A More discussions

A.1 Implementation details

Dataset selection. All of our experiments are performed on ImageNet-1k. All images come from validation set. We fix the random seed to be 2022. Then 100 classes are selected uniformly at random. For each class, an image from that class is selected uniformly at random.

Models. The models we use are pretrained models from `torchvison.models`. The `weights` parameter is set to `IMAGENET1K_V1`.

Feature transformation. For each image, we first crop it into size `(224,224,3)`. Then `quickshift` is used to segment the image into super-pixels. We use implementation from `scikit-image` and parameters are set as follows: `kernel_size=4, max_dist=200, ratio=0.2, random_seed=2023`. The setting we adopt is the same as the default setting in LIME except that we fix random seed. By fixing random seed, for the same image, we can always get the same super-pixels so that instability is only due to randomness in computing explanations. For different images, they are still segmented in different ways.

Computing explanations. Our implementation is based on LIME's original implementation. We fix `hide_color=None` so that the average value of each super-pixel will be used as reference for it when that super-pixel is removed. `distance_metric` to determine weight is set to `l2` which is suggested for image data in LIME [22]. Default value of `alpha` in Ridge regression is 1 if not otherwise mentioned. For each image, its most probable label is inferred from model f. Then, explanation w.r.t. this label is computed for that image. Explanations under 10 different random seeds are computed for each image. The parameter `random_seed` in `LimeImageExplainer` and its `explain_instance` function is fixed to these random seeds.

A.2 Stability of LIME and GLIME

In Figure 7, top-1,5,10 and average Jaccard index are presented. The average Jaccard index the the average of top-k Jaccard index for $k = 1, \cdots, d$. Results are similar with Figure 4a. By reformulating, GLIME produces more stable explanations than LIME.

A.3 LIME and GLIME-BINOMIAL converges to the same limit

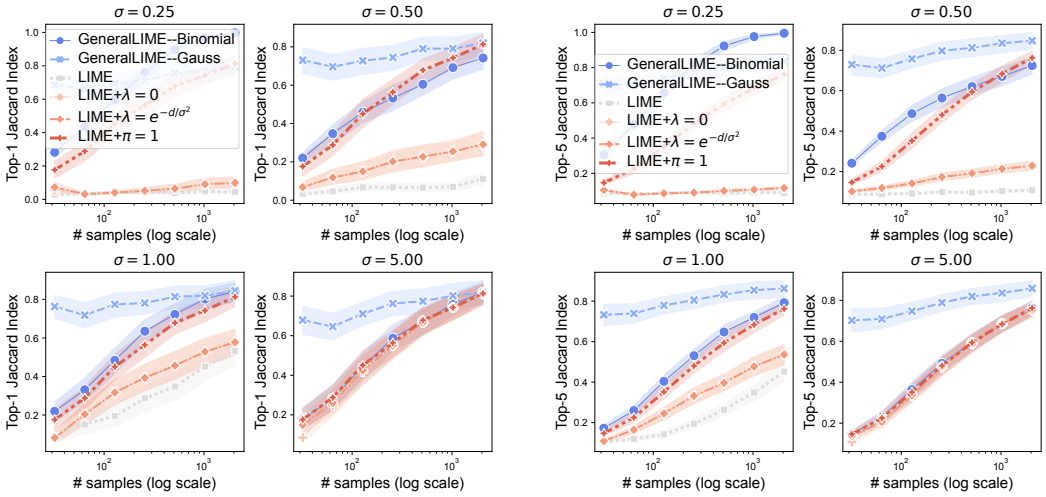
In Figure 8, we can observe the difference and correlation between the explanations produced by LIME and GLIME-BINOMIAL. As the sample size increases, LIME and GLIME-BINOMIAL become more similar and highly correlated. The difference between their explanations converges to zero rapidly when σ is very large (e.g., $\sigma = 5$). Since LIME converges very slow when σ is small, it may be intractable to sample until their difference fully converges. However, we can still observe that their correlation becomes stronger as more samples are used. This indicates that LIME and GLIME-BINOMIAL converge to the same limit as the sample size increases.

A.4 LIME explanation is different for different references.

Previous work [14] has pointed out that LIME is unstable with respect to references. As we argued in Section 4.2, this is due to sampling distribution of LIME is determined by reference chosen. In Figure 9, empirical evidence is presented. We choose six different references: black, white, red, blue, yellow image and the average value of removed super-pixel (which is the default setting for LIME). The average Jaccard index among explanations computed with these references is reported in Figure 9. Obviously, LIME is sensitive to references. Different references cause LIME to select different features as the most influential even when over 2000 samples are used. The top-1 Jaccard index is less than 0.7 when sample size is over 2000.

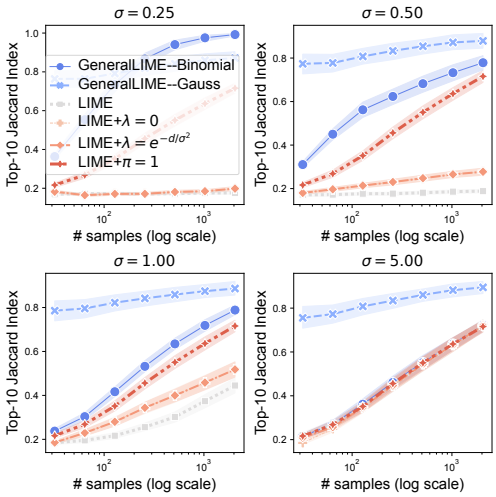
A.5 Local fidelity of GLIME

Figure 5 shows local fidelity of GLIME by sampling from ℓ_2 neighborhood $\{\mathbf{z} | \|\mathbf{z} - \mathbf{x}\|_2 \leq \epsilon\}$ of $\mathbf{x}$. Figure 10 and Figure 11 present local fidelity of GLIME by sampling from ℓ_1 neighborhood $\{\mathbf{z} | \|\mathbf{z} - \mathbf{x}\|_1 \leq \epsilon\}$ and ℓ_∞ neighborhood $\{\mathbf{z} | \|\mathbf{z} - \mathbf{x}\|_\infty \leq \epsilon\}$, respectively.

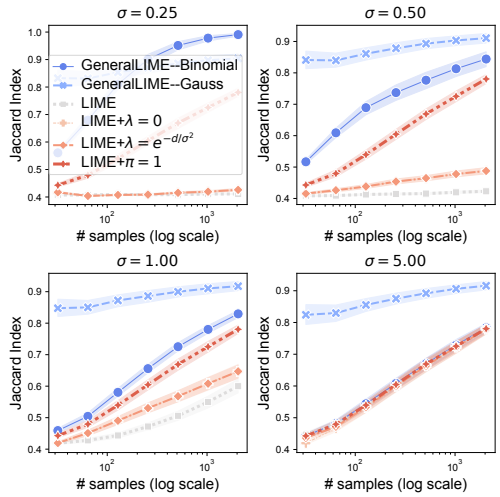


(a) Top-1 Jaccard index of different methods.

(b) Top-5 Jaccard index of different methods.



(c) Top-10 Jaccard index of different methods.



(d) Average Jaccard index of different methods.

Figure 7: Top-1,5,10 and average Jaccard index of different methods. Average Jaccard index is the average of top-1, $\cdots$, d Jaccard index.

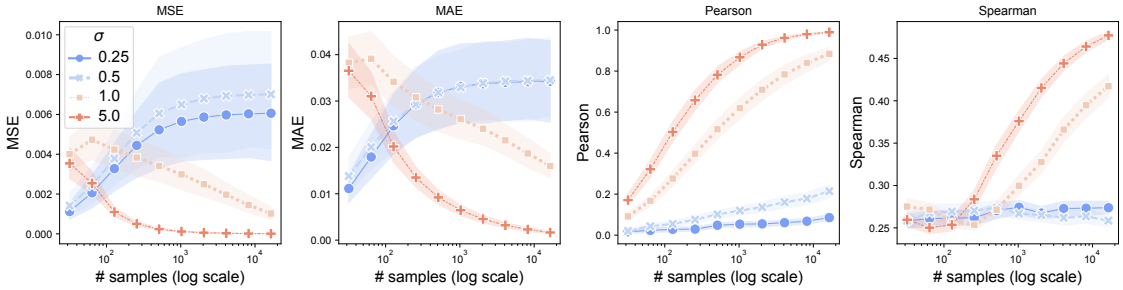


Figure 8: **Difference/correlation of LIME and GLIME-BINOMIAL explanations.** MSE, MAE are used to measure the difference between LIME and GLIME-BINOMIAL. Pearson and Spearman correlation are correlation measures. When the number of samples increases, their explanations become more similar. Difference/correlation converge faster when σ is larger.

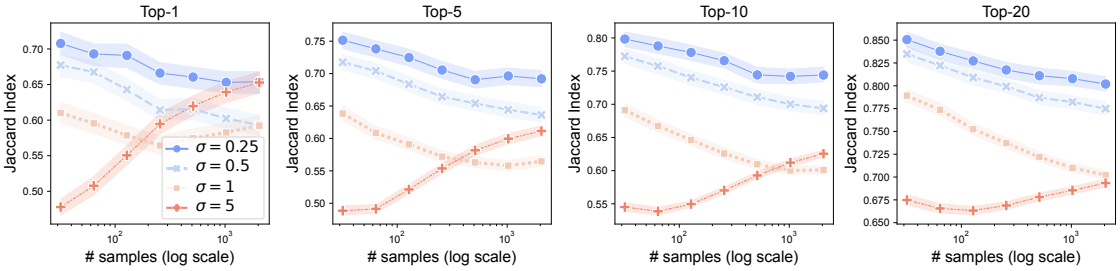


Figure 9: Top-K Jaccard index among explanations computed with different references. The top-1 Jaccard index is less than 0.7 even when sample size is over 2000.

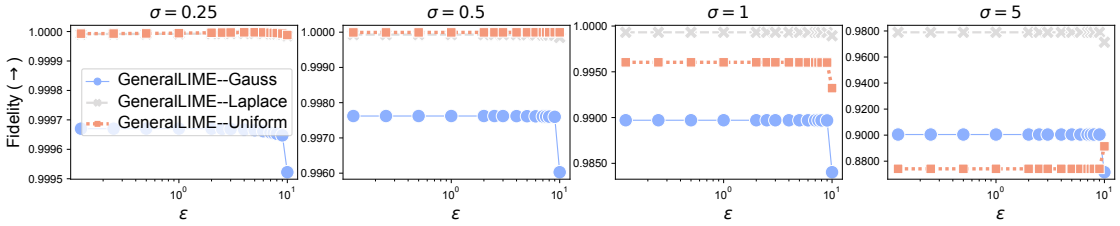


Figure 10: Local fidelity of GLIME in ℓ_1 neighborhood

By comparing Figure 5 and Figure 10, it can be observed that under the same σ, GLIME can explain local behaviors of f in ℓ_1 neighborhood with a larger radius than ℓ_2 neighborhood. This is due to the fact that under the same radius ϵ, $\{\mathbf{z}|\|\mathbf{z}-\mathbf{x}\|_2 \leq \epsilon\}$ defines a larger neighborhood than that of $\{\mathbf{z}|\|\mathbf{z}-\mathbf{x}\|_1 \leq \epsilon\}$.

Similarly, $\{\mathbf{z}|\|\mathbf{z}-\mathbf{x}\|_\infty \leq \epsilon\}$ defines a larger neighborhood than that of $\{\mathbf{z}|\|\mathbf{z}-\mathbf{x}\|_2 \leq \epsilon\}$. Under the same σ, the local fidelity peaks at a smaller radius ϵ for ℓ_∞ neighborhood than that of ℓ_2 neighborhood.

GLIME-LAPLACE generally has a better local fidelity than GLIME-GAUSS and GLIME-UNIFORM but for large ϵ, GLIME-GAUSS sometimes performs the best. Users should choose the sampling distribution according to the radius of the local neighborhood they aims to explain.

A.6 Previous methods GLIME unifies

KernelSHAP [19]. KernelSHAP is essentially LIME+Shapley value. LIME use a linear explnation model to locally approximate f. KernelSHAP seeks for a linear explanation model that satisfies axioms of Shapley values: local accuracy, missingness and consistency [19]. This is achieved by choosing the loss function $\ell(\cdot,\cdot)$, weighting function $\pi(\cdot)$ and regularization term R. The LIME choices for these parameters violate local accuracy and/or consistency [19]. However, the KernelSHAP choices for these parameters are proved to satisfiy these axioms (see Theorem 2 in [19]).

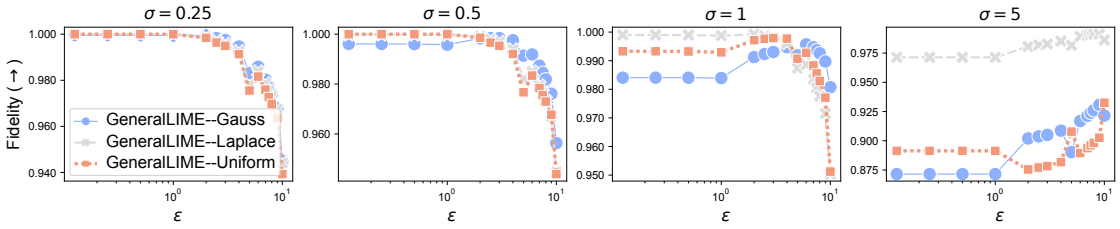
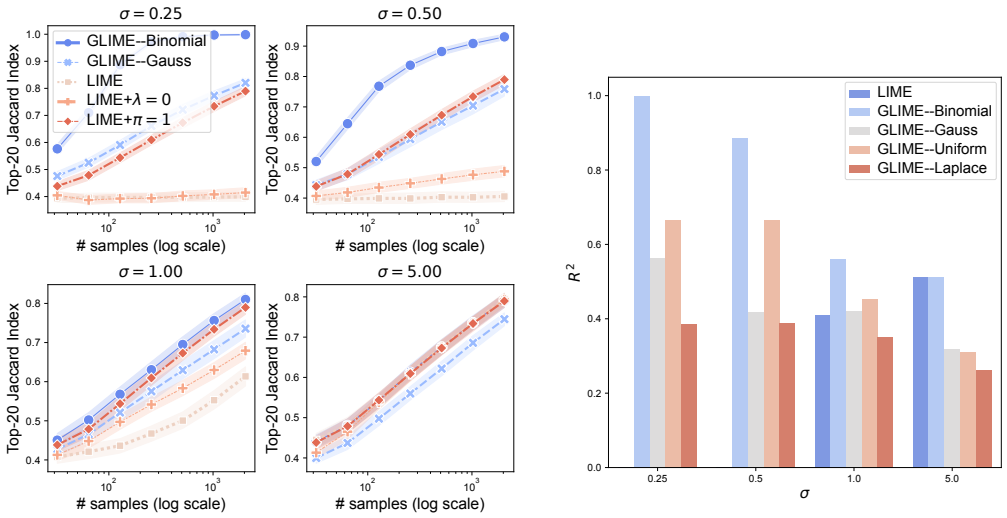


Figure 11: Local fidelity of GLIME in ℓ_∞ neighborhood



(a) **Stability of different methods (tiny Swin-Transformer).** Top-20 Jaccard index is reported. LIME+$\lambda = 0$ and LIME+$\pi = 1$ are LIME without regularization and weighting respectively. LIME is unstable when σ is small while GLIME is more stable for different σ. Without weighting or regularization, LIME becomes much more stable when σ is small. Regularization and weighting show little effect on stability of LIME when σ is large.

(b) **R^2 of LIME and several methods produced by GLIME with different sampling distributions (tiny Swin-Transformer).** 2048 samples are used to compute explanation and corresponding R^2 for each image and each method. LIME shows almost zero R^2 when $\sigma = 0.25, 0.5$. This indicates that LIME produce almost zero explanation. R^2 of LIME is generally lower than that of GLIME which demonstrates that GLIME improves local fidelity of LIME.

Figure 12: GLIME significantly improves stability and local fidelity upon LIME for different σ.

Gradient [2, 27]. It returns ∇f to measure the influence of each feature under infinitesimal perturbation [2, 27].

SmoothGrad [28]. Vanilla gradient explanations are shown to be noisy. SmoothGrad proposes to smooth out noise by averaging gradients at local neighborhood [28]. The feature importance is thus $\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2\mathbf{I})}[\nabla f(\mathbf{x}+\epsilon)]$.

DLIME [37]. Instead of random sampling, DLIME aims to design a deterministic way to obtain samples. DLIME first uses agglomerative Hierarchical Clustering to group the training data together and K-Nearest Neighbour to select the relevant cluster of the explained instance. Explanation is generated by compute a linear model on data points in the cluster found.

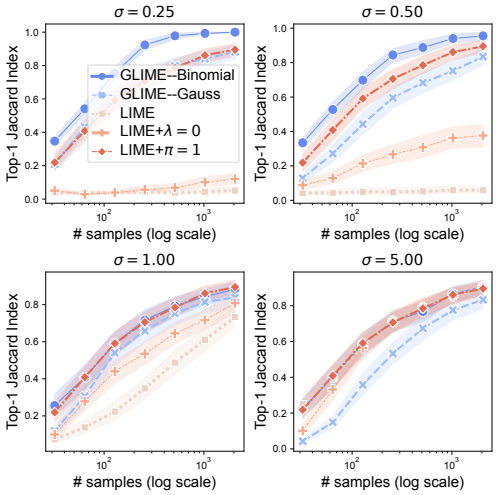
ALIME [24]. ALIME utilizes an auto-encoder to weight samples. An auto-encoder $\mathcal{AE}(\cdot)$ is first trained on training data. n nearest points of $\mathbf{x}$ are sampled from training dataset. The distances between samples and explained instance $\mathbf{x}$ are measured by ℓ_2 distance between their embeddings obtained by applying the auto-encoder $\mathcal{AE}(\cdot)$. For a sample $\mathbf{z}$, its distance with $\mathbf{x}$ is $\|\mathcal{AE}(\mathbf{z}) - \mathcal{AE}(\mathbf{x})\|_2$ and its weight is $\exp(-\|\mathcal{AE}(\mathbf{z}) - \mathcal{AE}(\mathbf{x})\|_2)$. Finally, the explanation is obtained by solving a weighted Ridge regression problem.

A.7 Results on tiny Swin-Transformer [18]

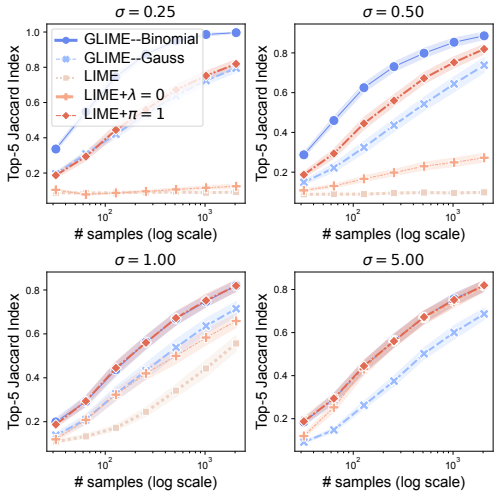
Results on tiny Swin-Transformer is similar with results on ResNet18 which further confirms that GLIME improves stability and local fidelity over LIME.

A.8 Comparing GLIME with ALIME

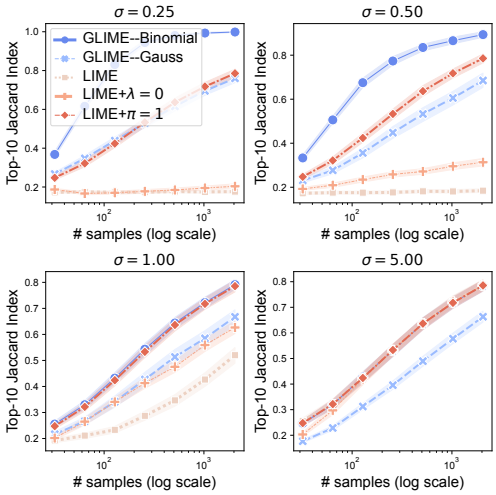
Although ALIME [24] improves stability and local fidelity over LIME, GLIME still outperforms ALIME. One major difference between ALIME and LIME is that ALIME uses an encoder to encode samples into embedding space and compute their distance with the input to be explained in embedding space $\|\mathcal{AE}(\mathbf{z}) - \mathcal{AE}(\mathbf{x})\|_2$ while LIME uses a binary vector $\mathbf{z} \in \{0,1\}^d$ to represent a sample, and



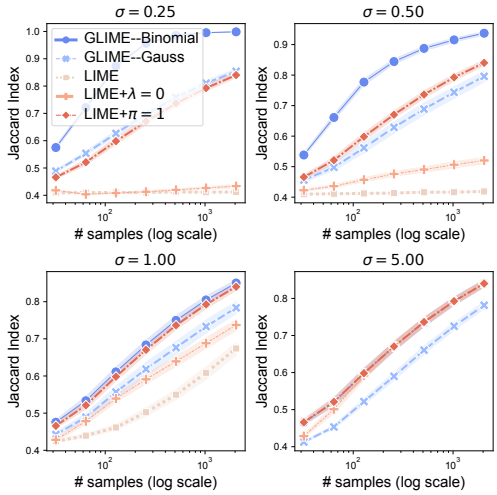
(a) Top-1 Jaccard index of different methods (tiny Swin-Transformer).



(b) Top-5 Jaccard index of different methods (tiny Swin-Transformer).



(c) Top-10 Jaccard index of different methods (tiny Swin-Transformer).



(d) Average Jaccard index of different methods (tiny Swin-Transformer).

Figure 13: Top-1,5,10 and average Jaccard index of different methods. Average Jaccard index is the average of top-1, $\cdots$, d Jaccard index (tiny Swin-Transformer).

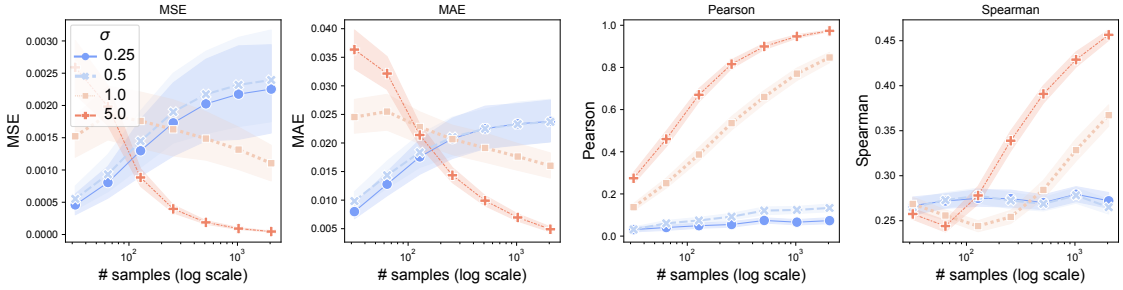


Figure 14: **Difference/correlation of LIME and GLIME-BINOMIAL explanations (tiny Swin-Transformer).** MSE, MAE are used to measure the difference between LIME and GLIME-BINOMIAL. Pearson and Spearman correlation are correlation measures. When the number of samples increases, their explanations become more similar. Difference/correlation converge faster when σ is larger.

use $\|\mathbf{1} - \mathbf{z}\|_2$ as the distance between the sample and the explained input. ALIME use distance in embedding space to weight samples. Therefore, if the samples generated by ALIME is distant from $\mathbf{x}$, sample weights may still be very small and may cause instability problem.

We conduct experiments to compare GLIME and ALIME. We utilize the VGG16 model provided by the repository[2] as the encoder in ALIME for our experiments on ImageNet. Table 1 shows the results of the experiments. It can be observed that although ALIME has improved stability compared to LIME, the improvement is still not as significant as that of GLIME, especially when σ is small or the sample size is small.

Table 1: **Top-20 Jaccard Index of GLIME-BINOMIAL, GLIME-GAUSS and ALIME.** GLIME-BINOMIAL and GLIME-GAUSS is much more stable than ALIME, especially when σ is small or only limited samples are available.

# samples		128	256	512	1024
$\sigma = 0.25$	GLIME-BINOMIAL	0.952	0.981	0.993	0.998
	GLIME-GAUSS	0.872	0.885	0.898	0.911
	ALIME	0.618	0.691	0.750	0.803
$\sigma = 0.5$	GLIME-BINOMIAL	0.596	0.688	0.739	0.772
	GLIME-GAUSS	0.875	0.891	0.904	0.912
	ALIME	0.525	0.588	0.641	0.688
$\sigma = 1$	GLIME-BINOMIAL	0.533	0.602	0.676	0.725
	GLIME-GAUSS	0.883	0.894	0.908	0.915
	ALIME	0.519	0.567	0.615	0.660
$\sigma = 5$	GLIME-BINOMIAL	0.493	0.545	0.605	0.661
	GLIME-GAUSS	0.865	0.883	0.898	0.910
	ALIME	0.489	0.539	0.589	0.640

A.9 Experiment Results on IMDb

Experiments on text data are conducting on text dara by utilizing the DistilBERT model. 100 data points are selected from the IMDb dataset as inputs for explanation. In our experiments, we compare the performance of GLIME-BINOMIAL and LIME, and the Jaccard Index results are presented in Figure 15. Our findings indicate that GLIME-BINOMIAL exhibits significantly higher stability than LIME across various values of σ and sample sizes. Particularly, when σ is small, GLIME-BINOMIAL demonstrates a substantial improvement in stability compared to LIME.

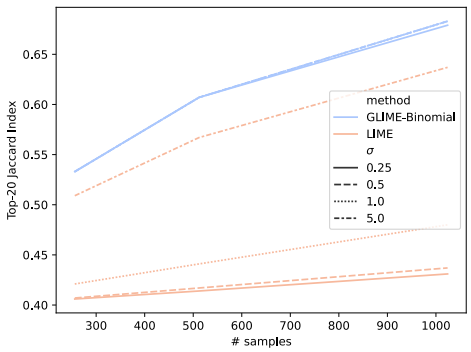
B Proofs

B.1 Equivalent GLIME formulation without $\pi(\cdot)$

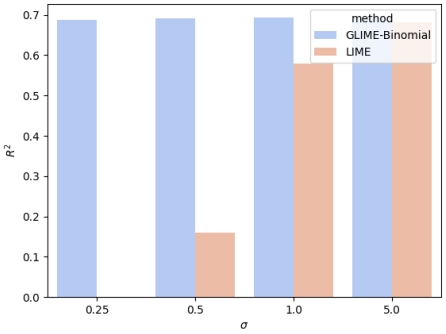
When $n \to \infty$, the regularization in Equation 2 could be omitted and the problem to solve is

$$
\begin{aligned}
\mathbf{w}^{\text{GLIME}} &= \arg\min_{\mathbf{v}} \mathbb{E}_{\mathbf{z}' \sim \mathcal{P}}[\pi(\mathbf{z}')\ell(f(\mathbf{z}), g(\mathbf{z}'))] + \lambda R(\mathbf{v}) \\
&= \arg\min_{\mathbf{v}} \int_{\mathbb{R}^d} \pi(\mathbf{z}')\ell(f(\mathbf{z}), g(\mathbf{z}'))\mathcal{P}(\mathbf{z})\mathrm{d}\mathbf{z} + \lambda R(\mathbf{v}) \\
&= \frac{\arg\min_{\mathbf{v}} \int_{\mathbb{R}^d} \ell(f(\mathbf{z}), g(\mathbf{z}'))\pi(\mathbf{z}')\mathcal{P}(\mathbf{z})\mathrm{d}\mathbf{z} + \lambda R(\mathbf{v})}{\int_{\mathbb{R}^d} \pi(\mathbf{u}')\mathcal{P}(\mathbf{u})\mathrm{d}\mathbf{u}} \\
&= \arg\min_{\mathbf{v}} \frac{\int_{\mathbb{R}^d} \ell(f(\mathbf{z}), g(\mathbf{z}'))\pi(\mathbf{z}')\mathcal{P}(\mathbf{z})\mathrm{d}\mathbf{z}}{\int_{\mathbb{R}^d} \pi(\mathbf{u}')\mathcal{P}(\mathbf{u})\mathrm{d}\mathbf{u}} + \frac{\lambda R(\mathbf{v})}{\int_{\mathbb{R}^d} \pi(\mathbf{u}')\mathcal{P}(\mathbf{u})\mathrm{d}\mathbf{u}} \\
&= \arg\min_{\mathbf{v}} \int_{\mathbb{R}^d} \ell(f(\mathbf{z}), g(\mathbf{z}'))\widetilde{\mathcal{P}}(\mathbf{z})\mathrm{d}\mathbf{z} + \frac{\lambda}{Z} R(\mathbf{v}) \quad \widetilde{\mathcal{P}}(\mathbf{z}) = \frac{\pi(\mathbf{z}')\mathcal{P}(\mathbf{z})}{Z}, Z = \int_{\mathbb{R}^d} \pi(\mathbf{u}')\mathcal{P}(\mathbf{u})\mathrm{d}\mathbf{u} \\
&= \arg\min_{\mathbf{v}} \mathbb{E}_{\mathbf{z}' \sim \widetilde{\mathcal{P}}}[\ell(f(\mathbf{z}), g(\mathbf{z}'))] + \frac{\lambda}{Z} R(\mathbf{v})
\end{aligned}
$$

[2] https://github.com/Horizon2333/imagenet-autoencoder



(a) **Stability of GLIME-BINOMIAL and LIME.** Top-20 Jaccard index is reported. GLIME is more stable for different σ while σ does not affect GLIME's stability. For LIME, it is more stable when σ is larger.



(b) **R^2 of LIME and GLIME-BINOMIAL with different sampling distributions.** 2048 samples are used to compute explanation and corresponding R^2 for each image and each method. LIME shows almost zero R^2 when $\sigma = 0.25$.

Figure 15: GLIME significantly improves stability and local fidelity upon LIME for different σ.

B.2 Equivalence between LIME and GLIME-BINOMIAL

For LIME, the probability that a sample $\mathbf{z}'$ with $\|\mathbf{z}'\|_0 = k$ could be sampled is $\frac{1}{2^d}$, so that

$$Z = \int_{\mathbb{R}^d} \pi(\mathbf{u}')\mathcal{P}(\mathbf{u})\mathrm{d}\mathbf{u} = \sum_{k=0}^{d} e^{(k-d)/\sigma^2} \frac{\binom{d}{k}}{2^d} = \frac{e^{-d/\sigma^2}}{2^d}(1+e^{1/\sigma^2})^d$$

Thus, we have

$$\widetilde{\mathcal{P}}(\mathbf{z}) = \frac{\pi(\mathbf{z}')\mathcal{P}(\mathbf{z})}{Z} = \frac{e^{(k-d)/\sigma^2}2^{-d}}{Z} = \frac{e^{k/\sigma^2}}{(1+e^{1/\sigma^2})^d}$$

Therefore, GLIME-BINOMIAL is equivalent to LIME.

B.3 LIME requires more samples to omit the influence of regularization.

Theorem B.1. *Suppose we have samples* $\{\mathbf{z}_i\}_{i=1}^n \sim Uni(\{0,1\}^d)$. *Let* $\alpha_0 = \mathbb{E}_{\mathbf{z}\sim Uni(\{0,1\}^d)}[\pi(\mathbf{z})\|\mathbf{z}\|_0/d]$. *For any* $t > 0, \delta \in (0,1)$, *if* $n \leq \frac{2\lambda t\alpha_0 - \sqrt{2\lambda t\alpha_0 \log\frac{1}{\delta}}}{2\alpha_0^2}$, *for LIME explanation, we have*

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}\pi(\mathbf{z}_i')\sum_j z_{ij}/d < \frac{\lambda}{n}t) \geq 1-\delta$$

Theorem B.2. *Suppose we have samples* $\{\mathbf{z}_i'\}_{i=1}^n \sim \mathcal{P}$ *such that* $\mathbf{z}_{i,j}', i = 1,\cdots,n, j = 1,\cdots,d$ *are sub-gaussian variables with mean* γ, *i.e.,* $\mathbb{E}[e^{(z_{i,j}')^2/\nu^2}] \leq 2, \forall t \in \mathbb{R}$. *Let* $\gamma = \mathbb{E}_{\mathbf{z}\sim\mathcal{P}}[\|\mathbf{z}\|_2^2/d]$. *For any* $t > 0, \delta \in (0,1)$, *if* $n \leq \frac{2c\lambda t\gamma - \sqrt{2cb^2\lambda t\gamma\log\frac{1}{\delta}}}{2c\gamma^2}$ *where* c *is an absolute constant, we have*

$$\mathbb{P}(\frac{1}{n}\sum_i \|\mathbf{z}_i\|_2^2/d < \frac{\lambda}{n}t) \geq 1-\delta$$

Corollary B.3. *Suppose we have samples* $\{\mathbf{z}_i\}_{i=1}^n \sim \tilde{\mathbb{P}}$. *Let* $\gamma = \mathbb{E}_{\mathbf{z}\sim\tilde{\mathbb{P}}}[\|\mathbf{z}\|_0/d]$. *For any* $t > 0, \delta \in (0,1)$, *if* $n \leq \frac{2\lambda t\gamma - \sqrt{2\lambda t\gamma\log\frac{1}{\delta}}}{2\gamma^2}$, *we have*

$$\mathbb{P}(\frac{1}{n}\sum_i \|\mathbf{z}_i\|_0/d < \frac{\lambda}{n}t) \geq 1-\delta$$

$\frac{1}{n}\sum_{i=1}^n \pi(\mathbf{z}_i')\sum_j z_{ij}/d$ can be regarded as the coefficient of $\|\mathbf{v}\|$ in the sum-of-square-term in Equation 1. Theorem B.1 shows the number of samples required for this coefficient to be large

than regularization strength λ. As will be proved $\alpha_0 = \frac{1}{2}\left(\frac{1+e^{-\frac{1}{\sigma^2}}}{2}\right)^{d-1}, \gamma \in (1/2, 1)$, comparing the results of LIME and GLIME-BINOMIAL, we can see that LUIME requires exponentially more samples than GLIME-BINOMIAL.

B.4 Proof of Theorem B.1

Proof. Let $y_i = \pi(\mathbf{z}'_i)\sum_j z_{ij}/d, \hat{y} = \frac{1}{n}\sum_i y_i$, then

$$\alpha_0 = \mathbb{E}[y_i] = \sum_{k=0}^{d} \frac{k}{d}\frac{\binom{d}{k}}{2^d} e^{\frac{k-d}{\sigma^2}} = \sum_{k=0}^{d} \frac{\binom{d-1}{k-1}}{2^d} e^{\frac{k-d}{\sigma^2}} = \frac{1}{2}\left(\frac{1+e^{-\frac{1}{\sigma^2}}}{2}\right)^{d-1}$$

For $n \le \frac{2\lambda t\alpha_0 - \sqrt{2\lambda t\alpha_0 \log\frac{1}{\delta}}}{2\alpha_0^2} \le \frac{\lambda t}{\alpha_0}$

$$\hat{y} > \frac{\lambda}{n}t \Longleftrightarrow \hat{y} - \alpha_0 > \frac{\lambda t}{n} - \alpha_0 > 0$$

Since $0 \le y_i \le 1$, thus by Hoeffding inequality, we have

$$\mathbb{P}(\hat{y} - \alpha_0 > \frac{\lambda t}{n} - \alpha_0) \le \exp\big(-2n(\frac{\lambda t}{n} - \alpha_0)^2\big)$$

Because

$$n < \frac{2\lambda t\alpha_0 - \sqrt{2\lambda t\alpha_0 \log\frac{1}{\delta}}}{2\alpha_0^2} \le \frac{4\lambda t\alpha_0 + \log\frac{1}{\delta} - \sqrt{8\lambda t\alpha_0 \log\frac{1}{\delta} + 16\lambda^2 t^2\alpha_0^2 + \log^2\frac{1}{\delta} - 16\lambda^2 t^2\alpha_0^2}}{4\alpha_0^2}$$

$$\Longleftrightarrow 2\alpha_0^2 n^2 - (4\lambda t\alpha_0 + \log\frac{1}{\delta})n + 2\lambda^2 t^2 \ge 0$$

$$\Longleftrightarrow \frac{2\alpha_0^2(n - \frac{\lambda t}{\alpha_0})^2}{n} \ge \log\frac{1}{\delta}$$

$$\Longleftrightarrow \exp\big(-2n(\frac{\lambda t}{n} - \alpha_0)^2\big) \le \delta$$

That is

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}\pi(\mathbf{z}'_i)\sum_j z_{ij}/d > \frac{\lambda}{n}t) \le \delta$$

□

B.5 Proof of Theorem B.2 and Corollary B.3

Proof of Theorem B.2. Let $w_i = \|\mathbf{z}_i\|_2^2/d, \hat{w} = \frac{1}{n}\sum_i w_i$. The distribution of w_i is as follows:

$$\mathbb{E}[w_i] = \mathbb{E}[\sum_j z_{i,j}^2/d] = \gamma, \mathbb{E}[\hat{w}] = \gamma$$

For $n \le \frac{2\lambda t\gamma - \sqrt{2\lambda t\gamma \log\frac{1}{\delta}}}{2\gamma^2} \le \frac{\lambda t}{\gamma}$

$$\hat{w} > \frac{\lambda}{n}t \Longleftrightarrow \hat{w} - \gamma > \frac{\lambda t}{n} - \gamma > 0$$

Since $0 \le y_i \le 1$, thus by Bernstein's inequality (refer to Theorem 2.8.1 in [34]), we have

$$\mathbb{P}(\hat{w} - \gamma > \frac{\lambda t}{n} - \gamma) \le \exp\big(-2\frac{nc}{b^2}(\frac{\lambda t}{n} - \gamma)^2\big)$$

where c is an absolute constant. Because

$$n < \frac{2\lambda t\gamma c - \sqrt{2cb^2\lambda t\gamma \log\frac{1}{\delta}}}{2\gamma^2 c} \leq \frac{4\lambda t\gamma c + b^2\log\frac{1}{\delta} - \sqrt{8cb^2\lambda t\gamma\log\frac{1}{\delta} + 16c^2\lambda^2t^2\gamma^2 + b^4\log^2\frac{1}{\delta} - 16c^2\lambda^2t^2\gamma^2}}{4\gamma^2 c}$$

$$\iff 2c\gamma^2 n^2 - (4\lambda t\gamma c + b^2\log\frac{1}{\delta})n + 2c\lambda^2t^2 \geq 0$$

$$\iff \frac{2\gamma^2 c(n - \frac{\lambda t}{\gamma})^2}{nb^2} \geq \log\frac{1}{\delta}$$

$$\iff \exp\big(-2\frac{nc}{b^2}(\frac{\lambda t}{n} - \gamma)^2\big) \leq \delta$$

That is

$$\mathbb{P}(\frac{1}{n}\sum_i \|\mathbf{z}_i\|_2^2/d < \frac{\lambda}{n}t) \geq 1 - \delta$$

□

Proof of Corollary B.3. Since $\mathbf{z}_i' \in \{0,1\}^d$, $\|\mathbf{z}_i'\|_2^2 = \|\mathbf{z}_i'\|_0$. We can prove that $\gamma = \mathbb{E}_{\mathbf{z}\sim\tilde{\mathbb{P}}}[\|\mathbf{z}\|_0/d] \in (\frac{1}{2}, 1)$

$$\mathbb{E}_{\mathbf{z}\sim\tilde{\mathbb{P}}}[\|\mathbf{z}\|_0/d] = \sum_{k=0}^{d}\frac{k}{d}\frac{\binom{d}{k}e^{k/\sigma^2}}{(1+e^{1/\sigma^2})^d} = \sum_{k=1}^{d}\frac{\binom{d-1}{k-1}e^{k/\sigma^2}}{(1+e^{1/\sigma^2})^d} = \frac{(1+e^{1/\sigma^2})^(d-1)e^{1/\sigma^2}}{(1+e^{1/\sigma^2})^d} = \frac{e^{1/\sigma^2}}{1+e^{1/\sigma^2}}$$

Since

$$\frac{e^{1/\sigma^2}}{1+e^{1/\sigma^2}} = \frac{1}{1+e^{-1/\sigma^2}} \in (\frac{1}{2}, 1),$$

we have $\gamma \in (\frac{1}{2}, 1)$. □

B.6 Proof of Theorem 4.1

Theorem B.4. *Suppose samples $\{\mathbf{z}_i'\}_{i=1}^n \sim Uni(\{0,1\}^d)$ are used to compute LIME explanation. For any $\epsilon > 0, \delta \in (0,1)$, if $n = \Omega((1+\lambda)^2\epsilon^{-2}d^5 2^{4d}e^{4/\sigma^2}\log(4d/\delta))$, we have $\mathbb{P}(\|\hat{\mathbf{w}}^{LIME} - \mathbf{w}^{LIME}\|_2 < \epsilon) \geq 1 - \delta$. $\mathbf{w}^{LIME} = \lim_{n\to\infty}\hat{\mathbf{w}}^{LIME}$.*

Proof. To compute LIME explanation with n samples, we solve

$$\hat{\mathbf{w}}^{\text{LIME}} = \arg\min_{\mathbf{v}} \frac{1}{n}\sum_{i=1}^{n}\pi(\mathbf{z}_i')(f(\mathbf{z}_i) - \mathbf{v}^\top\mathbf{z}_i')^2 + \frac{\lambda}{n}\|\mathbf{v}\|_2^2$$

Denote $L = \frac{1}{n}\sum_{i=1}^n \pi(\mathbf{z}_i')(f(\mathbf{z}_i) - \mathbf{v}^\top\mathbf{z}_i')^2 + \frac{\lambda}{n}\|\mathbf{v}\|_2^2$, set the gradient of L w.r.t $\mathbf{v}$ to zero, we have

$$-2\frac{1}{n}\pi(\mathbf{z}_i')(f(\mathbf{z}_i) - \mathbf{v}^\top\mathbf{z}_i')\mathbf{z}_i' + \frac{2}{n}\lambda\mathbf{v} = 0$$

$$\implies \hat{\mathbf{w}}^{\text{LIME}} = \left(\frac{1}{n}\sum_{i=1}^{n}\pi(\mathbf{z}_i')\mathbf{z}_i'(\mathbf{z}_i')^\top + \frac{\lambda}{n}\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\pi(\mathbf{z}_i')\mathbf{z}_i' f(\mathbf{z}_i)\right)$$

Denote $\mathbf{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n \pi(\mathbf{z}_i')\mathbf{z}_i'(\mathbf{z}_i')^\top + \frac{\lambda}{n}, \mathbf{\Gamma}_n = \frac{1}{n}\sum_{i=1}^n \pi(\mathbf{z}_i')\mathbf{z}_i' f(\mathbf{z}_i), \mathbf{\Sigma} = \lim_{n\to\infty}\mathbf{\Sigma}_n, \mathbf{\Gamma} = \lim_{n\to\infty}\mathbf{\Gamma}_n$, then

$$\hat{\mathbf{w}}^{\text{LIME}} = \mathbf{\Sigma}_n^{-1}\mathbf{\Gamma}_n, \mathbf{w}^{\text{LIME}} = \mathbf{\Sigma}^{-1}\mathbf{\Gamma},$$

To prove concentration of $\hat{\mathbf{w}}^{\text{LIME}}$, we follows the proofs in [8]: (1) We first prove the concentration of $\mathbf{\Sigma}_n$ (2) then bound $\|\mathbf{\Sigma}^{-1}\|_F^2$ (3) then prove the concentration of $\mathbf{\Gamma}_n$ (4) and finally use the following inequality:

$$\|\mathbf{\Sigma}_n^{-1}\mathbf{\Gamma}_n - \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\| \leq 2\|\mathbf{\Sigma}^{-1}\|_F\|\mathbf{\Gamma}_n - \mathbf{\Gamma}\|_2 + 2\|\mathbf{\Sigma}^{-1}\|_F^2\|\mathbf{\Gamma}\|\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|$$

when $\|\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}_n - \mathbf{\Sigma})\| \leq 0.32$ [8].

Before proving concentration, we first derive the expression for $\mathbf{\Sigma}$.

Expression of $\mathbf{\Sigma}$.

$$\mathbf{\Sigma}_n = \begin{bmatrix} \frac{1}{n}\sum_i \pi(\mathbf{z}'_i)(\mathbf{z}'_{i1})^2 + \frac{\lambda}{n} & \frac{1}{n}\sum_i \pi(\mathbf{z}'_i)\mathbf{z}'_{i1}\mathbf{z}'_{i2} & \cdots & \frac{1}{n}\sum_i \pi(\mathbf{z}'_i)\mathbf{z}'_{i,1}\mathbf{z}'_{id} \\ \frac{1}{n}\sum_i \pi(\mathbf{z}'_i)\mathbf{z}'_{i1}\mathbf{z}'_{i2} & \frac{1}{n}\sum_i \pi(\mathbf{z}'_i)(\mathbf{z}'_{i2})^2 + \frac{\lambda}{n} & \cdots & \frac{1}{n}\sum_i \pi(\mathbf{z}'_i)\mathbf{z}'_{i2}\mathbf{z}'_{id} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_i \pi(\mathbf{z}'_i)\mathbf{z}'_{i1}\mathbf{z}'_{id} & \frac{1}{n}\sum_i \pi(\mathbf{z}'_i)\mathbf{z}'_{i2}\mathbf{z}'_{id} & \cdots & \frac{1}{n}\sum_i \pi(\mathbf{z}'_i)(\mathbf{z}'_{id})^2 + \frac{\lambda}{n} \end{bmatrix}$$

By taking $n \to \infty$, we have

$$\mathbf{\Sigma}_n \to \mathbf{\Sigma} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_2 \\ \alpha_2 & \alpha_1 & \cdots & \alpha_2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_2 & \alpha_2 & \cdots & \alpha_1 \end{bmatrix} = (\alpha_1 - \alpha_2)\mathbf{I} + \alpha_2 \mathbf{1}\mathbf{1}^\top$$

where

$$\begin{aligned}
\alpha_1 =& \mathbb{E}_{\mathbf{z}' \sim \text{Uni}(\{0,1\}^d)}[\pi(\mathbf{z}')z'_i] = \mathbb{E}_{\mathbf{z}' \sim \text{Uni}(\{0,1\}^d)}[\pi(\mathbf{z}')(z'_i)^2] \\
=& \sum_{k=0}^{d} e^{(k-d)/\sigma^2} \mathbb{P}(z'_i = 1 | \|\mathbf{z}'\|_0 = k) \mathbb{P}(\|\mathbf{z}'\|_0 = k) \\
=& \sum_{k=0}^{d} e^{(k-d)/\sigma^2} \frac{k}{d} \frac{\binom{d}{k}}{2^d} \\
=& \sum_{k=0}^{d} e^{(k-d)/\sigma^2} \frac{\binom{d-1}{k-1}}{2^d} \\
=& \sum_{k=0}^{d} e^{(k-1)/\sigma^2} e^{(1-d)/\sigma^2} \frac{\binom{d-1}{k-1}}{2^d} \\
=& e^{(1-d)/\sigma^2} \frac{(1 + e^{\frac{1}{\sigma^2}})^{d-1}}{2^d} = \frac{(1 + e^{-\frac{1}{\sigma^2}})^{d-1}}{2^d} \\
\alpha_2 =& \mathbb{E}_{\mathbf{z}' \sim \text{Uni}(\{0,1\}^d)}[\pi(\mathbf{z}')z'_i z'_j] \\
=& \frac{1}{Z} \sum_{k=0}^{d} e^{(k-d)/\sigma^2} \mathbb{P}(z'_i = 1, z'_j = 1 | \|\mathbf{z}'\|_0 = k) \mathbb{P}(\|\mathbf{z}'\|_0 = k) \\
=& \sum_{k=0}^{d} e^{(k-d)/\sigma^2} \frac{k(k-1)}{d(d-1)} \frac{\binom{d}{k}}{2^d} \\
=& \sum_{k=0}^{d} e^{(k-d)/\sigma^2} \frac{\binom{d-2}{k-2}}{2^d} \\
=& \sum_{k=0}^{d} e^{(k-2)/\sigma^2} e^{(2-d)/\sigma^2} \frac{\binom{d-2}{k-2}}{2^d} \\
=& e^{(2-d)/\sigma^2} \frac{(1 + e^{\frac{1}{\sigma^2}})^{d-2}}{2^d} = \frac{(1 + e^{-\frac{1}{\sigma^2}})^{d-2}}{2^d}
\end{aligned}$$

By Sherman-Morrison formula, we have

$$\mathbf{\Sigma}^{-1} = ((\alpha_1 - \alpha_2)\mathbf{I} + \alpha_2 \mathbf{1}\mathbf{1}^\top)^{-1} = \frac{1}{\alpha_1 - \alpha_2}(\mathbf{I} + \frac{\alpha_2}{\alpha_1 - \alpha_2}\mathbf{1}\mathbf{1}^\top)^{-1} = \frac{1}{\alpha_1 - \alpha_2}(\mathbf{I} - \frac{\frac{\alpha_2}{\alpha_1 - \alpha_2}\mathbf{1}\mathbf{1}^\top}{1 + \frac{\alpha_2}{\alpha_1 - \alpha_2}d}) = (\beta_1 - \beta_2)\mathbf{I} + \beta_2 \mathbf{1}\mathbf{1}^\top$$

where

$$\beta_1 = \frac{\alpha_1 + (d-2)\alpha_2}{(\alpha_1 - \alpha_2)(\alpha_1 + (d-1)\alpha_2)}, \beta_2 = -\frac{\alpha_2}{(\alpha_1 - \alpha_2)(\alpha_1 + (d-1)\alpha_2)}$$

In the following, we will prove the concentration of $\hat{\mathbf{w}}^{\text{LIME}}$.

Concentration of $\mathbf{\Sigma}_n$. Since $0 \leq \pi(\cdot) \leq 1, \mathbf{z}_i \in \{0,1\}^d$, we have each element in $\mathbf{\Sigma}_n$ is in $[0, 1+\frac{\lambda}{n}]$. In addition, as

$$\frac{1}{2^d} \leq \alpha_1 = \frac{(1+e^{-\frac{1}{\sigma^2}})^{d-1}}{2^d} \leq \frac{2^{d-1}}{2^d} = \frac{1}{2}$$
$$\frac{1}{2^d} \leq \alpha_2 = \frac{(1+e^{-\frac{1}{\sigma^2}})^{d-2}}{2^d} \leq \frac{2^{d-2}}{2^d} = \frac{1}{4}$$
$$\frac{e^{-1/\sigma^2}}{2^d} \leq \alpha_1 - \alpha_2 = e^{-\frac{1}{\sigma^2}} \frac{(1+e^{-\frac{1}{\sigma^2}})^{d-2}}{2^d} \leq \frac{1}{4}$$

we have elements in $\mathbf{\Sigma}$ are in range $[0, \frac{1}{4}]$. Thus, elements in $\mathbf{\Sigma}_n - \mathbf{\Sigma}$ are in range $[-\frac{1}{4}, 1+\frac{\lambda}{n}]$.

By matrix Heoffding's inequality [32], we have $\forall t > 0$

$$\mathbb{P}(\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|_2 \geq t) \leq 2d \exp(-\frac{nt^2}{8(1+\lambda/n)^2 d^2})$$

Bounding $\|\mathbf{\Sigma}^{-1}\|_F^2$.

$$\|\mathbf{\Sigma}^{-1}\|_F^2 = d\beta_1^2 + (d^2-d)\beta_2^2$$

Because

$$\frac{d}{2^d} \leq \alpha_1 + (d-1)\alpha_2 \leq \frac{2+(d-1)}{4} = \frac{d+1}{4}$$
$$\frac{d-1}{2^d} \leq \alpha_1 + (d-2)\alpha_2 \leq \frac{2+(d-2)}{4} = \frac{d}{4}$$

we have

$$|\beta_1| = |\frac{\alpha_1 + (d-2)\alpha_2}{(\alpha_1-\alpha_2)(\alpha_1+(d-1)\alpha_2)}| \leq |\frac{1}{\alpha_1-\alpha_2}| \leq 2^d e^{1/\sigma^2}, \beta_1^2 \leq 2^{2d} e^{2/\sigma^2}$$

$$|\beta_2| = |-\frac{\alpha_2}{(\alpha_1-\alpha_2)(\alpha_1+(d-1)\alpha_2)}| = |e^{1/\sigma^2}\frac{1}{(\alpha_1+(d-1)\alpha_2)}| \leq d^{-1}2^d e^{1/\sigma^2}, \beta_2^2 \leq d^{-2}2^{2d}e^{2/\sigma^2}$$

so that

$$\|\mathbf{\Sigma}^{-1}\|_F^2 = d\beta_1^2 + (d^2-d)\beta_2^2 \leq d2^{2d}e^{2/\sigma^2} + (d^2-d)d^{-2}2^{2d}e^{2/\sigma^2} \leq 2d2^{2d}e^{2/\sigma^2}$$

Concentration of $\mathbf{\Gamma}_n$. Since $|f| \leq 1$, we have elements in $\mathbf{\Gamma}_n$ and $\mathbf{\Gamma}$ are all in range $[0,1]$. By matrix Hoeffding's inequality [32], we have $\forall t > 0$

$$\mathbb{P}(\|\mathbf{\Gamma}_n - \mathbf{\Gamma}\| \geq t) \leq 2d \exp(-\frac{nt^2}{8d})$$

Concentration of $\hat{\mathbf{w}}^{\text{LIME}}$. When $\|\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}_n - \mathbf{\Sigma})\| \leq 0.32$ [8], we have

$$\|\mathbf{\Sigma}_n^{-1}\mathbf{\Gamma}_n - \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\| \leq 2\|\mathbf{\Sigma}^{-1}\|_F\|\mathbf{\Gamma}_n - \mathbf{\Gamma}\|_2 + 2\|\mathbf{\Sigma}^{-1}\|_F^2\|\mathbf{\Gamma}\|\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|$$

Because

$$\|\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}_n - \mathbf{\Sigma})\| \leq \|\mathbf{\Sigma}^{-1}\|\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\| \leq 2^{1/2}d^{1/2}2^d e^{1/\sigma^2}\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|$$

By concentration of $\mathbf{\Sigma}_n$, we have when $n \geq n_1 = 2^5(1+\lambda)^2 d^3 2^{2d} e^{2/\sigma^2}\log(4d/\delta), t = t_1 = 5^{-2}2^{2.5}d^{-0.5}2^{-d}e^{-1/\sigma^2}$,

$$\mathbb{P}(\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|_2 \geq t) \leq 2d\exp(-\frac{nt^2}{8(1+\lambda/n)^2d^2}) \leq 2d\exp(-\frac{nt^2}{8(1+\lambda)^2d^2}) \leq \frac{\delta}{2}$$

Then with probability at least $1-\frac{\delta}{2}$, we have

$$\|\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}_n - \mathbf{\Sigma})\| \leq \|\mathbf{\Sigma}^{-1}\|\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\| \leq 2^{1/2}d^{1/2}2^d e^{1/\sigma^2}\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\| \leq 0.32$$

When $n \geq n_2 = 2^8\epsilon^{-2}d^2 2^{2d}e^{2/\sigma^2}\log(4d/\delta), t_2 = 2^{-2.5}d^{-0.5}2^{-d}e^{-1/\sigma^2}\epsilon$, we have

$$\mathbb{P}(\|\mathbf{\Gamma}_n - \mathbf{\Gamma}\| \geq t_2) \leq 2d\exp(-\frac{n_2 t_2^2}{8d}) \leq \frac{\delta}{2}$$

In this case, with probability at least $1-\frac{\delta}{2}$ we have

$$\|\mathbf{\Sigma}^{-1}\|\|\mathbf{\Gamma}_n-\mathbf{\Gamma}\|\leq\frac{\epsilon}{4}$$

Because $\|\mathbf{\Gamma}\|\leq\sqrt{d}$, by choosing $n\geq n_3=2^7(1+\lambda)^2\epsilon^{-2}d^5 2^{4d}e^{4/\sigma^2}\log(4d/\delta), t_3=2^{-3}d^{-1.5}2^{-2d}e^{-2/\sigma^2}\epsilon$,

$$\mathbb{P}(\|\mathbf{\Sigma}_n-\mathbf{\Sigma}\|_2\geq t_3)\leq 2d\exp(-\frac{n_3t_3^2}{8(1/4+\lambda/n_3)^2d^2})\leq 2d\exp(-\frac{n_3t_3^2}{8(1+\lambda)^2d^2})\leq\frac{\delta}{2}$$

and with probability at least $1-\delta/2$

$$\|\mathbf{\Sigma}^{-1}\|^2\|\mathbf{\Gamma}\|\|\mathbf{\Sigma}_n-\mathbf{\Sigma}\|\leq\frac{\epsilon}{4}$$

Therefore, overall, we can choose $n\geq\max\{n_1,n_2,n_3\}$, and then we have $\forall\epsilon>0,\delta\in(0,1)$

$$\mathbb{P}(\|\mathbf{\Sigma}_n^{-1}\mathbf{\Gamma}_n-\mathbf{\Sigma}^{-1}\mathbf{\Gamma}\|\geq\epsilon)\leq\delta$$

□

B.7 Proof of Theorem 4.2 and Corollary 4.3

Theorem B.5. *Suppose $\mathbf{z}'\sim\mathcal{P}$ such that the largest eigenvalue of $\mathbf{z}'(\mathbf{z}')^\top$ is bounded by R and $\mathbb{E}[\mathbf{z}'(\mathbf{z}')^\top]=(\alpha_1-\alpha_2)\mathbf{I}+\alpha_2\mathbf{1}\mathbf{1}^\top, \|Var(\mathbf{z}'(\mathbf{z}')^\top)\|_2\leq\nu^2$, $|(\mathbf{z}'f(\mathbf{z}))_i|\leq M$ for some $M>0$. $\{\mathbf{z}_i'\}_{i=1}^n$ are i.i.d. samples from $\mathcal{P}$ and are used to compute* GLIME *explanation $\hat{\mathbf{w}}^{\text{GLIME}}$. For any $\epsilon>0,\delta\in(0,1)$, if $n=\Omega(\epsilon^{-2}M^2\nu^2d^3\gamma^4\log(4d/\delta))$ where $\gamma^2=d\beta_1^2+(d^2-d)\beta_2^2, \beta_1=(\alpha_1+(d-2)\alpha_2)/\beta_0, \beta_2=-\alpha_2/\beta_0, \beta_0=(\alpha_1-\alpha_2)(\alpha_1+(d-1)\alpha_2))$, we have $\mathbb{P}(\|\hat{\mathbf{w}}^{\text{GLIME}}-\mathbf{w}^{\text{GLIME}}\|_2<\epsilon)\geq 1-\delta$. $\mathbf{w}^{\text{GLIME}}=\lim_{n\to\infty}\hat{\mathbf{w}}^{\text{GLIME}}$.*

Proof. The proof is similar with the proof of Theorem 4.1. By the same derivation, we can obtain that

$$\mathbf{\Sigma}=(\alpha_1+\lambda-\alpha_2)\mathbf{I}+\alpha_2\mathbf{1}\mathbf{1}^\top, \mathbf{\Sigma}^{-1}=(\beta_1-\beta_2)\mathbf{I}+\beta_2\mathbf{1}\mathbf{1}^\top$$

$$\beta_1=\frac{\alpha_1+\lambda+(d-2)\alpha_2}{(\alpha_1+\lambda-\alpha_2)(\alpha_1+\lambda+(d-1)\alpha_2)}, \beta_2=-\frac{\alpha_2}{(\alpha_1+\lambda-\alpha_2)(\alpha_1+\lambda+(d-1)\alpha_2)}$$

Since $\lambda_{\max}(z'(z')^\top)\leq R$ and $\|\text{Var}(\mathbf{z}'(\mathbf{z}')^\top)\|_2\leq\nu^2$, by matrix Hoeffding's inequality [31], we have $\forall t>0$

$$\mathbb{P}(\|\mathbf{\Sigma}_n-\mathbf{\Sigma}\|_2\geq t)\leq 2d\exp(-\frac{nt^2}{8\nu^2})$$

Applying Hoeffding's inequality coordinate by coordinate, we have

$$\mathbb{P}(\|\mathbf{\Gamma}_n-\mathbf{\Gamma}\|\geq t)\leq 2d\exp(-\frac{nt^2}{8M^2d^2})$$

$$\|\mathbf{\Sigma}^{-1}\|_F^2=d\beta_1^2+(d^2-d)\beta_2^2=\gamma^2$$

By choosing $n\geq n_1=2^5\gamma^2\nu^2\log(4d/\delta), t_1=2^3 5^{-2}\gamma^{-1}$, we have

$$\mathbb{P}(\|\mathbf{\Sigma}_n-\mathbf{\Sigma}\|_2\geq t_1)\leq 2d\exp(-\frac{n_1t_1^2}{8\nu^2})\leq\frac{\delta}{2}$$

with probability at least $1-\delta/2$

$$\|\mathbf{\Sigma}^{-1}(\mathbf{\Sigma}_n-\mathbf{\Sigma})\|\leq\|\mathbf{\Sigma}^{-1}\|\cdot\|\mathbf{\Sigma}_n-\mathbf{\Sigma}\|\leq\gamma t_1=0.32$$

Let $n\geq n_2=2^5\epsilon^{-2}M^2d^2\gamma^2\log(4d/\delta), t_2=2^{-2}\epsilon\gamma^{-1}$, we have

$$\mathbb{P}(\|\mathbf{\Gamma}_n-\mathbf{\Gamma}\|\geq t_2)\leq 2d\exp(-\frac{n_2t_2^2}{8M^2d^2})\leq\frac{\delta}{2}$$

with probability at least $1-\delta/2$

$$\|\mathbf{\Sigma}\|\|\mathbf{\Gamma}_n - \mathbf{\Gamma}\| \leq \gamma t_2 \leq \frac{\epsilon}{4}$$

Since $\|\mathbf{\Gamma}\| \leq M$, by choosing $n \geq n_3 = 2^5\epsilon^{-2}M^2\nu^2 d\gamma^4 \log(4d/\delta), t_3 = 2^{-2}\epsilon M^{-1}d^{-0.5}\gamma^{-2}$, we have

$$\mathbb{P}(\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\|_2 \geq t_3) \leq 2d\exp(-\frac{n_3 t_3^2}{2\nu^2}) \leq \frac{\delta}{2}$$

and with probability at least $1-\delta/2$,

$$\|\mathbf{\Sigma}^{-1}\|^2\|\mathbf{\Gamma}\|\|\mathbf{\Sigma}_n - \mathbf{\Sigma}\| \leq \gamma^2 M d^{0.5} t_3 = \frac{\epsilon}{4}$$

Therefore, by choosing $n = \max\{n_1, n_2, n_3\}$, we have

$$\mathbb{P}(\|\mathbf{\Sigma}_n^{-1}\mathbf{\Gamma}_n - \mathbf{\Sigma}^{-1}\mathbf{\Gamma}\| \geq \epsilon) \leq \delta$$

□

Corollary B.6. *Suppose $\{\mathbf{z}_i'\}_{i=1}^n$ are i.i.d. samples from $\mathbb{P}(\mathbf{z}', \|\mathbf{z}'\|_0 = k) = e^{k/\sigma^2}/(1+e^{1/\sigma^2})^d, k = 1,\ldots,d$ are used to compute* GLIME-BINOMIAL *explanation. For any $\epsilon > 0, \delta \in (0,1)$, if $n = \Omega(\epsilon^{-2}d^5 e^{4/\sigma^2}\log(4d/\delta))$, we have $\mathbb{P}(\|\hat{\mathbf{w}}^{Binomial} - \mathbf{w}^{Binomial}\|_2 < \epsilon) \geq 1-\delta$. $\mathbf{w}^{Binomial} = \lim_{n\to\infty}\hat{\mathbf{w}}^{Binomial}$.*

Proof. For GLIME-BINOMIAL, we have each coordinate of $\mathbf{z}'(\mathbf{z}')^\top$ follows a Bernoulli distribution so that variance of $\mathbf{z}'(\mathbf{z}')^\top$ and $(\mathbf{z}' f(\mathbf{z}'))_i$ are bounded. We also have

$$\|\mathbf{\Gamma}\| \leq \sqrt{d},$$

$$\begin{aligned}
\alpha_1 =& \mathbb{E}[(z_i^2)'] = \mathbb{E}[z_i'] = \frac{e^{1/\sigma^2}}{1+e^{1/\sigma^2}} \\
=& \sum_{k=0}^{d}\mathbb{P}(z_i'=1|\|\mathbf{z}'\|_0=k)\mathbb{P}(\|\mathbf{z}'\|_0=k) \\
=& \sum_{k=0}^{d}\frac{k}{d}\frac{\binom{d}{k}e^{k/\sigma^2}}{(1+e^{1/\sigma^2})^d} \\
=& \sum_{k=0}^{d}\frac{\binom{d-1}{k-1}e^{k/\sigma^2}}{(1+e^{1/\sigma^2})^d} \\
=& \frac{(1+e^{1/\sigma^2})^{d-1}}{(1+e^{1/\sigma^2})^d}e^{1/\sigma^2} = \frac{e^{1/\sigma^2}}{1+e^{1/\sigma^2}} \\
\alpha_2 =& \mathbb{E}[z_i'z_j'] = \frac{e^{1/\sigma^2}}{1+e^{1/\sigma^2}} \\
=& \sum_{k=0}^{d}\mathbb{P}(z_i'=1,z_j'=1|\|\mathbf{z}'\|_0=k)\mathbb{P}(\|\mathbf{z}'\|_0=k) \\
=& \sum_{k=0}^{d}\frac{k(k-1)}{d(d-1)}\frac{\binom{d}{k}e^{k/\sigma^2}}{(1+e^{1/\sigma^2})^d} \\
=& \sum_{k=0}^{d}\frac{\binom{d-2}{k-2}e^{k/\sigma^2}}{(1+e^{1/\sigma^2})^d} \\
=& \frac{(1+e^{1/\sigma^2})^{d-2}}{(1+e^{1/\sigma^2})^d}e^{2/\sigma^2} = \frac{e^{2/\sigma^2}}{(1+e^{1/\sigma^2})^2} = \alpha_1^2
\end{aligned}$$

$$|\beta_1|^2 = |\frac{\alpha_1+\lambda+(d-2)\alpha_2}{(\alpha_1+\lambda-\alpha_2)(\alpha_1+\lambda+(d-1)\alpha_2)}|^2 \leq |\frac{1}{\alpha_1+\lambda-\alpha_2}| \leq \frac{1}{|\alpha_1-\alpha_2|} = e^{-1/\sigma^2}(1+e^{1/\sigma^2})^2 \leq 4e^{1/\sigma^2}$$

$$
\begin{aligned}
|\beta_2|^2 =& |-\frac{\alpha_2}{(\alpha_1+\lambda-\alpha_2)(\alpha_1+\lambda+(d-1)\alpha_2)}|^2 \\
\leq& \frac{\alpha_2^2}{(\alpha_1-\alpha_2)(\alpha_1+\lambda+(d-1)\alpha_2)^2} \\
=& \frac{\alpha_1\alpha_2}{(1-\alpha_1)(\alpha_1+(d-1)\alpha_2)^2} \\
\leq& \frac{\alpha_1\alpha_2}{(1-\alpha_1)((d-1)\alpha_2)^2} = \frac{e^{-1/\sigma^2}(1+e^{1/\sigma^2})^2}{(d-1)^2} \leq \frac{2^2 e^{1/\sigma^2}}{(d-1)^2}
\end{aligned}
$$

Therefore,

$$
d\beta_1^2 + (d^2-d)\beta_2^2 \leq de^{1/\sigma^2} + e^{1/\sigma^2}\frac{d}{d-1} \leq de^{1/\sigma^2}
$$

□

B.8 Formulation of SmoothGrad

Proposition B.7. *SmoothGrad is equivalent to* GLIME *formulation with* $\mathbf{z} = \mathbf{z}' + \mathbf{x}$ *where* $\mathbf{z}' \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, $\ell(f(\mathbf{z}), g(\mathbf{z}')) = (f(\mathbf{z}) - g(\mathbf{z}'))^2$ *and* $\pi(\mathbf{z}) = 1, \Omega(\mathbf{v}) = 0$.

The explanation returned by GLIME *for* f *at* $\mathbf{x}$ *with infinitely many samples under the above setting is*

$$
\mathbf{w}^* = \frac{1}{\sigma^2}\mathbb{E}_{\mathbf{z}'\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}[\mathbf{z}' f(\mathbf{z}'+\mathbf{x})] = \mathbb{E}_{\mathbf{z}'\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}[\nabla f(\mathbf{x}+\mathbf{z}')]
$$

which is exactly SmoothGrad explanation. When $\sigma \to 0$, $\mathbf{w}^* \to \nabla f(\mathbf{x}+\mathbf{z})|_{\mathbf{z}=\mathbf{0}}$.

Proof. To prove this proposition, we first derive the expression of GLIME explanation $\mathbf{w}^*$.

Exact Expression of $\boldsymbol{\Sigma}$. $\mathbf{z}'_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}), \forall i = 1, \cdots, n$. In this case

$$
\hat{\boldsymbol{\Sigma}}_n = \begin{bmatrix} \frac{1}{n}\sum_k (z_{k1}^2)' & \cdots & \frac{1}{n}\sum_k z'_{l1}z'_{kd} \\ \vdots & \ddots & \vdots \\ \frac{1}{n}\sum_k z'_{kd}z'_{k1} & \cdots & \frac{1}{n}\sum_k (z_{kd}^2)' \end{bmatrix}
$$

Then we have

$$
\boldsymbol{\Sigma} = \mathbb{E}_{\mathbf{z}'\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}[\mathbf{z}'(\mathbf{z}')^\top] = \begin{bmatrix} \sigma^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma^2 \end{bmatrix}
$$

$$
\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{\sigma^2} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma^2} \end{bmatrix}
$$

As a direct consequence, we have

$$
\mathbf{w}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Gamma} = \frac{1}{\sigma^2}\mathbb{E}_{\mathbf{z}'\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}[\mathbf{z}' f(\mathbf{x}+\mathbf{z}')] = \mathbb{E}_{\mathbf{z}'\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})}[\nabla f(\mathbf{x}+\mathbf{z}')]
$$

The last equality directly follows from Stein's lemma [17]. □