STAGE-REGULARIZED NEURAL STEIN CRITICS FOR TESTING GOODNESS-OF-FIT OF GENERATIVE MODELS

Yao Xie[1], Matthew Repasky[1], and Xiuyuan Cheng[2]

[1]School of Industrial & Systems Engineering, Georgia Tech

[2]Department of Mathematics, Duke University

INTRODUCTION

The Stein discrepancy provides a means to assess the Goodness-of-Fit (GoF) of statistical models. The Stein discrepancy only requires the model score function, making it naturally applicable to energy-based models (EBMs), which are described only up to a normalizing constant. Neural network stein critics trained using a novel **staged regularization scheme** are used to compute the Stein discrepancy. The resultant critics **localize the discrepancy between distributions induced by generative models** of image data.

NEURAL STEIN CRITICS

The Stein discrepancy between distributions p and q evaluated at *critic function* $\mathbf{f} \in \mathcal{F}$ is:

$$\mathrm{SD}[\mathbf{f}] \coloneqq \mathbb{E}_{x\sim p}\mathbf{s}_q(x)\cdot\mathbf{f}(x) + \nabla\cdot\mathbf{f}(x) = \mathbb{E}_{x\sim p}T_q\mathbf{f}(x),$$

where $\mathbf{s}_q = \nabla q/q$ is the score of q. The Stein discrepancy over function class $\mathcal{F}$ is

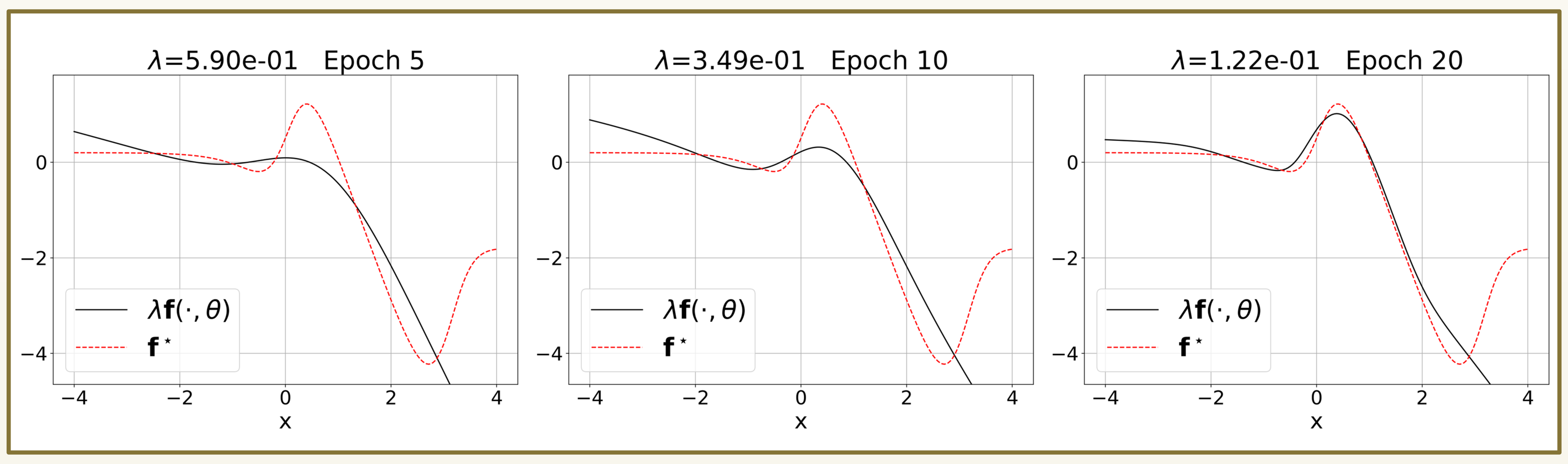
$$\mathrm{SD}_{\mathcal{F}}(p,q) \coloneqq \sup_{\mathbf{f}\in\mathcal{F}} \mathrm{SD}[\mathbf{f}]\,.$$

If $\mathcal{F} = L^2$, the spaced of squared-integrable vector fields, then $\mathbf{f}^* = \mathbf{s}_q - \mathbf{s}_p$. The optimal neural Stein critic is related to the $\mathbf{f}$ which minimizes the regularized functional:

$$\mathcal{L}_\lambda[\mathbf{f}] \coloneqq -\mathrm{SD}[\mathbf{f}] + \frac{\lambda}{2}\mathbb{E}_{x\sim p}\|\mathbf{f}(x)\|^2,$$

which is minimal at $\mathbf{f}_\lambda^* \coloneqq \lambda^{-1}\mathbf{f}^*$. A neural network Stein critic $\mathbf{f}(\cdot,\theta)$ can be trained to minimize $\mathcal{L}_\lambda$.

STAGED REGULARIZATION



Large λ early in training can be **approximated by neural tangent kernel** (NTK) theory: $\mathbf{f}(\cdot,\theta)$ reaches its optimum in $\sim 1/\lambda$ time. Weaker λ may be necessary to **go beyond the kernel learning regime**. Log-linear staging is used:

$$\Lambda(B_i;\lambda_{\text{init}},\lambda_{\text{term}},\beta) = \max\{\lambda_{\text{init}}\cdot\beta^i,\lambda_{\text{term}}\}.$$

Let $\beta \in (0,1)$ be the decay rate, B be the staging period in batches, and $B_i = i\cdot B$.

GOODNESS-OF-FIT

GoF assesses $H_0\!: p = q$ versus $H_1\!: p \neq q$ given a model q and $X = \{x_i\}$ drawn from p. A test statistic $\hat{T}(X)$ is used to reject H_0 if $\hat{T} > t_{\text{thresh}}$. Given neural Stein critic $\mathbf{f}(\cdot,\theta)$:

$$\hat{T} \coloneqq \frac{1}{n_{\text{GoF}}}\sum_{i=1}^{n_{\text{GoF}}} T_q\mathbf{f}(x)\,,$$

which is an estimator of the Stein discrepancy.

ENERGY-BASED MODEL EVALUATION

Let q be a model of data distribution p of the form:

$$q(x) = Z^{-1}\cdot\exp\left(-E_\phi(x)\right),$$

where Z is a normalizing constant that is not required to compute the score $\mathbf{s}_q = -\nabla E_\phi(x)$. Stein critics can be used to assess the local discrepancy between p and q:

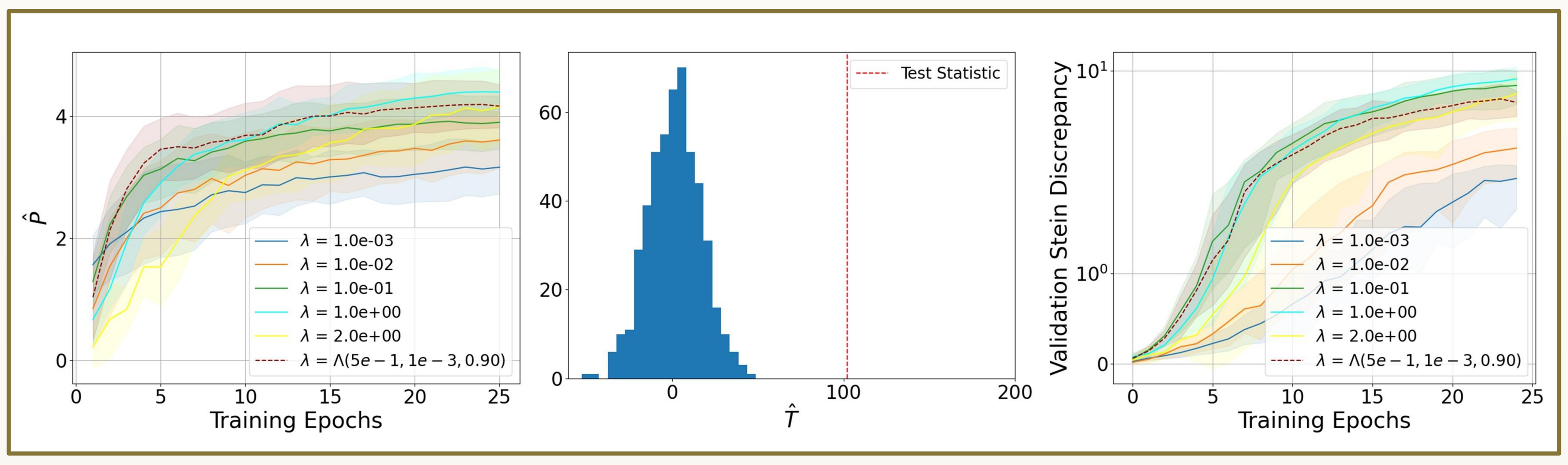
$$w(x) = T_q\mathbf{f}(x,\theta).$$

Let $\bar{w}_p$ and $\sigma(w_p)$ be the mean and standard deviation of $w(x)$ computed on $x\sim p$, and likewise $\sigma(w_q)$ for $x\sim q$. Metric $\hat{P}$ reflects the discrepancy between distributions p and q in terms of test statistic $\hat{T}$:

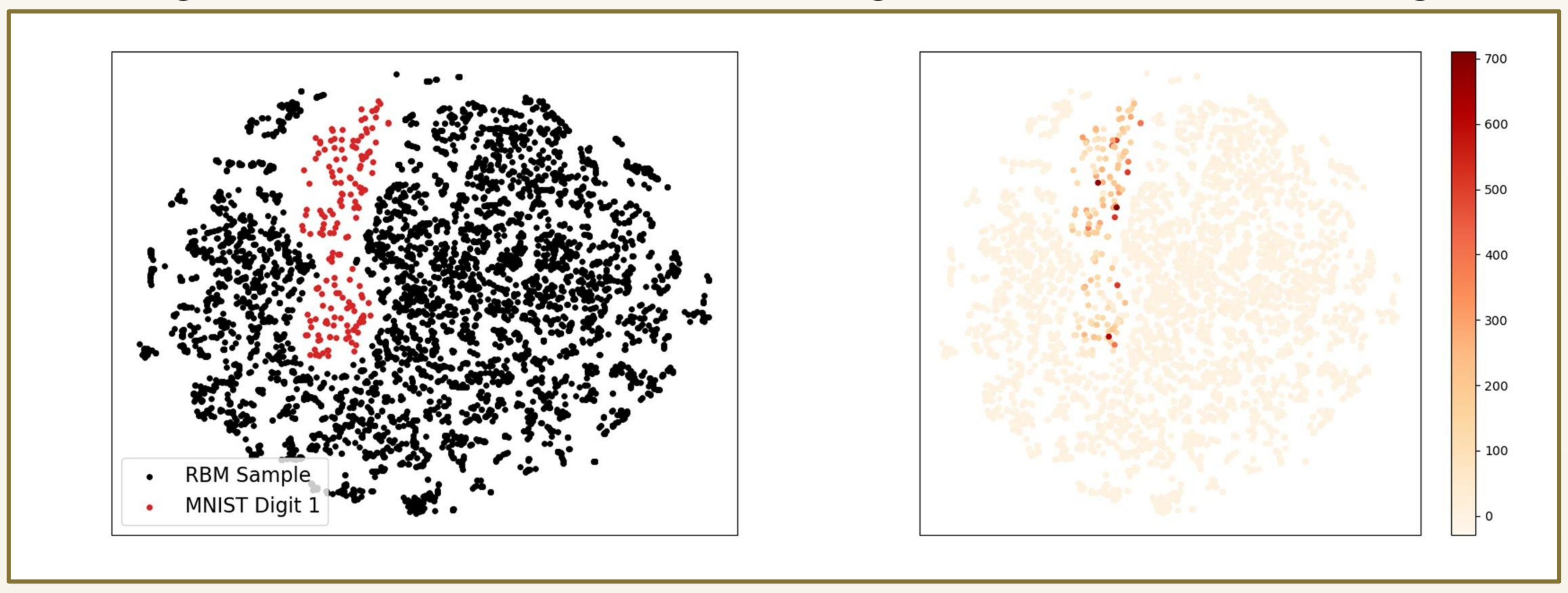
$$\hat{P} = \frac{\bar{w}_p}{\sigma(w_p)+\sigma(w_q)}.$$

EXPERIMENT

Let q be an EBM representing MNIST digits “1” and p be a mixture of 97% q and 3% true digits “1”. Neural Stein critics are trained using 2,000 samples from p. In training, staged-regularized critics more rapidly yield high-power discriminators compared to fixed-λ regularization.



The staged-regularized critic is applied to a validation set to yield $w(x)$ for $x\sim p$. In a t-SNE embedding of this set, the true MNIST digits are highlighted in red on the left. On the right, the value of $w(x)$ is larger for true MNIST digits.



The critic evaluated at anomalous points reveals the ability of neural Stein critics to localize disparity.

CONCLUSION

Staged regularization results in **more efficient learning** of neural Stein critics which can **localize discrepancy** in distributions represented by generative models.

REFERENCES

[1] M. Repasky, X. Cheng, & Y. Xie. “Stage-Regularized Neural Stein Critics For Testing Goodness-Of-Fit Of Generative Models.” ICASSP, 2024.

[2] M. Repasky, X. Cheng, & Y. Xie. "Neural Stein Critics With Staged L2-Regularization." IEEE Transactions on Information Theory, 2023.

[3] W. Grathwohl, et al. "Learning the stein discrepancy for training and evaluating energy-based models without sampling." ICML, 2020.