts, and this retraining process may in turn invalidate some of the previously prescribed explanations, thus rendering them unactionable. For example, a user may be informed that their low salary was the primary reason for the rejection of their loan application, thus prompting them to increase their income. However, an update to the model could result in a shift of model internals such that the user's credit score, and not their income is now the primary factor for the rejection. In this case, the user's action of increasing their income would be less likely to result in a positive outcome once the updated model was deployed.

The aforementioned scenario could have been avoided if the underlying model and its explanations remained (relatively) unchanged in the face of model retraining due to dataset shifts. Such *explanation stability* implies actionability because if explanations remain (relatively) unchanged despite model retraining, they are likely to be actionable over an extended period of time. Therefore, it is necessary to develop a systematic understanding of explanation stability in the face of dataset shifts, characterize the conditions that lead to unstable explanations, and find ways to mitigate this challenge. Despite its significance, there is little to no research on characterizing the factors that influence explanation stability.

Our work addresses the aforementioned critical gaps by providing one of the first theoretical and empirical characterizations of the factors influencing explanation stability in the face of dataset shifts. To this end, we conduct rigorous

Submitted to the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023). **To be used for reviewing only**.

theoretical analysis to demonstrate that model curvature and robustness, weight decay parameters, and the magnitude of the dataset shift are key factors that determine the extent of explanation (in)stability. Our theoretical analysis emphasizes how seemingly inconsequential modeling decisions which may not impact predictive accuracy can heavily influence other critical aspects such as explanation stability. We also conduct extensive experimentation with real-world datasets to validate our theoretical insights with various state-of-the-art explanation methods (e.g., gradient and perturbation based methods such as LIME, SHAP, SmoothGrad etc.) and explanation stability metrics (e.g., ℓ_2, top-k consistency). We also empirically analyze how other training decisions (e.g., learning rate and number of training epochs) impact explanation stability. In summary, our work makes the following key contributions:

1. We provide one of the first theoretical and empirical characterizations of the factors influencing explanation stability in the face of dataset shifts.
2. We conduct a rigorous theoretical analysis to demonstrate that model curvature and robustness, weight decay parameters, and the magnitude of the dataset shift are key factors influencing the extent of explanation (in)stability (§3).
3. We carry out extensive experimentation with multiple real-world datasets to validate our theoretical results (§4.1). Our empirical findings suggest that standard neural network training pipelines exhibit low explanation stability, and confirm that our theoretical results are valid even when some of the underlying assumptions do not hold in practice (§4.2).
4. We also empirically analyze the impact of other training hyperparameters such as learning rates and number of training epochs on explanation stability (§4.3).

2 RELATED WORK

Explanation Methods A variety of post-hoc techniques have been proposed to explain complex models [10, 20, 31]. These techniques differ in their access to the complex model (i.e., black box vs. access to internals), scope of approximation (e.g., global vs. local), search technique (e.g., perturbation-based vs. gradient-based), explanation families (e.g., linear vs. non-linear), etc. For instance, LIME [30] and SHAP [24] are *perturbation-based*, *local explanation* approaches that learn a linear model locally around each prediction. Other *local explanation* methods capture feature importances by computing the gradient with respect to the input [32, 33, 35, 37]. Counterfactual explanation methods, on the other hand, capture the changes that need to be made to a given instance in order to flip its prediction [18, 19, 23, 27, 39, 40, 41]. In this work, we focus on analyzing the stability of explanations output by perturbation-based and gradient-based local explanation methods.

Robustness of Explanations Various notions of explanation robustness have been suggested in the literature. One line of work focuses on input perturbations by showing how to craft adversarial manipulations of input perturbations that change their explanations but not predictions [9, 14]. Another approach is to alter the underlying model to maintain accuracy, but change explanations to a desired target [3, 17, 34]. Multiple works suggest using smooth or low-curvature models to improve both model and explanation robustness [9, 36]. However, these works focus on adversarial manipulation of either the model or the input, while we are interested in more realistic model shifts occurring due to natural dataset shifts. In contrast, algorithmic approaches for generating counterfactual explanations that are robust to model shifts has been studied by prior research [29, 38]. However, unlike prior works, we focus on feature-attribution methods and theoretically characterize the factors that impact explanation stability.

More generally, neural networks are known to be non-robust to small training modifications, e.g., different random initializations or alternate model selection due to underspecification [4, 8, 22, 25]. However, relatively few works have studied how *explanations* change in the same setting. These works only show empirically that explanations are not robust to underspecification [7] or random model initializations [5]. These works are orthogonal to ours, as we focus on theoretically and empirically characterizing specific properties of models and datasets that impact explanation stability.

Algorithmic Stability Algorithmic stability is a classical learning theoretic framework that characterizes the consistency of outputs of learning algorithms when trained on different but similarly distributed data [6]. In contrast, our study does not make the assumption that the shifted data is similarly distributed. [16] studied the hypothesis stability of stochastic gradient descent (SGD) and showed that practical choices such as learning rate and number of epochs of training can impact the stability of model parameters. Recent work has applied the algorithmic stability perspective to post hoc explanations [12] by requiring that feature attributions be stable to a re-sampling of the underlying dataset. However, none of these prior works study realistic model shifts occurring due to natural dataset shifts or characterize the factors influencing explanation stability, which is the main focus of our work.

3 THEORETICAL ANALYSIS

In this section, we provide an analytical characterization of explanation shifts $e_1 \rightarrow e_2$ and model shifts $f_1 \rightarrow f_2$ that happen due to underlying dataset shifts $\mathcal{D}_1 \rightarrow \mathcal{D}_2$. Concretely, we address the following question:

How much do the explanations for a model f_1 trained on dataset $\mathcal{D}_1$ change when fine-tuning[1] on a slightly shifted dataset $\mathcal{D}_2$, resulting in a new model f_2?

Let the dataset $\mathcal{D}_1$ be composed of input-output pairs $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^d$, and y is either a real value or a one-hot vector depending on the task (regression or classification). For simplicity of analysis, in this section we shall focus on gradient explanations $e = \nabla_{\mathbf{x}} f \in \mathbb{R}^d$, where f is a scalar-valued neural network function $f(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R}$ mapping D-dimensional inputs to scalar outputs. Let $\ell(\cdot, \theta) : \mathbb{R}^d \to \mathbb{R}$ denote f composed with a loss function such as cross-entropy loss. Here, θ represents the parameters of the function being minimized, e.g., a deep neural network.

Our goal here is to relate the explanation shift to the dataset shift (i.e., $\mathrm{d}(\mathcal{D}_1, \mathcal{D}_2)$, for some distance measure d) and the learning algorithm. To this end, our overall strategy is:

1. In §3.1, we bound the *parameter shift* $\|\theta_2 - \theta_1\|_2$ in terms of the *dataset shift* $d(\mathcal{D}_1, \mathcal{D}_2)$.
2. In §3.2, we bound the *explanation shift* $\|\nabla_{\mathbf{x}} f(\mathbf{x}, \theta_2) - \nabla_{\mathbf{x}} f(\mathbf{x}, \theta_1)\|_2$ in terms of the *parameter shift* $\|\theta_2 - \theta_1\|_2$ derived above.

While the existence of these shifts are qualitatively evident, we make novel contributions in that (a) we quantify the parameter and gradient changes and (b) we show that these changes are affected by factors in the modelling process.

3.1 BOUNDING THE PARAMETER SHIFT

We begin by defining the distance between datasets $\mathrm{d}(\mathcal{D}_1, \mathcal{D}_2)$ using Hungarian distance, as defined in Defn. 1. The idea behind the Hungarian distance is simple: we find the "alignment" between data points of two datasets that minimizes the average ℓ_2 distance between a point in $\mathcal{D}_1$ and its counterpart in $\mathcal{D}_2$.

Definition 1. *Given datasets $\mathcal{D}_1 = \{(x_i, y_i)_{i=1,\ldots,N}\}$ and $\mathcal{D}_2 = \{(x'_i, y'_i)_{i=1,\ldots,N}\}$, their* Hungarian distance *is*

$$d(\mathcal{D}_1, \mathcal{D}_2) = \min_{P(\mathcal{D}_2)} \sum_{i=1}^{N} \|x_i - x'_i\|_2$$

where the minimum is taken over $P(\mathcal{D}_2)$, all permutations on the data points of $\mathcal{D}_2$.

Using this distance metric, we derive an expression for the parameter change $\|\theta_2 - \theta_1\|_2$ in terms of the dataset distance $d(\mathcal{D}_1, \mathcal{D}_2)$. We use the following assumptions to derive this result: (1) we minimize a regularized loss ℓ_{reg} that involves weight decay, i.e, $\ell_{reg}(\theta) = \ell(\theta) + \gamma\|\theta\|_2^2$; (2) ℓ is locally quadratic; (3) the learning algorithm returns a unique minimum θ given a dataset $\mathcal{D}$.

Theorem 1. *Given the assumptions stated above, and that $\mathcal{L}_x(\theta_1)$ is the Lipschitz constant of the model with parameters θ_1, we have*

$$\|\theta_2 - \theta_1\|_2 \leq \sqrt{\frac{\mathcal{L}_x(\theta_1) d(\mathcal{D}_1, \mathcal{D}_2)}{\gamma}} + C$$

where γ is the weight decay regularization constant, and C is a small problem-dependent constant.

Proof Idea. The idea of the proof is to first estimate $\ell_{\mathcal{D}_2}(\theta_1)$, i.e., the initial loss of the model θ_1, before fine-tuning on $\mathcal{D}_2$. Assuming, for the sake of explanation, that the optimal loss value obtained by θ_2 on $\mathcal{D}_2$ is zero, we have obtained the change in loss value from $\theta_1 \to \theta_2$. To derive the change in parameters from the change in loss derived above, we use a second order Taylor series expansion (with the assumption that the loss is locally quadratic). In particular, we lower bound the Hessian with its lowest eigenvalue using the weight decay term γ. The "problem-dependent constant" C arises from deviating from the assumption that the optimal loss value is zero. The complete proof is provided in the supplementary material. □

The above theorem achieves our first goal of relating parameter change to dataset shift. Intuitively, this tells us that parameter shift depends directly on dataset shift (as expected), directly on the Lipschitz constant (which depends on the model's robustness), and inversely on the weight decay parameter.

3.2 BOUNDING THE EXPLANATION SHIFT

We now consider the problem of bounding the explanation shift $\|\nabla_{\mathbf{x}} f(\cdot, \theta_2) - \nabla_{\mathbf{x}} f(\cdot, \theta_1)\|_2$ given an estimate of the parameter shift $\|\theta_2 - \theta_1\|_2$. To this end, we first define a quantity called the *gradient-parameter Lipschitz constant*, defined in Defn. 2. To the best of our knowledge, this quantity has not been considered in previous works.

Definition 2. *We define the "gradient-parameter Lipschitz constant" $\mathcal{L}_{\Theta,\mathcal{D}}$ wrt an input distribution $\mathcal{D}$, and parameter set Θ as follows:*

$$\mathcal{L}_{\Theta,\mathcal{D}} = \mathbb{E}_{\theta \in \Theta} \mathbb{E}_{\mathbf{x} \sim \mathcal{D}} \|\nabla_\theta \nabla_x f(x; \theta)\|_2$$

Intuitively, this quantity captures the sensitivity of gradient explanations to small changes in the parameter θ. In contrast to usual definitions of Lipschitz constants, this is defined (1)

[1] Here, fine-tuning refers to initializing model parameters of f_2 with those of f_1, and then training f_2 using D_2 until its loss converges. We provide practical details of this in §4.1.

locally with respect a particular distribution and parameter set, and (2) using the mean instead of the sup. This allows us to derive the following relationship between average gradient difference and parameter shift.

Lemma 1. *The gradient-parameter Lipschitz has the following property:*

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\|\nabla_x f(x;\theta_1) - \nabla_x f(x;\theta_2)\|_2 \leq \mathcal{L}_{\Theta,\mathcal{D}} \times \|\theta_2 - \theta_1\|_2$$

where $\Theta = \{\lambda\theta_1 + (1-\lambda)\theta_2 \mid \lambda \in [0,1]\}$.

This follows immediately from the fundamental theorem of integral calculus, and the full proof is given in the supplementary material. This result reveals an intuitive fact that larger parameter shifts lead to larger gradient shifts, but mediated by the gradient-parameter Lipschitz. This quantity is difficult to analyse for general neural networks, and we provide below an analysis for the case of 1-hidden layer neural networks, with a specific simplifying assumption on the data distribution.

Theorem 2. *Assume that we have a 1-hidden layer neural network with weights θ, and random inputs $\mathbf{x} \sim \mathcal{N}(0, I)$[2]. Further assume that we use an activation function σ with well-defined second derivatives (e.g: softplus). For this case, the gradient-parameter Lipschitz constant is*

$$\mathbb{E}_{\mathbf{x},\theta}\|\nabla_\theta \nabla_{\mathbf{x}} f(\mathbf{x},\theta)\|_2 \leq (\mathbb{E}_\theta\|\theta\|_2) + \beta\,(\mathbb{E}_\theta\phi(\theta))$$

where β is the maximum curvature of activation function σ, and $\phi(\theta)$ is the path-norm [26] of the model.

Proof Idea. The proof involves computing the gradient-parameter derivatives for the 1-hidden layer neural network case. Our assumption regarding smoothness of activation function ensures that $\sigma'' \leq \beta$, i.e., that the second derivative of the activation function is upper bounded, which is true for smooth non-linearities like softplus. Another simplifying assumption regarding the distribution of the inputs ($\mathbf{x} \sim \mathcal{N}(0, I)$) helps us dramatically simplify the expression for the expected value of the gradient norm terms.

The path-norm for a 1-hidden layer neural network (for weights W_2, W_1) is given by $\phi(\theta) = \sqrt{\sum_{i,j}(W_2^j W_1^{i,j})^2}$ and has been linked to model generalization [26]. The complete proof is provided in the supplementary material. □

These results achieve our goal of relating the change in gradients to parameter change. Taken together with the results in §3.1, this achieves our overall goal of relating gradient change to dataset shift. Intuitively, this tells us the following regarding gradient shift:

1. It depends directly on the *dataset distance* $d(\mathcal{D}_2, \mathcal{D}_1)$, which is as expected: the larger the dataset shift, the larger can be the parameter shift, and explanation shift.
2. It depends inversely on the *weight decay* parameter γ, and directly on the norm of weights $\|\theta\|_2$. Intuitively, large weight decay shifts the optima to be closer to zero, reducing the norm of weights, which makes all optima closer to each other.
3. It depends directly on the *Lipschitz constant* of the first model, which is connected to the model's robustness. Intuitively, if the model is already robust to small input changes, we don't expect it to shift too much upon fine-tuning on slightly shifted data.
4. It depends directly on the *smoothness constant* β of the activation function used in the model. The smoother the model, the smaller are its second derivatives, and thus its gradient-input Lipschitz.

These insights motivate the core thesis of the paper - that explanation stability depends on properties beyond predictive performance, and that specific algorithm choices that do not necessarily improve accuracy, such as weight decay, smoothness constant and robustness-inducing losses, affect explanation and model stability to data shifts.

We conclude by making two remarks regarding the theory. First, it only highlights *sufficient conditions* for explanation stability, not necessary conditions. Meaning, there can be ways to achieve gradient stability without the parameters θ_1, θ_2 being close, or without neural nets being Lipschitz, or when ReLU or weight decay are not used, but these lie outside the scope of this theory. Second, the focus of these bounds is not to produce numerically tight estimates for explanation stability, rather, it is intended as a tool to make *qualitative predictions* regarding specific modelling interventions that can improve explanation stability in practice. Section 4 explores this in more detail.

4 EXPERIMENTAL EVALUATION

In this section, we describe our general evaluation setup. In §4.1, we experimentally validate that the modeling interventions identified in §3 improve explanation stability. Next, in §4.2, we show that these relationships extend to realistic training setups and more-sophisticated explanation techniques resulting in top-K feature attributions. Finally, in §4.3, we show that other hyperparameters not discussed in our theory – e.g., learning rate and number of training epochs – also impact gradient stability.

Datasets We evaluate explanation stability on three binary datasets with continuous features. We use the WHO life expectancy dataset (n=2928, number of features d=18) [28], which evaluates whether a country's life expectancy is above the global median based on health and economic factors;

[2]Covariance of I is chosen for notational brevity

the HELOC dataset (n=9871, d=23) [13], which evaluates loan acceptance given the applicant's financial information; and the Adult Income dataset (n=32561, d=6) [11], which evaluates whether a person's income is above $50,000. We model dataset shifts in two ways:

- *Synthetic Noise.* We add Gaussian $\mathcal{N}(0, \sigma^2)$ noise to all training samples, with varying levels of noise σ.
- *Temporal Shift.* We create a temporal shift dataset from the WHO data by using pre-2012 data as the "original" dataset and the full dataset as the "shifted" dataset.

Models We evaluate the stability of medium-sized neural networks containing 5 layers of 50 nodes. By default, we use ReLU activation; however, to see the effects of model curvature on explanation stability, we use softplus (SP) for some experiments. For the experiments using SP, we vary β between 2 and 10 depending on the dataset and the experiment's goals. Smaller values of β correspond to lower-curvature models, whereas SP with $\beta = \infty$ is equivalent to ReLU. Unless otherwise specified, all models we compare within each section achieve similar predictive accuracy.

Experimental procedure For each experiment, we average over 10 trials to reduce the impact that random seeds have on the outcomes. Each trial in the synthetic noise experiments consists of training one "base" model plus 10 "noisy" models, each corresponding to a different random perturbation of the data. In the temporal shift experiments, each trial compares the base model (trained on the partial dataset) to a new model (trained on the full dataset). In the results, we display the mean of all trials along with the middle 50% of values.

4.1 VALIDATING OUR THEORETICAL RESULTS

Experiment setup In this section, we test our theoretical results from §3 by evaluating the impact of *weight decay*, *model curvature*, and *size of the data shift* on explanation stability. To ensure that our analysis is sound, we adhere to the theoretical assumptions as closely as possible. In particular, we focus on *model parameters* (θ) and *gradients* (g) rather than top-K feature attributions, since top-K explanations are not differentiable functions. We train each base model for a large number of epochs, to ensure convergence of the training loss to a minimum, ignoring any overfitting on the test set. To fine-tune each base model, we initialize a new model with the same parameters as the base model before training it on the shifted data. Additionally, we start fine-tuning with a lower learning rate and train for fewer epochs, given that the initial loss is already close to a minimum.

For brevity, we omit similar results from the Adult dataset, with complete results in the appendix, where the specific hyperparameters used for each dataset are also provided.

Weight decay results Figure 1 and Table 1 show the effect of weight decay for the synthetically- and naturally-shifted datasets, respectively. We see that larger weight decay values correspond to orders of magnitude smaller mean gradient and parameter changes, which is in line with the theoretical relationship between weight decay and stability. As the amount of synthetic noise grows, we observe that both parameters and gradients diverge more between the base model and fine-tuned model. This pattern supports our hypotheses that (a) the loss landscape changes more with larger data shift and (b) a larger change in model parameters is correlated with a larger change in test sample gradients.

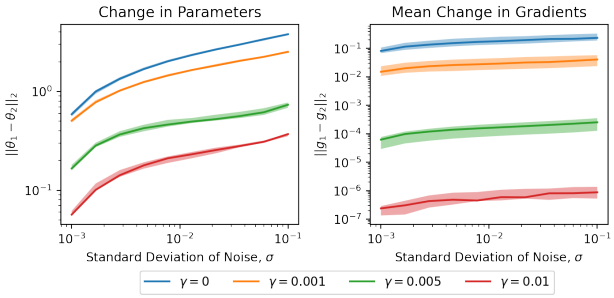

Figure 1: Effect of weight decay value on the mean gradient and parameter changes for the HELOC dataset after fine-tuning on a shifted dataset. All models use ReLU activation. The x-axis is the size of the data shift, represented by the standard deviation σ of Gaussian $\mathcal{N}(0, \sigma^2)$ noise.

Table 1: Effect of weight decay (γ) on WHO (with fine-tuning).

	$\gamma = 0$	0.001	0.01
$\|\theta - \theta'\|_2$	1.69 ±0.09	1.53 ±0.03	0.95 ±0.02
$\|g - g'\|_2$	0.201 ±0.065	0.122 ±0.034	0.010 ±0.002

Curvature results Figure 2 and Table 2 show that using low-curvature training techniques increases gradient stability. In particular, SP does better than ReLU, particular with smaller values of β. For both the real-world and synthetic shifts, the gradient changes are less distinct (i.e., the confidence intervals overlap more) than the parameter changes, but still shows a trend in the predicted direction.

Table 2: Effect of curvature on WHO (with fine-tuning).

	ReLU	SP ($\beta = 10$)	SP ($\beta = 5$)
$\|\theta - \theta'\|_2$	1.53 ±0.03	1.37 ±0.05	1.23 ±0.02
$\|g - g'\|_2$	0.122 ±0.034	0.096 ±0.026	0.070 ±0.024

4.2 GOING BEYOND THEORETICAL RESULTS

In this section, we evaluate whether our theoretical predictions still hold if we deviate from the theoretical assump-

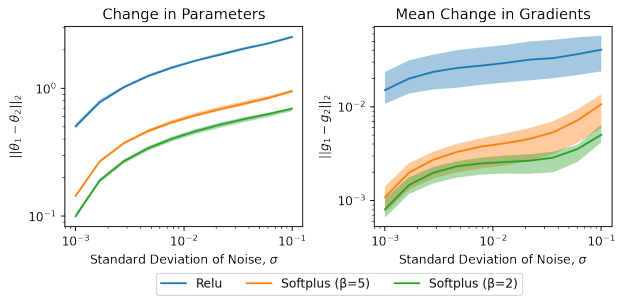

Figure 2: Effect of model curvature on the mean gradient and parameter changes for the HELOC dataset after fine-tuning on a shifted dataset. The x-axis is the size of the data shift, represented by the standard deviation σ of Gaussian $\mathcal{N}(0, \sigma^2)$ noise.

tions, namely: fine-tuning, exactly reaching a local minimum, and using non-differentiable top-K feature attribution metrics rather than gradients ℓ_2 distance.

Experiment setup We maintain our experimental setup from §4.1, except that we train each model for a smaller number of epochs (until accuracy stabilizes), before retraining a model from scratch on the shifted data, rather than fine-tuning (hyperparameters are located in the appendix). We use the same random network initialization for both training and retraining, as prior work has shown that feature attributions are not stable to changes in random seed [5].

To complete the analysis, we consider four feature attribution techniques: input gradients, input gradients with SmoothGrad, LIME, and kernel SHAP [24, 30, 35, 33]. We use three top-K stability metrics adapted from Brunet et al. [7]. They each take in a pair of top-K feature sets corresponding to two models. Sign Agreement (SA) returns the fraction of top-K features that appear in both models' top-K features and have the same signed value. Consistent Direction of Contribution (CDC) is binary (per-sample) and measures whether all features in the top-K (for either model) have the same signed value in the other model. Signed-Set

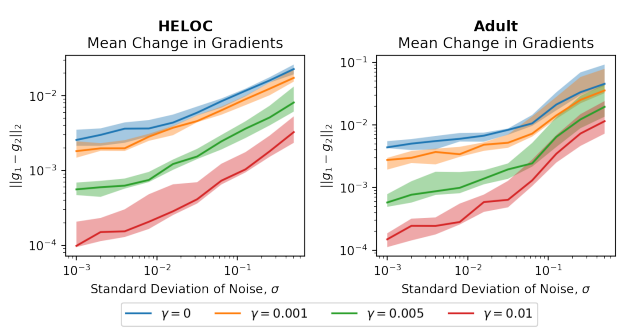

Figure 3: Effect of weight decay on parameter stability when retraining on the Adult and HELOC datasets. The x-axis is the size of the data shift, represented by the standard deviation σ of $\mathcal{N}(0, 1)$ noise.

Agreement (SSA) is also binary (per-sample) and is 1 if the two model's top-K feature sets contain the same features and the features have the same signed value (the order of the top-K features does not matter).

Relaxing Training Assumptions From Table 3, we see that increasing the algorithm's weight decay significantly decreases both parameters and gradients. Figure 3 confirms that a larger weight decay during training also increases gradient stability for models with synthetic noise (analogous results for parameter stability are included in the appendix). Figure 4 and Table 4 show that the curvature trends discovered in §4.1 still hold, i.e., lower model curvature induces higher gradient stability.

Table 3: Effect of weight decay (γ) on WHO (with retraining).

	$\gamma = 0$	0.001	0.01
$\|\theta - \theta'\|_2$	3.97 ±0.13	3.50 ±0.19	1.79 ±0.17
$\|g - g'\|_2$	0.295 ±0.075	0.150 ±0.034	0.003 ±0.001

Table 4: Effect of curvature on WHO (with retraining).

	ReLU	SP ($\beta = 10$)	SP ($\beta = 5$)
$\|\theta - \theta'\|_2$	3.50 ±0.19	2.60 ±0.05	1.97 ±0.12
$\|g - g'\|_2$	0.150 ±0.034	0.102 ±0.040	0.076 ±0.024

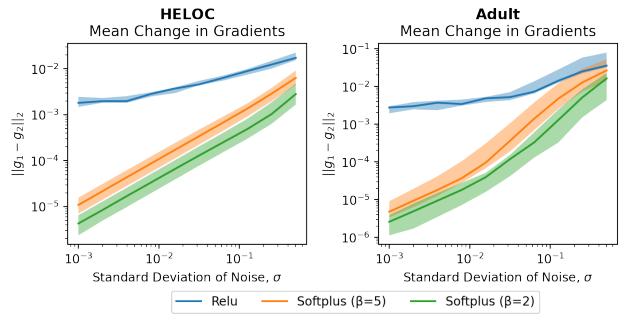

Figure 4: Effect of model curvature on gradient stability when retraining on the HELOC and Adult datasets. The x-axis is the size of the data shift, represented by the standard deviation σ of $\mathcal{N}(0, 1)$ noise.

Although curvature and weight decay increase explanation stability upon retraining, we note that the parameter stability scores are much lower for retraining than for fine-tuning given fixed curvatures and weight decays. For example, the mean change in parameters when using ReLU with a weight decay of 0.001 is 1.53 ±0.03 when fine-tuning, versus 3.50 ±0.19 when retraining. The difference in gradient change between the two techniques is less stark, however, suggesting that even though the fine-tuned models are more similar to each other overall, the retrained models are sufficiently similar on the parts of the input domain represented in the test set.

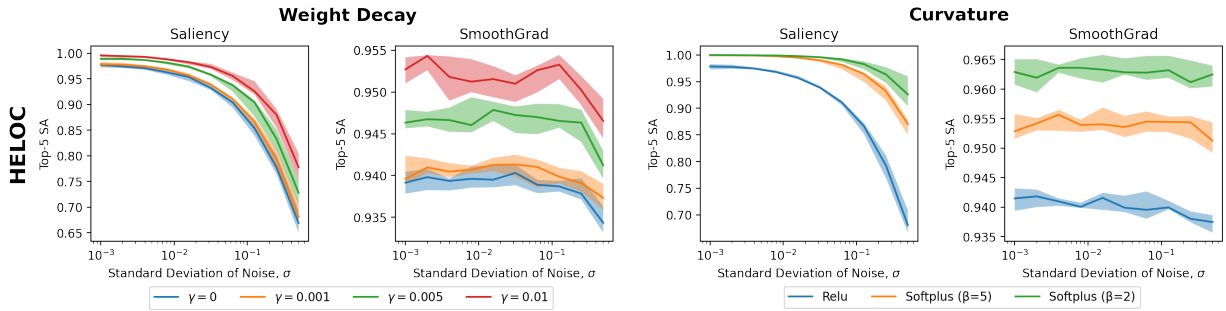

Figure 5: Top-5 SA for the HELOC dataset under model retraining. The left two graphs show the effect of weight decay and the right two graphs show the effect of curvature. Graphs 1 and 3 show Top-5 SA for saliency, while graphs 2 and 4 show Top-5 SA for SmoothGrad. Confidence intervals represent the middle 50% of values. See the appendix for other datasets and explanation techniques.

Using Diverse Explanation Techniques Figure 5 shows Top-5 SA stability for two popular explanation techniques, namely, Saliency (input gradients) and SmoothGrad [21]. As in the previous experiments, models with low curvature and large weight decay are more effective at preserving explanation stability. In case of saliency, the explanation stability starts to decrease sharply for shift magnitudes larger than $\mathcal{N}(0, 0.1)$. For SmoothGrad, the noise has less of an effect, which we hypothesize is because SmoothGrad is more stable overall as it computes an averaged gradient over a small local neighborhood of a given instance (see Table 5). Lowering the model's curvature and increasing the weight decay similarly improve explanation stability under data shifts for the other metrics and explanation techniques (see the appendix for a full set of graphs).

Table 5 gives a more detailed breakdown of top-K feature attribution stability across all datasets and explanation techniques. The table corresponds to a single modeling choice – ReLU with weight decay 0.001. We notice trends based on the *explanation technique*, *explanation metric*, and *dataset*.

The different explanation techniques exhibit vastly different stabilities under dataset shifts. Notably, SmoothGrad outperforms all other explanation techniques, with over 90% of test samples exhibiting Top-3 and Top-5 explanation stability for all of the metrics we used.

To compare the explanation metrics, we look at the relative difference between SSA and each of SA and CDC (since SSA is strictly stronger than either of the other metrics). We see that neither SA nor CDC is consistently more stable, suggesting that the failure to have identical top-K feature sets stems both from different features ranking as the most influential (i.e., what SA captures) and also whether the top features have the same sign in the other model (i.e., CDC).

We also see that different datasets exhibit different levels of explanation stability. The top-3 and top-5 scores for the Adult datasets are very high; however, this dataset only has six features, so it is a much lower bar to achieve top-k SA, in particular. We see that SSA scores for Adult are much lower than the SA scores (although still higher than the SSA scores from other datasets), indicating that even with only 6 features, the learnt models often disagree on their impacts.

4.3 SENSITIVITY ANALYSIS

In this section, we analyze the impact of other training hyperparameters on explanation stability throughout the entire training process. While we see strong empirical trends for the impact of weight decay and curvature on explanation stability, practitioners should remain conscious of the effects that training hyperparameters can also have. In Section 4.2, base models were trained with the same set of hyperparameters as the models trained on shifted data (retrained models). However, the particular set of hyperparameters chosen at this stage can also impact explanation stability significantly.

Given our observation that parameter stability generally promotes explanation stability, we expect a lower number of epochs and a lower learning rate to both increase the likelihood of explanation stability under dataset shift by decreasing the divergence between optimization trajectories in parameter space. Based on the geometry of divergent optimization paths, we anticipate that if the base model is trained for a longer period, the retrained model's parameters that are most similar to the base model will occur relatively sooner in the retraining process. This emphasizes the challenge of retraining models that both perform well and remain close to the base model.

We track explanations during each epoch of retraining, measuring similarity using gradient similarity and top-K SA consistency. Each trial consists of varying a single hyperparameter while keeping all others fixed. For each value of the varied hyperparameter, we compare one base model with 10 models retrained on synthetically shifted HELOC data.

Learning Rate For fixed synthetic noise ($\sigma = 0.1$), we verify in Figure 6 that increasing the learning rate of all models effectively worsens gradient stability (base models are trained for 30 epochs). We observe that the peaks in

Table 5: Explanation stability scores for all datasets. The synthetically-shifted datasets are modified by $\mathcal{N}(0, 0.1^2)$ noise. Models use ReLU activation and weight decay $\gamma = 0.001$. All values averaged across all test samples and then across all trials. Error bounds are such that the lower and upper bound values represent the 25th and 75th percentile respectively.

Dataset	Explanation technique	Top-3 SA	Top-3 CDC	Top-3 SSA	Top-5 SA	Top-5 CDC	Top-5 SSA
WHO	Saliency	0.63±0.01	0.83±0.03	0.18 ±0.03	0.61±0.03	0.61±0.10	0.03±0.01
	SmoothGrad	0.94 ±0.00	0.94±0.01	0.91 ±0.00	0.94±0.00	0.92±0.00	0.90±0.00
	LIME	0.69 ±0.09	0.35 ±0.24	0.27±0.20	0.59±0.57	0.08±0.05	0.06 ±0.05
	K.SHAP	0.58 ±0.10	0.60±0.39	0.13±0.08	0.61±0.10	0.41±0.31	0.05 ±0.05
HELOC	Saliency	0.51±0.02	0.88±0.03	0.07±0.01	0.55±0.02	0.71±0.05	0.01±0.00
	SmoothGrad	0.94±0.00	0.96±0.00	0.90±0.00	0.94±0.00	0.94±0.00	0.90±0.00
	LIME	0.63±0.02	0.45±0.04	0.18±0.03	0.66±0.02	0.16±0.03	0.07±0.02
	K.SHAP	0.64±0.02	0.91±0.03	0.19±0.03	0.70±0.02	0.79±0.05	0.09±0.02
Adult	Saliency	0.87±0.01	0.92±0.01	0.62±0.04	0.85±0.01	0.57±0.04	0.38±0.03
	SmoothGrad	0.98±0.00	0.98±0.00	0.96±0.00	0.98±0.00	0.96±0.00	0.94±0.00
	LIME	0.95±0.00	0.86±0.02	0.86±0.02	0.97±0.00	0.84±0.02	0.84±0.02
	K.SHAP	0.88±0.01	0.88±0.04	0.65±0.04	0.84±0.02	0.47±0.10	0.32±0.08

gradient stability occur sooner when the learning rate is higher. This is because a high learning rate stretches the optimization path of the base model, taking it much further away from the initialization in parameter space. Since the retrained models and the base models all take divergent paths (due to the dataset shifts), the likelihood that they achieve similar parameters decreases as the learning rate increases.

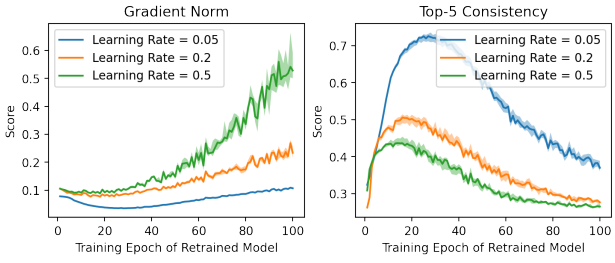

Figure 6: Sensitivity to learning rate. Gradient similarity (y-axis) of retrained models are shown per epoch of training (x-axis).

Base Model Epochs Figure 7 verifies that as the number of base model epochs is increased, explanation stability suffers, with significantly larger gradient changes and lower top-5 consistency scores. For a sufficiently small base model epoch of 20, we observe that similarity also peaks at around 20 training epochs. However, if we were to increase the base model epoch to 60, as shown in green, we observe that the peak still occurs at around the same value. This means that, after just 20 epochs, the optimization path of the retrained model must have diverged so much from the base model that any further training leads only to a decrease in stability.

The overall trend shown in these results is that a longer training process has two implications for gradient stability. Firstly, it worsens the best possible stability value that can be attained. Secondly, it also fails to achieve this best value unless the retraining process is shortened appropriately.

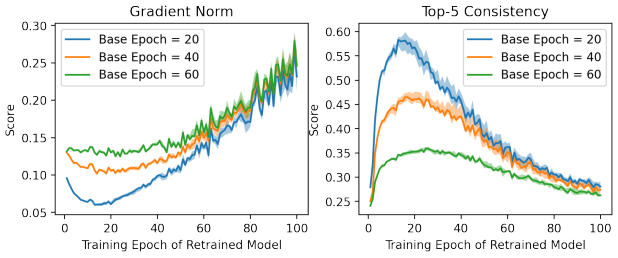

Figure 7: Sensitivity to the base model's number of epochs. Gradient similarity (y-axis) of retrained models are shown per epoch of training (x-axis). Learning rate is 0.2.

5 CONCLUSIONS

In this paper, we showed that model explanations are often not stable under model updates due to small dataset shifts, but that modelling choices like using low-curvature activation functions and and large weight decay values can greatly increase their stability. The message for ML practitioners is clear: in order to preserve explanations upon model retraining due to data shifts at little cost to accuracy, use softplus with the lowest possible β, and the largest weight decay that does not affect predictive accuracy. If possible, practitioners should also fine-tune the model (§4.1) rather than retrain from scratch to avoid instability due to the multiplicity of optimal solutions. We also took a first step towards characterizing how different explanation techniques yield variable explanation stability rates under dataset shifts, noting that SmoothGrad outperforms other methods. Future work can investigate how to develop novel explanation techniques which preserve explanation stability despite changes in the underlying models and data.

References

[1] Blueprint for an AI Bill of Rights.

[2] General Data Protection Regulation (GDPR).

[3] C. J. Anders, P. Pasliev, A.-K. Dombrowski, K.-R. Müller, and P. Kessel. Fairwashing explanations with off-manifold detergent. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

[4] E. Black and M. Fredrikson. Leave-one-out unfairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 285–295. Association for Computing Machinery, 2021.

[5] E. Black, K. Leino, and M. Fredrikson. Selective ensembles for consistent predictions. In *International Conference on Learning Representations*, 2022.

[6] O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

[7] M.-E. Brunet, A. Anderson, and R. Zemel. Implications of model indeterminacy for explanations of automated decisions. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[8] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, F. Hormozdiari, N. Houlsby, S. Hou, G. Jerfel, A. Karthikesalingam, M. Lucic, Y. Ma, C. McLean, D. Mincu, A. Mitani, A. Montanari, Z. Nado, V. Natarajan, C. Nielson, T. F. Osborne, R. Raman, K. Ramasamy, R. Sayres, J. Schrouff, M. Seneviratne, S. Sequeira, H. Suresh, V. Veitch, M. Vladymyrov, X. Wang, K. Webster, S. Yadlowsky, T. Yun, X. Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022.

[9] A.-K. Dombrowski, M. Alber, C. Anders, M. Ackermann, K.-R. Müller, and P. Kessel. Explanations can be manipulated and geometry is to blame. *Advances in neural information processing systems*, 32, 2019.

[10] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv: Machine Learning*, 2017.

[11] D. Dua and C. Graff. UCI machine learning repository, 2017.

[12] T. Fel, D. Vigouroux, R. Cadène, and T. Serre. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 720–730, January 2022.

[13] FICO. Explainable machine learning challenge, 2018.

[14] A. Ghorbani, A. Abid, and J. Zou. Interpretation of neural networks is fragile. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[15] T. Han, S. Srinivas, and H. Lakkaraju. Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[16] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

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

[18] A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 353–362, New York, NY, USA, 2021. Association for Computing Machinery.

[19] A.-H. Karimi, J. von Kügelgen, B. Schölkopf, and I. Valera. Algorithmic recourse under imperfect causal knowledge: a probabilistic approach. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 265–277. Curran Associates, Inc., 2020.

[20] P. W. Koh and P. Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1885–1894. JMLR.org, 2017.

[21] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020.

[22] J. F. Kolen and J. B. Pollack. Back propagation is sensitive to initial conditions. In *Proceedings of the 1990 Conference on Advances in Neural Information Processing Systems 3*, NIPS-3, page 860–867, San Francisco, CA, USA, 1990. Morgan Kaufmann Publishers Inc.

[23] A. V. Looveren and J. Klaise. Interpretable counterfactual explanations guided by prototypes. *ArXiv*, abs/1907.02584, 2019.

[24] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[25] J. Mehrer, C. J. Spoerer, N. Kriegeskorte, and T. C. Kietzmann. Individual differences among deep neural network models. *Nature communications*, 11(1):5725, November 2020.

[26] B. Neyshabur, R. Tomioka, and N. Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pages 1376–1401. PMLR, 2015.

[27] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, page 344–350, New York, NY, USA, 2020. Association for Computing Machinery.

[28] K. Rajarshi. Life expectancy (WHO), 2017.

[29] K. Rawal, E. Kamar, and H. Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts, 2020.

[30] M. T. Ribeiro, S. Singh, and C. 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*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[31] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.

[32] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[33] K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

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

[35] D. Smilkov, N. Thorat, B. Kim, F. Viegas, and M. M. Wattenberg. Smoothgrad: removing noise by adding noise. In *ICML Workshop on Visualization for Deep Learning*, Sydney, Australia, 2017.

[36] S. Srinivas, K. Matoba, H. Lakkaraju, and F. Fleuret. Efficient training of low-curvature neural networks. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[37] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In D. Precup and Y. W. Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 06–11 Aug 2017.

[38] S. Upadhyay, S. Joshi, and H. Lakkaraju. Towards robust and reliable algorithmic recourse. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16926–16937. Curran Associates, Inc., 2021.

[39] B. Ustun, A. Spanghler, and Y. Liu. Actionable recourse in linear classification. In *ACM Conference on Fairness, Accountability, and Transparency (FAT '19)*, 2019.

[40] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard journal of law & technology*, 31:841–887, 04 2018.

[41] W. Yang, J. Li, C. Xiong, and S. C. H. Hoi. Mace: An efficient model-agnostic framework for counterfactual explanation. *ArXiv*, abs/2205.15540, 2022.