On Minimizing the Impact of Dataset Shifts on Actionable Explanations (Supplementary material)

Anna P. Meyer*[1] **Dan Ley***[2] **Suraj Srinivas**[2] **Himabindu Lakkaraju**[2]

[1]Dept. of Computer Sciences, University of Wisconsin - Madison, USA
[2] Harvard University, Boston, MA, USA

A PROOFS

In this section, we provide the proofs for the theory presented in the main paper. We present the notations here again for convenience in Table 1. As with the main paper, we first bound parameter shift in §A.1 and then explanation shift in §A.2.

A.1 BOUNDING THE PARAMETER SHIFT

In order to bound the parameter shift, we first prove an intermediate result connecting the loss on $\mathcal{D}_2$ of the optimum obtained by minimizing the loss on $\mathcal{D}_1$.

Lemma 1. *The loss $\ell_{\mathcal{D}_2}$ on θ_1 is given by*

$$\ell_{\mathcal{D}_2}(\theta_1) \leq \epsilon_1 + \gamma\|\theta_1\|^2 + \mathcal{L}_x(\theta_1) d(\mathcal{D}_1, \mathcal{D}_2)$$

Proof.

$$\begin{aligned}
\ell_{\mathcal{D}_2}(\theta_1) &= \ell_{\mathcal{D}_1}(\theta_1) + (\ell_{\mathcal{D}_2}(\theta_1) - \ell_{\mathcal{D}_1}(\theta_1)) \\
&\leq \ell_{\mathcal{D}_1}(\theta_1) + |\ell_{\mathcal{D}_2}(\theta_1) - \ell_{\mathcal{D}_1}(\theta_1)| \\
&\leq \epsilon_1 + \gamma\|\theta_1\|^2 + \frac{1}{n}\sum_{i=1}^{n} |\ell(\theta_1, z_i) - \ell(\theta_1, z_i')| \quad \text{(Convexity of abs fn)} \\
&\leq \epsilon_1 + \gamma\|\theta_1\|^2 + \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_x(\theta_1)\|z_i - z_i'\|_2 \quad \text{(Defn of Lipschitz)} \\
&\leq \epsilon_1 + \gamma\|\theta_1\|^2 + \mathcal{L}_x(\theta_1)\underbrace{\frac{1}{n}\sum_{i=1}^{n}\|z_i - z_i'\|}_{d(\mathcal{D}_1,\mathcal{D}_2)}
\end{aligned}$$

□

In the result above, $\epsilon_1 + \gamma\|\theta_1\|^2$ is the optimal value of the regularized loss, and the term following that denotes the shift from the optimal value incurred. As mentioned in the paper, we use the following additional assumptions to derive the

*Equal Contribution

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Table 1: Notations

z_i	$(x_i, y_i);\ x_i \in \mathbb{R}^d;\ y \in \mathbb{R}$	data point
$\mathcal{D}_1$	$\{z_1, ..z_n\}$	original training set
$\mathcal{D}_2$	$\{z'_1, ...z'_n\}$	new training set
$\ell_{reg}(\theta, z_i)$	$\ell(\theta, z_i) + \gamma\|\theta\|_2^2;\ \ \ell(\theta, z_i) \in \mathbb{R}^+$	regularized loss function
$\ell_{\mathcal{D}_1}(\theta)$	$\frac{1}{n}\sum_{i=1}^{n} \ell_{reg}(\theta, z_i)$	loss incurred by θ on dataset D_1
θ_1	$\arg\min_\theta \ell_{\mathcal{D}_1}(\theta)$	optimal weights on dataset D_1
$\epsilon_1 + \gamma\|\theta_1\|^2$	$\ell_{\mathcal{D}_1}(\theta_1) = \min_\theta \ell_{\mathcal{D}_1}(\theta)$	optimal loss value on dataset $\mathcal{D}_1$

main 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. Assuming the loss ℓ_{D_2} is locally quadratic around the optimal solution θ_2, we can employ a second order Taylor series expansion as follows:

$$\ell_{D_2}(\theta_1) = \ell_{D_2}(\theta_2) + \nabla_\theta \ell_{D_2}(\theta_2)^\top(\theta_1 - \theta_2) + \frac{1}{2}(\theta_1 - \theta_2)^\top H_{\theta_2}(\theta_1 - \theta_2) \tag{1}$$

$$= \epsilon_2 + \gamma\|\theta_2\|_2^2 + \frac{1}{2}(\theta_1 - \theta_2)^\top H_{\theta_2}(\theta_1 - \theta_2) \quad \text{(global optimality of } \theta_2 \text{ on } S_2) \tag{2}$$

In the statement above we have used the fact that $\nabla_\theta \ell_{\mathcal{D}_2}(\theta_2) = 0$ because of first order optimality criteria, and that $\epsilon_2 + \gamma\|\theta_2\|^2 = \ell_{\mathcal{D}_2}(\theta_2)$, is the optimal value. Now we use Lemma 1 to substitute $\ell_{\mathcal{D}_2}(\theta_1)$, and we have:

$$\epsilon_2 + \gamma\|\theta_2\|_2^2 + \frac{1}{2}(\theta_1 - \theta_2)^\top H_{\theta_2}(\theta_1 - \theta_2) \leq \epsilon_1 + \gamma\|\theta_1\|_2^2 + \mathcal{L}_x(\theta_1) d(\mathcal{D}_1, \mathcal{D}_2) \tag{3}$$

$$\frac{1}{2}\lambda_{\min}(H_{\theta_2})\|\theta_1 - \theta_2\|_2^2 \leq (\epsilon_1 - \epsilon_2) + \gamma(\|\theta_1\|_2^2 - \|\theta_2\|_2^2) + \mathcal{L}_x(\theta_1) d(\mathcal{D}_1, \mathcal{D}_2) \tag{4}$$

To derive the second step we have used the following fact that a lower bound on the quadratic term is given by the lowest eigenvalue i.e., $\arg\min_{\theta; \|\theta\|=1} \theta^\top H \theta = \lambda_{\min} \implies \theta^\top H \theta \geq \lambda_{\min}\|\theta\|^2$. We now observe the following relationship for λ_{min}:

$$\lambda_{\min}(\nabla_\theta^2 \ell_{reg}(\theta_2)) = \lambda_{\min}(\nabla_\theta^2 \ell(\theta_2)) + 2\gamma \geq 2\gamma \tag{5}$$

as $\nabla_\theta^2 \ell(\theta_2)$ is positive definite due to global optimality. Putting eqn. 5 in eqn. 4, we have the final result. Here, the “problem dependent constant” $C = \sqrt{\frac{\ell_{\mathcal{D}_1}(\theta_1) - \ell_{\mathcal{D}_2}(\theta_2)}{\gamma}}$, which we can expect to be quite small in practice if the optimal values $\ell_{\mathcal{D}_1}(\theta_1) \approx \ell_{\mathcal{D}_2}(\theta_2)$. Thus in practice we can assume C to be very small.

□

A.2 BOUNDING THE EXPLANATION SHIFT

Lemma 2. *The parameter-input Lipschitz has the following property:*

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

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

Proof. The proof follows from the fundamental theorem of integral calculus. Let $g(\theta, \mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x}, \theta)$ for convenience.

$$\begin{aligned}
g(\theta_2, \mathbf{x}) - g(\theta_1, \mathbf{x}) &= \left(\int_{\lambda=0}^{1} \nabla_\theta g(\theta_\lambda, \mathbf{x}) d\lambda\right)^\top (\theta_2 - \theta_1) \\
\|g(\theta_2, \mathbf{x}) - g(\theta_1, \mathbf{x})\|_2 &\leq \| \int_{\lambda=0}^{1} \nabla_\theta g(\theta_\lambda, \mathbf{x}) d\lambda\|_2 \|\theta_2 - \theta_1\|_2 \quad \text{(Cauchy Schwartz)} \\
&\leq \left(\int_{\lambda=0}^{1} \|\nabla_\theta g(\theta_\lambda, \mathbf{x})\|_2 d\lambda\right) \|\theta_2 - \theta_1\|_2 \quad \text{(Jensen's inequality)} \\
\mathbb{E}_{\mathbf{x}\in\mathcal{D}}\|g(\theta_2, \mathbf{x}) - g(\theta_1, \mathbf{x})\|_2 &\leq \left(\int_{\lambda=0}^{1} \mathbb{E}_{\mathcal{D}}\|\nabla_\theta g(\theta_\lambda, \mathbf{x})\|_2 d\lambda\right) \|\theta_2 - \theta_1\|_2 \quad \text{(Swapping expectation and integral)} \\
\mathbb{E}_{\mathbf{x}\in\mathcal{D}}\|g(\theta_2, \mathbf{x}) - g(\theta_1, \mathbf{x})\|_2 &\leq \mathbb{E}_\Theta \mathbb{E}_{\mathcal{D}}\|\nabla_\theta g(\theta, \mathbf{x})\|_2 \|\theta_2 - \theta_1\|_2 \quad \text{(defn of expectation)}
\end{aligned}$$

□

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

$$E_{\mathbf{x}}\|\nabla_\theta \nabla_{\mathbf{x}} \ell(\mathbf{x}, \theta)\|_2 \leq E_{\mathbf{x}}\|\nabla_\theta \nabla_{\mathbf{x}} \ell(\mathbf{x}, \theta)\|_F \leq \|\theta\|_2 + \beta\phi(\theta)$$

where β *is an maximum curvature, and* $\phi(\theta)$ *is the path-norm [29] of the model.*

Proof. We derive this expression for the case of a scalar valued neural network $f(\mathbf{x}) = W_2\sigma(W_1\mathbf{x})$, where $W_2 \in \mathbb{R}^{h\times d}, W_1 \in \mathbb{R}^{1\times h}$, where h is the number of hidden layers, and d is the input dimensionality. Here, $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ is a point-wise smooth non-linearity with a well-defined curvature, such as softplus, but not ReLU. For such a non-linearity, let $\sigma''(x) := \frac{\partial^2 \sigma(x)}{\partial x^2} \leq M$. For softplus, this $M = \beta$ hyper-parameter is implicit in its definition [43].

By straightforward calculus, its gradient and the "parameter-input derivatives" are given by

$$\begin{aligned}
\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_i} &= \sum_{k=1}^{h} W_2^k \sigma'(W_1\mathbf{x})^k W_1^{i,k} \\
\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \partial W_2^j} &= \sigma'(W_1\mathbf{x})^j W_1^{i,j} \\
\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \partial W_1^{i,j}} &= W_2^j \sigma'(W\mathbf{x})^j + W_2^j \sigma''(W_1\mathbf{x})^j W_1^{i,j} \mathbf{x}_i \\
\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \partial W_1^{m,j}} &= W_2^j \sigma''(W_1\mathbf{x})^j W_1^{i,j} \mathbf{x}_m \quad (\forall m \neq i)
\end{aligned}$$

[1] Covariance of I is chosen for notational brevity

Now, assume that $\mathbf{x} \sim \mathcal{N}(0, I)$, i.e., the input is an independent normal distribution. Let us now compute the squared terms for the parameter-input derivatives to eventually be able to compute its Frobenius norm.

$$\begin{aligned}
\left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \partial W_2^j}\right)^2 &\leq (\sigma'(W_1\mathbf{x})^j W_1^{i,j})^2 \leq (W_1^{i,j})^2 \quad (\sigma' \leq 1 \text{ for softplus}) \\
\mathbb{E}_{\mathbf{x}} \left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \partial W_1^{i,j}}\right)^2 &= \mathbb{E}_{\mathbf{x}} (W_2^j \sigma'(W\mathbf{x})^j + W_2^j \sigma''(W_1\mathbf{x})^j W_1^{i,j} \mathbf{x}_i)^2 \\
&\leq (W_2^j)^2 + (W_2^j \sigma''(W_1\mathbf{x})^j W_1^{i,j})^2 \quad (\text{Using } \mathbf{x} \sim \mathcal{N}(0, I)) \\
&\leq (W_2^j)^2 + \beta^2 (W_2^j W_1^{i,j})^2 \quad (\sigma'' \leq \beta) \\
\mathbb{E}_{\mathbf{x}} \left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \partial W_1^{m,j}}\right)^2 &\leq \beta^2 (W_2^j W_1^{i,j})^2 \quad (\forall m \neq i)
\end{aligned}$$

Computing the Frobenius norm, we have

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}} \|\nabla_\theta \nabla_{\mathbf{x}} f(\mathbf{x})\|_F^2 &= \sum_{i,j,k} \mathbb{E}_X \left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \partial W_1^{j,k}}\right) + \sum_{i,j} \mathbb{E}_X \left(\frac{\partial^2 f(\mathbf{x})}{\partial \mathbf{x}_i \partial W_2^j}\right) \\
&\leq \left(\sum_{j,k} (W_2^k)^2 + \sum_{j=i} \beta^2 (W_2^k W_1^{j,k})^2 + \sum_{j \neq i} \beta^2 (W_2^k W_1^{j,k})^2\right) + \sum_{i,j} (W_1^{i,j})^2 \\
&\leq \left(\sum_{k} (W_2^k)^2 + \sum_{i,j} (W_1^{i,j})^2\right) + \beta^2 \sum_{j,k} (W_2^k W_1^{j,k})^2 \\
&\leq \|\theta\|^2 + \beta^2 \phi^2(\theta)
\end{aligned}$$

Here, we use the fact that $\phi(\theta) = \sqrt{\sum_{j,k} (W_2^k W_1^{j,k})^2}$ is the 2-path-norm [29] for the purpose of characterizing generalization in neural networks. Finally, we re-write the above using square roots:

$$\begin{aligned}
\mathbb{E}_{\mathbf{x}} \|\nabla_\theta \nabla_{\mathbf{x}} f(\mathbf{x})\|_F &\leq \sqrt{\mathbb{E}_{\mathbf{x}} \|\nabla_\theta \nabla_{\mathbf{x}} f(\mathbf{x})\|_F^2} \\
&\leq \sqrt{\|\theta\|_2^2 + \beta^2 \phi^2(\theta)} \\
&\leq \|\theta\|_2 + \beta \phi(\theta)
\end{aligned}$$

□

B ADDITIONAL EXPERIMENTS

Additional data from §4.1 The specific hyperparameters that we use to a) train base models, and b) fine-tune these models, are shown in Table 2. These hyperparameters are chosen such that a minimum can be attained by the optimizer (SGD), as is required to verify the theory. The learning rates shown in the fine-tuning process for HELOC and Adult (indicated by a star in Table 2) are multiplied by a scalar that is dependent on the amount of noise added to the dataset. Intuitively, adding more noise causes a larger increase in loss, so we increase the learning rate used during fine-tuning in order to converge to the new minimum within 50 epochs. The alternative approach, used in the WHO experiments, is to use more epochs during fine-tuning, and instead start with a constant learning rate that is higher but decays more, such that we again see convergence.

Table 2: Details of (dataset specific) hyperparameters used in §4.1 fine-tuning. Base models are trained for a large enough number of epochs to ensure convergence to a minimum. Decay value indicates the amount of decay applied to the learning rate, and step size indicates the number of epochs between applying decay.

	WHO		HELOC		Adult	
	Base	Fine-Tune	Base	Fine-Tune	Base	Fine-Tune
Epochs	4000	1000	500	50	500	50
Learning Rate	0.5	0.1	0.5	0.01*	0.5	0.01*
Decay Value	0.9	0.9	0.95	0.95	0.95	0.95
Step Size	40	40	100	100	100	100

* these learning rates vary as described in this appendix.

Table 3: Accuracy for fine-tuning in the WHO dataset. Confidence intervals are the middle 50% of values. For softplus, $\gamma = 0.001$.

ReLU / γ=0		ReLU / γ=0.001		ReLU / γ=0.01		SP / $\beta = 10$		SP / $\beta = 5$	
Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
99.1 ±0.1	92.8±0.5	98.9 ±0.2	92.9 ±0.4	96.5±0.2	92.7 ±0.3	97.2 ±0.2	93.2 ±0.2	95.3±0.1	91.8±0.3

Accuracy data for WHO in the fine-tuning experiments is shown in Table 3. Larger weight decay values and lower model curvatures do reduce train accuracy, as these modifications make it more difficult to overfit the model in reaching an exact minimum. However, test accuracy remains very similar across techniques. For the experiments that add synthetic noise, the final accuracy after fine-tuning is consistent regardless of how much noise we add, up to a standard deviation of 0.1. The accuracy is in the mid-70% range for HELOC and the mid-80% range for Adult.

Figure 1 shows the effect of weight decay and curvature on the Adult dataset for the fine-tuning experiments (omitted in the main text). In the case of curvature, the trend towards higher gradient stability given less curvature is not very pronounced, which we attribute to the difficulty associated with training a model to an exact minimum. In the presence of larger noise, the closest minimum in the loss landscape during fine-tuning may also shift to a different minimum.

Additional data from §4.2 For a given dataset, the hyperparameters for the retraining experiments are fixed between base models and retrained models in this section. In the HELOC and Adult datasets, we train any particular model for 30 epochs with a learning rate of 0.2, and no decay. For the WHO dataset, we use 80 epochs, an initial learning rate 0.8, with decay value 0.8 and step size 10 (i.e. the learning rate is multiplied by 0.8 every 10 epochs).

Accuracy data for the retraining experiments is in Table 4 for WHO, and Figure 2 for HELOC and Adult. In the latter, the base model accuracy is essentially identical to the lower extreme on the x-axis (when very little/no noise has been added to the training data). In all datasets, while there exists some disparity between training and test data, this is much smaller than in the fine-tuning experiments. Importantly, test accuracies remain approximately constant across model types (e.g. lower or higher weight decay or curvature). We see a slight drop in accuracy for the Adult dataset using the softplus model with $\beta = 2$. Despite how similar the test accuracies of all shifted models are, we observe in the main text and in the following

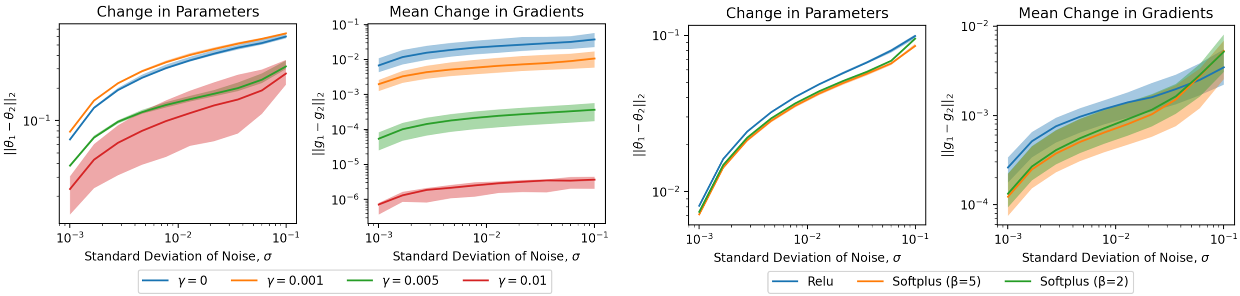

Figure 1: Effect of weight decay value (left) and model curvature (right) on gradient and parameter stability for the Adult dataset. All models use ReLU activation. Shifted models were trained via fine-tuning the original model. The x-axis is the size of the data shift, represented by the standard deviation σ of Gaussian $\mathcal{N}(0, \sigma^2)$ noise.

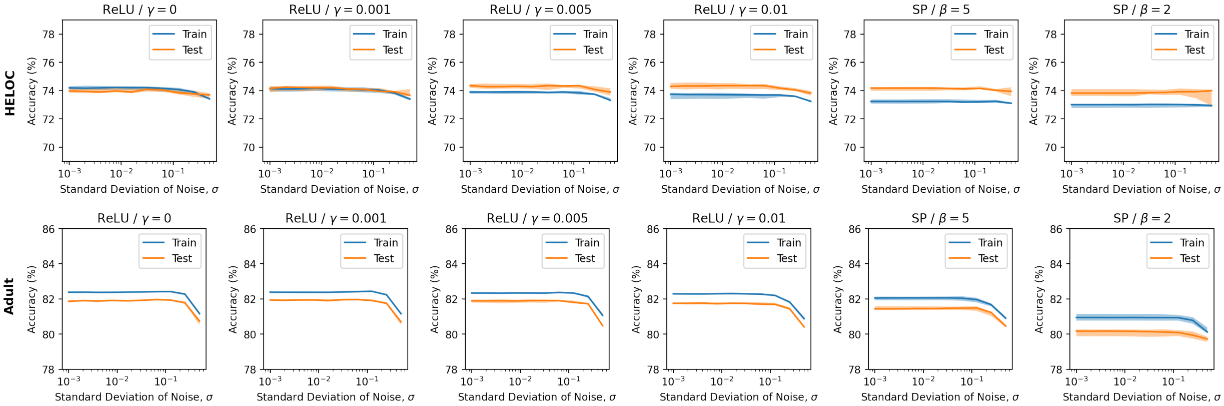

Figure 2: Accuracy for retraining (HELOC and Adult datasets). Confidence intervals are the middle 50% of values.

Table 4: Accuracy for of base models (upper row) and shifted models (lower row) in retraining experiments. Confidence intervals are the middle 50% of values. For softplus, $\gamma = 0.001$.

	ReLU / γ=0		ReLU / γ=0.001		ReLU / γ=0.01		SP / $\beta = 10$ large*		SP / $\beta = 5$	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Base	98.3 ±0.2	93.3 ±0.2	97.7±0.2	93.9±0.1	94.7±0.0	93.4±0.4	95.8 ±0.0	92.9 ±0.4	94.5±0.0	91.5±0.2
Shifted	98.0 ±0.2	93.7 ±0.4	97.4±0.2	93.7±0.4	94.7±0.2	92.5±0.3	95.7 ±0.0	93.2±0.2	94.7 ±0.1	92.6 ±0.2

figures that gradient stability across test inputs can still vary greatly, independent of test accuracy.

The left and right sections of Figure 3 show how parameters change given different weight decays and curvatures, respectively, as the data shift grows. This illustrates the same trends that we see in the main text, namely, that increasing weight decay or reducing curvature results in a smaller change in parameters, while adding more noise to the data results in a larger change.

Figures 4 and 5 show the Top-5 consistency scores of HELOC and Adult, respectively, each for Saliency and SmoothGrad techniques. We show the same for LIME and K.SHAP in Figure 6 for HELOC and Figure 7 for Adult. Overall, SmoothGrad demonstrates consistently strong performance with respect to explanation stability, while LIME and K.SHAP show poorer performance with higher variability. For all explanation techniques, we see the same consistent trends with regards to weight decay and curvature, though the effect of each of these is occasionally less distinct in the cases of LIME and K.SHAP (i.e. when the explanation technique is inherently less stable, which we attribute mostly to the number of samples used in LIME, the countermeasures we propose are sometimes less effective).

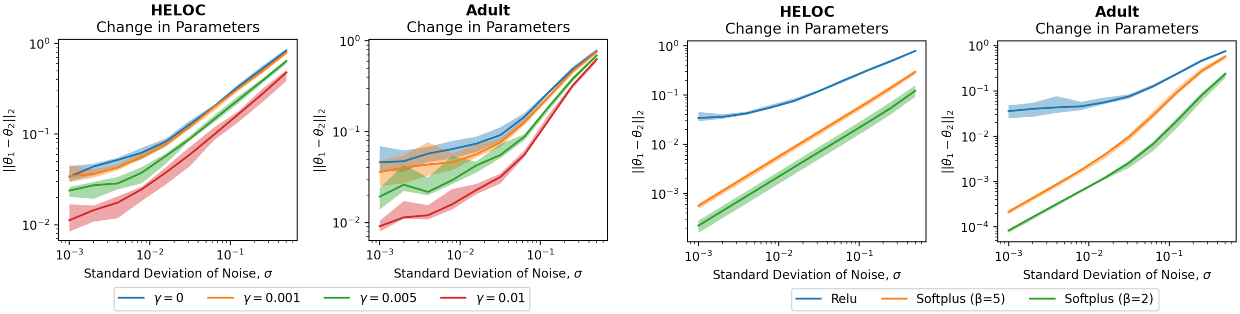

Figure 3: Effect of weight decay (left) and model curvature (right) on parameter stability when retraining on the HELOC and Adult datasets. The x-axis is the size of the data shift, represented by the standard deviation σ of $\mathcal{N}(0, 1)$ noise.

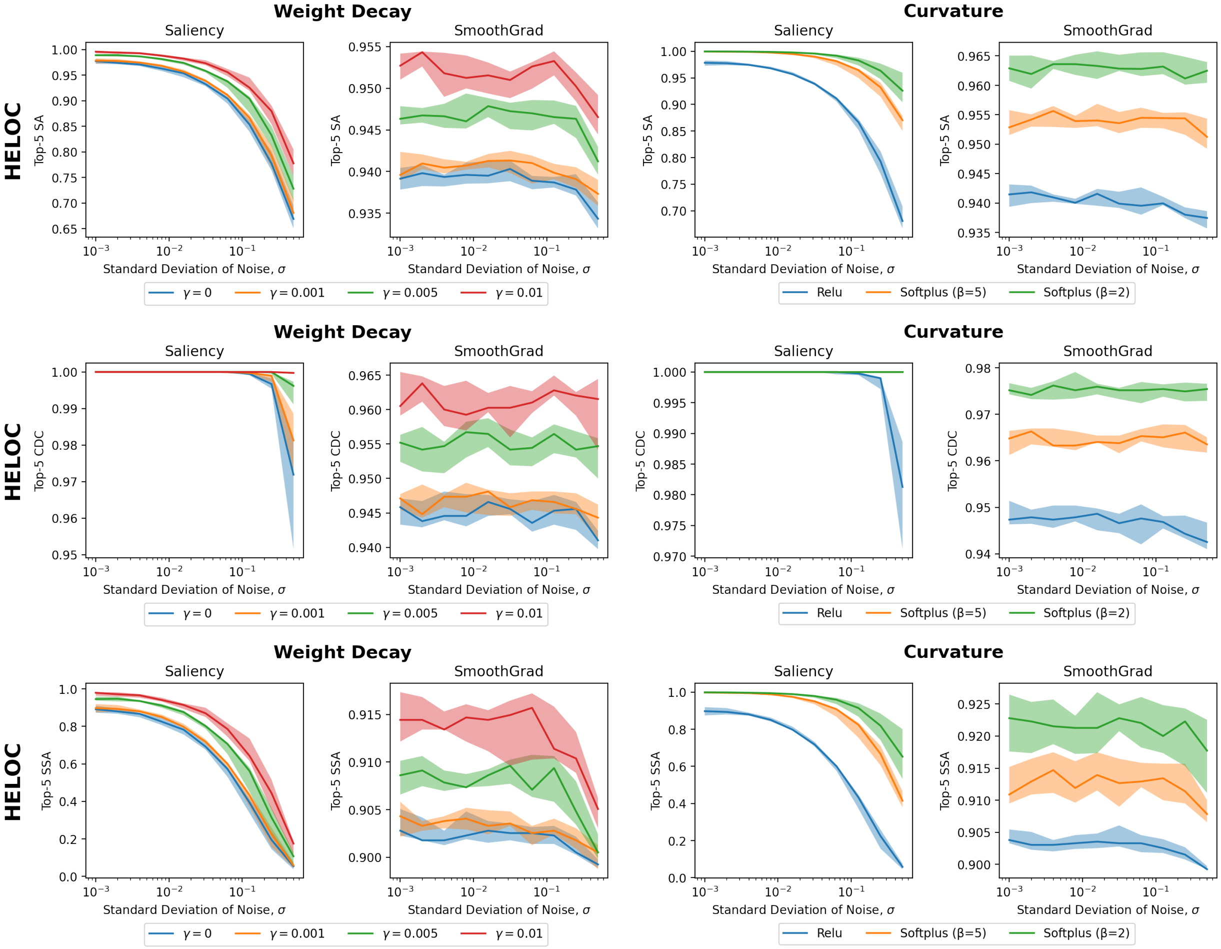

Figure 4: Top-5 consistency. From the top, HELOC SA, HELOC CDC, HELOC SSA (the graph for HELOC SA appears in the main text). Each row shows the effects of weight decay and curvature as data shift grows for salience and SmoothGrad top-5 metrics. Confidence intervals represent the middle 50% of values.

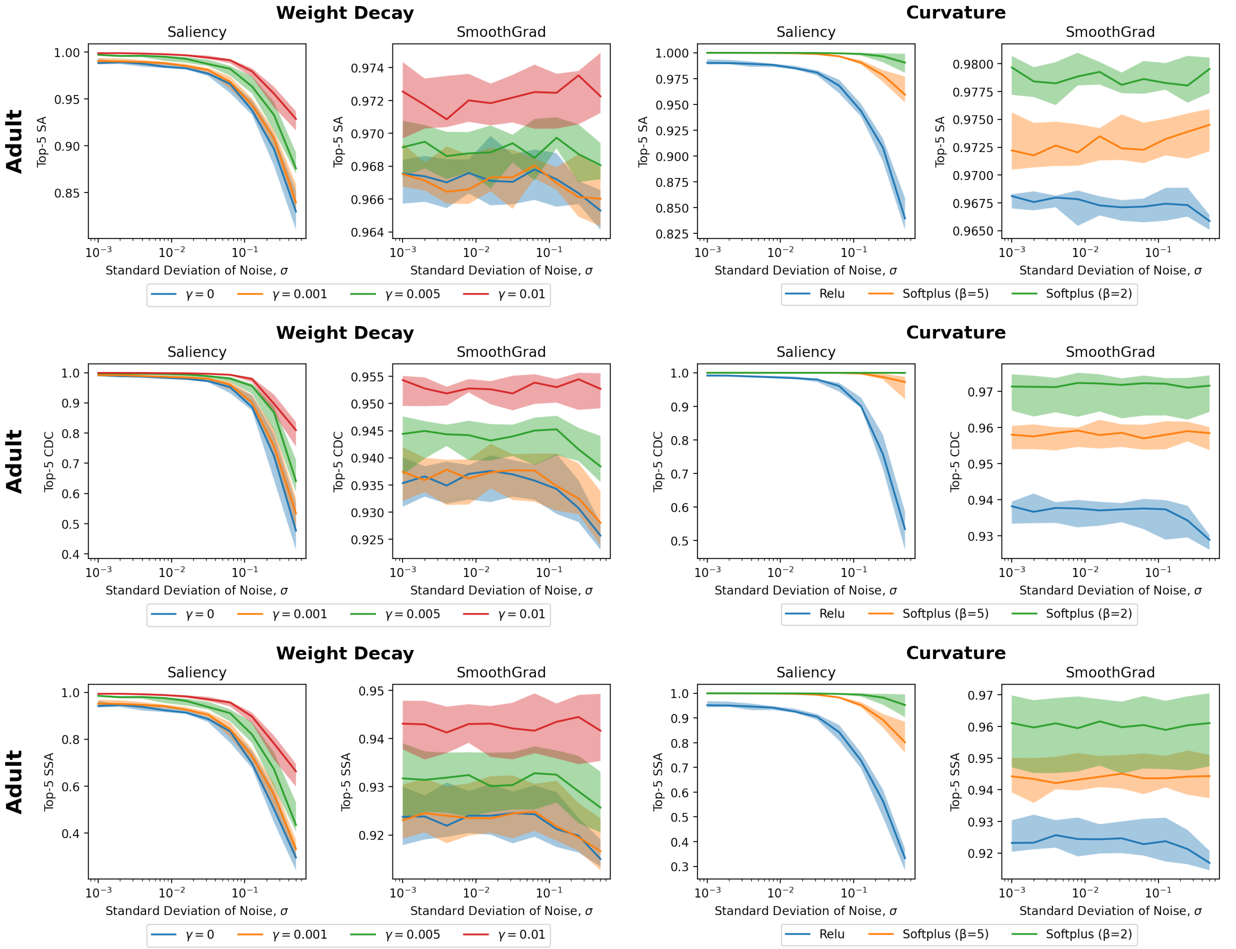

Figure 5: Top-5 consistency. From the top, Adult SA, Adult CDC, Adult SSA. Each row shows the effects of weight decay and curvature as data shift grows for salience and SmoothGrad top-5 metrics. Confidence intervals represent the middle 50% of values.

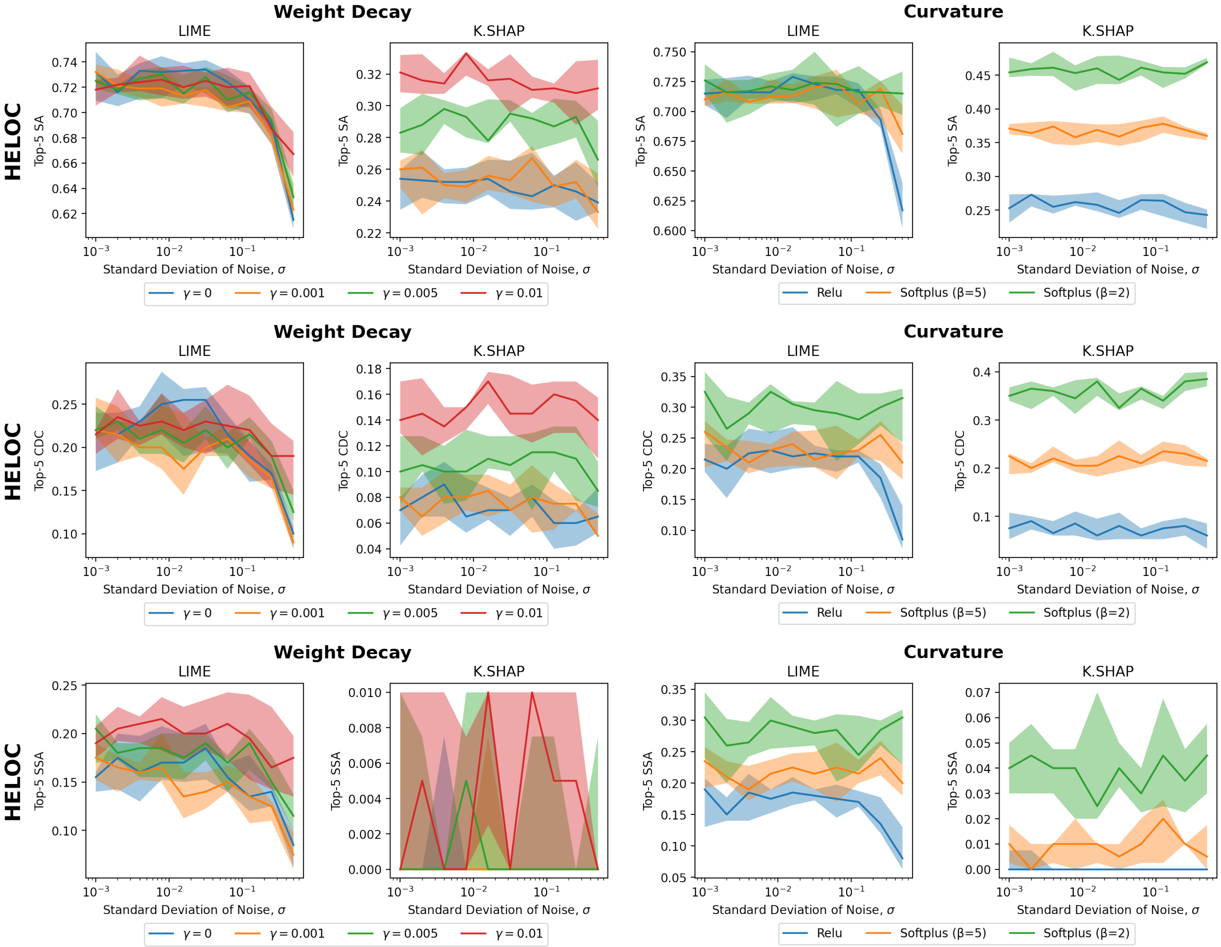

Figure 6: Top-5 consistency. From the top, HELOC SA, HELOC CDC, HELOC SSA. Each row shows the effects of weight decay and curvature as data shift grows for LIME and K.SHAP top-5 metrics. Confidence intervals represent the middle 50% of values.

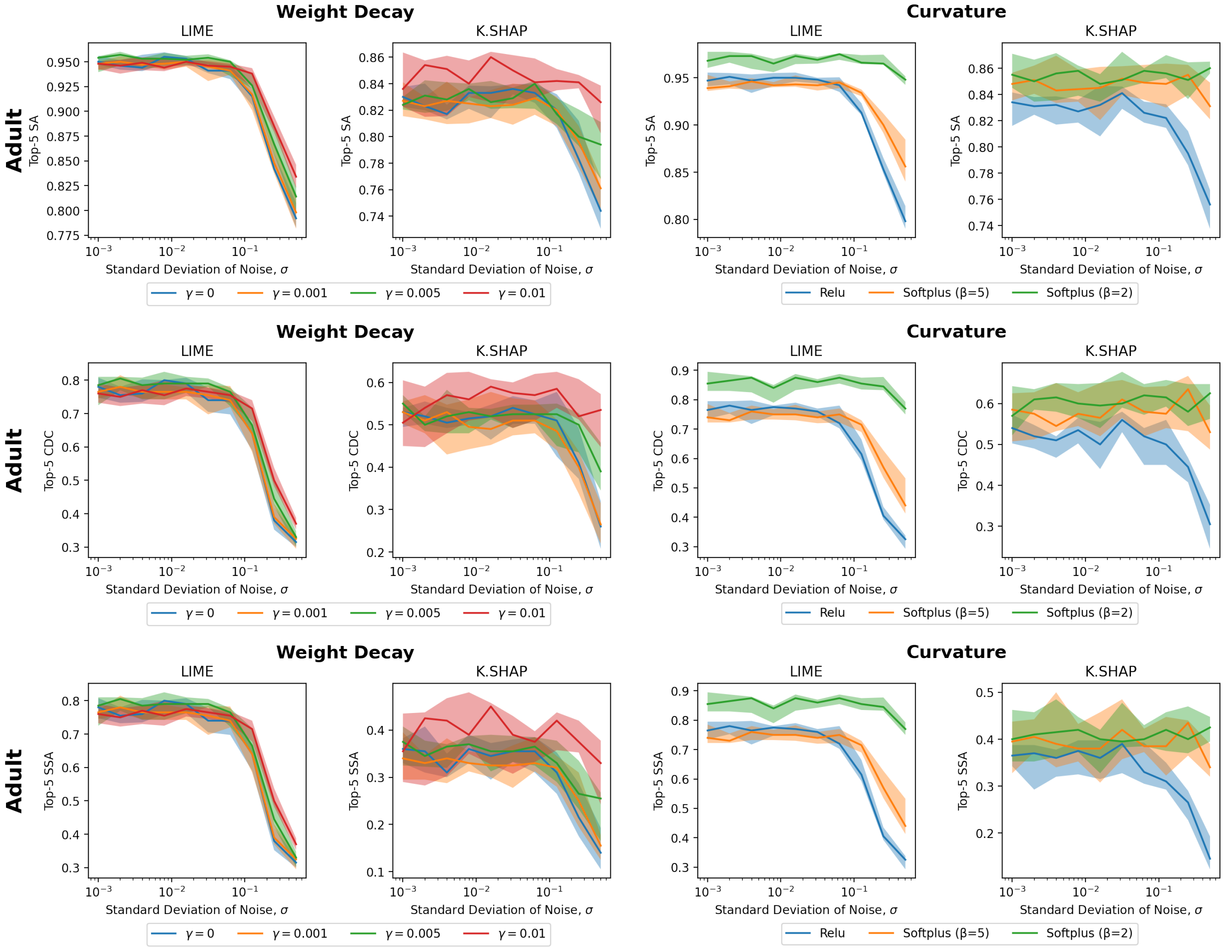

Figure 7: Top-5 consistency. From the top, Adult SA, Adult CDC, Adult SSA. Each row shows the effects of weight decay and curvature as data shift grows for LIME and K.SHAP top-5 metrics. Confidence intervals represent the middle 50% of values.

Table 5: Explanation Stability for WHO

		ReLU ($\gamma = 0$)	ReLU (γ=0.001)	ReLU (γ=0.01)	SP ($\beta = 10$)	SP ($\beta = 5$)
SA	Salience	0.63±0.01	0.64±0.01	0.73±0.01	0.72±0.02	0.77±0.02
	SmoothGrad	0.94±0.00	0.94±0.00	0.95±0.00	0.95±0.00	0.95±0.00
	LIME	0.59±0.02	0.59±0.03	0.70±0.03	0.65±0.03	0.69±0.02
	SHAP	0.52±0.01	0.56±0.02	0.67±0.03	0.61±0.02	0.65±0.02
CDC	Salience	0.80±0.04	0.83±0.04	0.95±0.01	0.94±0.02	0.97±0.01
	SmoothGrad	0.94±0.00	0.94±0.01	0.95±0.00	0.95±0.00	0.96±0.00
	LIME	0.52±0.03	0.52±0.05	0.57±0.05	0.58±0.05	0.65±0.04
	SHAP	0.62±0.02	0.70±0.04	0.87±0.03	0.77±0.03	0.86±0.04
SSA	Salience	0.19±0.04	0.18±0.03	0.30±0.02	0.29±0.01	0.39±0.03
	SmoothGrad	0.91±0.00	0.91±0.00	0.91±0.00	0.91±0.00	0.92±0.00
	LIME	0.27±0.02	0.25±0.02	0.32±0.03	0.30±0.03	0.35±0.03
	SHAP	0.19±0.04	0.21±0.01	0.29±0.04	0.23±0.03	0.28±0.03

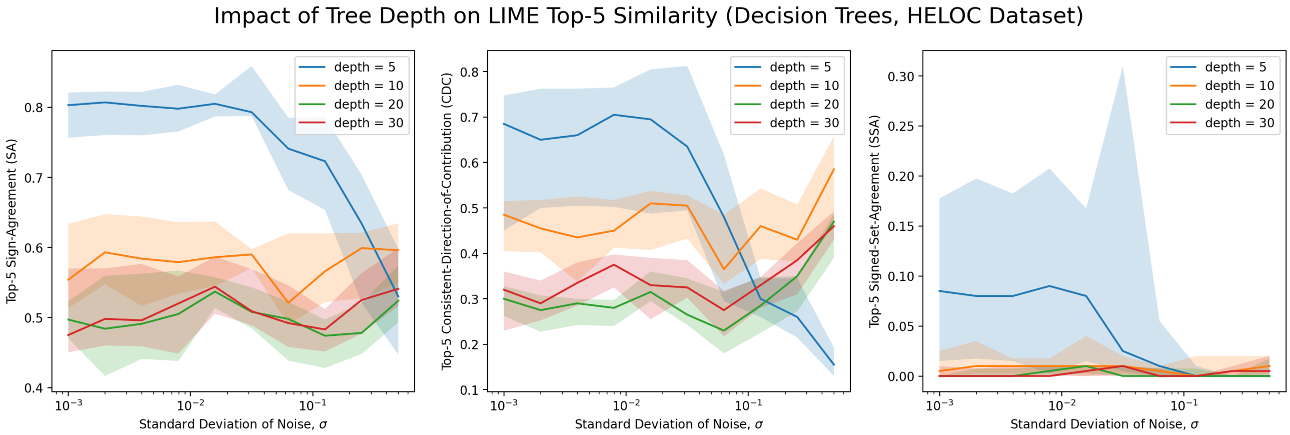

Figure 8: Various decision tree depths for the HELOC dataset. Similarity of LIME explanations between base models (original dataset) and retrained models (averaged across 10 random noise seeds for each value of σ). Left to right show Top-5 SA, CDC, and SSA metrics.

Table 5 shows the explanation stability scores for various metrics and explanation techniques for the WHO dataset. We make the following observations: first, for salience and SHAP (and to a lesser extent, LIME), using a weight decay of 0.01 is much more stable than using smaller weight decays. SmoothGrad is already very stable, so increasing the weight decay has no effect. Second, lowering the model curvature greatly increases the stability of CDC and SSA across explanation techniques, but has less effect on SA. This trend suggests that increasing the model curvature has limited impact on what features are in the top-K (i.e., what SA measures), but has a large impact on keeping the same sign for influential features (even though they might not be in the top K), which is what CDC measures.

Additional experiments with tree-based models Our theoretical results primarily focus on differentiable models, which encompass popular classes like neural networks, though the trends we have identified may not be universally applicable to all model classes. That said, it is worth noting that prior works analyzing the behavior and stability of ML models and explanation methods also typically rely on linear or differentiable models [9, 33]. We show that differentiable models with softplus activation functions and higher weight decay values, both of which can be viewed as a way to limit the functional complexity of the model, tend to promote explanation stability. Thus, we posit that explanations of tree-based models could similarly be made more robust by imposing analogous limits on their complexity.

We empirically evaluate explanation stability when retraining decision trees, random forests, and XGBoost models, controlling for tree depth in the first two cases and the λ regularization term in XGBoost. Higher λ values indicate greater regularization and lower complexity. We evaluated the stability of post-hoc explanations on the HELOC dataset with synthetic noise $\sim \mathcal{N}(0, \sigma^2)$, conducting empirical evaluations using LIME and SHAP. Please refer to §4 for detailed information on the setup of synthetic noise experiments. Interestingly, our findings show mixed results. While larger values of λ (corresponding to higher regularization constants) or smaller tree depths tend to yield more robust explanations in some

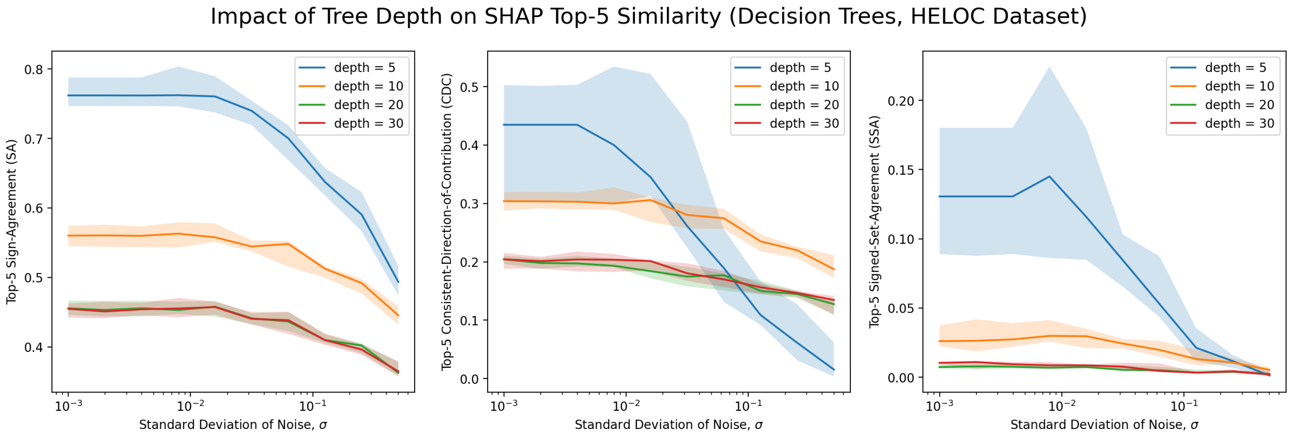

Figure 9: Various decision tree depths for the HELOC dataset. Similarity of SHAP explanations between base models (original dataset) and retrained models (averaged across 10 random noise seeds for each value of σ). Left to right show Top-5 SA, CDC, and SSA metrics.

cases, the outcomes vary for different explanation methods and dataset shifts.

Figures 8 and 9 demonstrate the impact of decision tree depth on LIME and SHAP, respectively, for Top-5 SA, CDC, and SSA metrics. Decision trees with lower depth exhibit improved explanation stability for smaller noise values, though the effects can break down as the dataset shifts by larger amounts. While these results support our findings for neural networks to a moderate degree, further research is required to investigate the behaviour that occurs for larger noise values (recall that our theory applies to small dataset shifts).

Figures 10 and 11 demonstrate the impact of random forest tree depth on LIME and SHAP, respectively, for Top-5 SA, CDC, and SSA metrics. Observe how the explanation stability of random forests is in general much higher than for decision trees– this would support the idea that the ensemble approach of random forests, which reduces functional diversity and complexity through averaging across multiple decision trees, indeed promotes explanation stability. However, the relative effects of depth are mixed. For LIME, at depth 5, explanations are relatively less stable, though for higher depths, explanations retained strong similarity until sufficiently large synthetic shifts were added ($\sigma > 0.1$). For SHAP, we observe the trends that we expect (higher depth reduces stability), though in the case of CDC this was reversed, albeit at very high stability values. Again, further research is required to investigate the behaviour that occurs for larger shifts.

Figures 12 and 13 demonstrate the impact of the XGBoost ℓ_2 regularization term λ on LIME and SHAP, respectively, for Top-5 SA, CDC, and SSA metrics. For LIME, the outcomes appear to be quite volatile, showing a relatively unpredictable range of optimal λ values for explanation stability. Interestingly, the added synthetic noise didn't significantly destabilize the explanations, and in some instances, it surprisingly seemed to bolster stability. This counter-intuitive finding underscores the need for further investigation to understand the underlying phenomena driving these trends (though LIME itself may be the significant driving factor). In contrast, for SHAP, the results are much clearer, revealing that higher regularization indeed promotes more stable explanations. Nonetheless, as anticipated, the stability decreases as the dataset undergoes more significant shifts. Notably, the performance of SHAP appeared to be superior to LIME.

We want to emphasize that these results are preliminary - more work is needed to generalize these results to larger experiments, to extend these results to real-world (not synthetic) data shifts, and to develop theory for when explanations are be more stable for tree-based models. Notably, we did not perform experiments with the WHO dataset (our real-world shift example) because the results of performing a single experiment were too noisy (with synthetic shifts, the random initialization of the noise can be used to average over multiple trials).

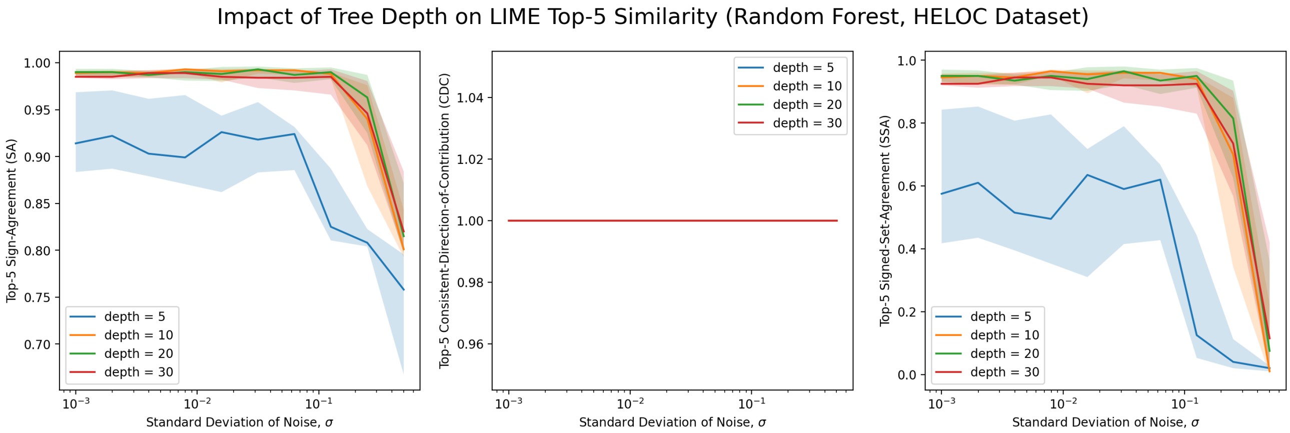

Figure 10: Various random forest depths for the HELOC dataset. Similarity of LIME explanations between base models (original dataset) and retrained models (averaged across 10 random noise seeds for each value of σ). Left to right show Top-5 SA, CDC, and SSA metrics.

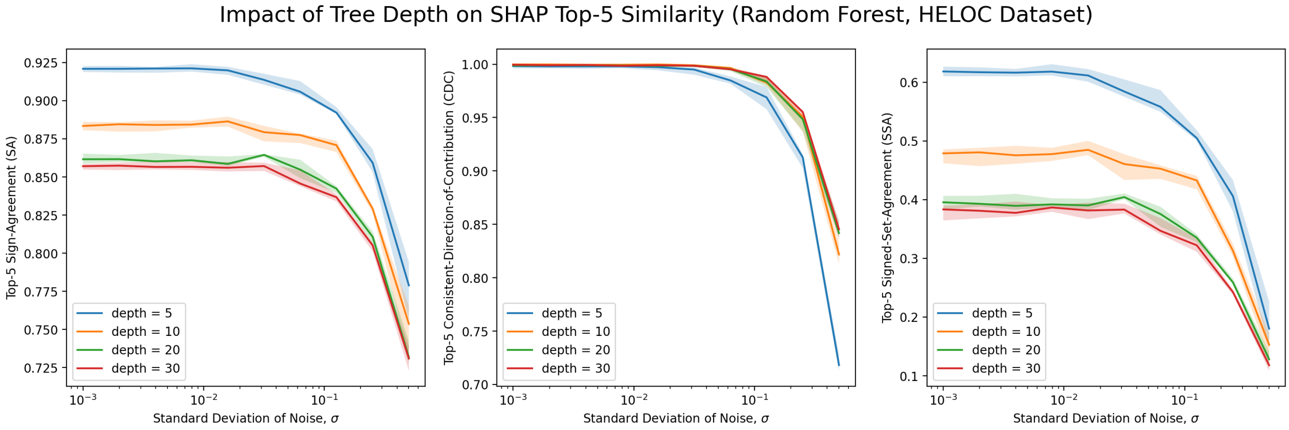

Figure 11: Various random forest depths for the HELOC dataset. Similarity of SHAP explanations between base models (original dataset) and retrained models (averaged across 10 random noise seeds for each value of σ). Left to right show Top-5 SA, CDC, and SSA metrics.

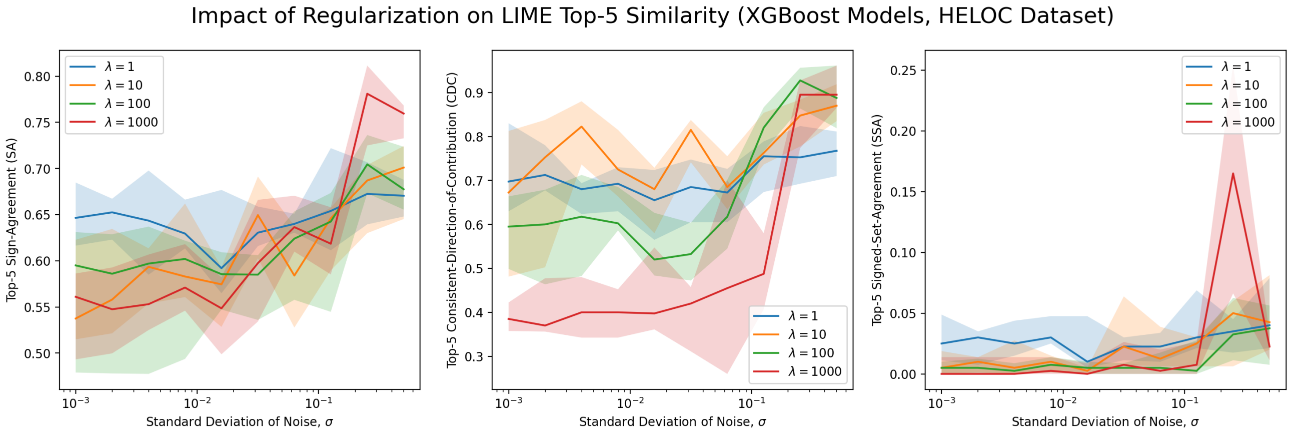

Figure 12: Various XGBoost λ values for the HELOC dataset. Similarity of LIME explanations between base models (original dataset) and retrained models (averaged across 10 random noise seeds for each value of σ). Left to right show Top-5 SA, CDC, and SSA metrics.

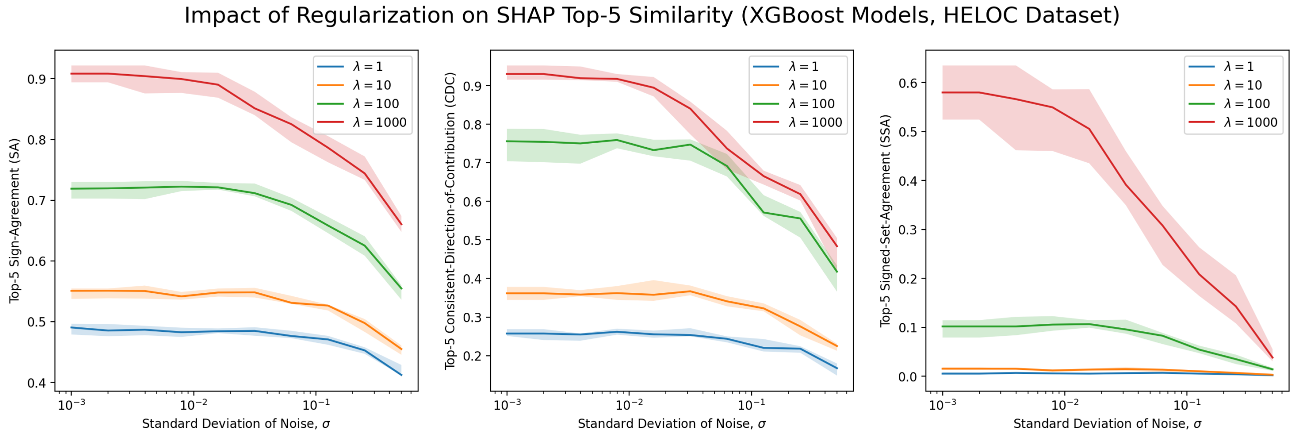

Figure 13: Various XGBoost λ values for the HELOC dataset. Similarity of SHAP explanations between base models (original dataset) and retrained models (averaged across 10 random noise seeds for each value of σ). Left to right show Top-5 SA, CDC, and SSA metrics.