ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10445927

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

Matthew Repasky[⋆] *Xiuyuan Cheng*[†] *and Yao Xie*[⋆]

[⋆]H. Milton Stewart School of Industrial and Systems Engineering,
Georgia Institute of Technology, Atlanta, GA, 30332
[†]Department of Mathematics,
Duke University, Durham, NC, 27708

ABSTRACT

Learning to differentiate model distributions from observed data is a fundamental problem in statistics and machine learning, and high-dimensional data remains a challenging setting for such problems. Metrics that quantify the disparity in probability distributions, such as the Stein discrepancy, play an important role in high-dimensional statistical testing. This paper presents a method based on neural network Stein critics to distinguish between data sampled from an unknown probability distribution and a nominal model distribution with a novel staging of the weight of regularization. The benefit of using staged L^2 regularization in training such critics is demonstrated on evaluating generative models of image data.

Index Terms— Stein Discrepancy, Goodness-of-fit Test, Generative Models

1. INTRODUCTION

Minimizing the discrepancy between target and model probability distributions can be used to construct probability density models given observed data. Generally, generative models require discriminative critics to distinguish between data and a distribution [1]. A wide array of integral probability metrics quantify distances on probability distributions [2], including the Stein discrepancy [3]. Computing the Stein discrepancy only requires knowledge of the score function of the model distribution, avoiding the need to integrate normalizing constants. This makes the Stein discrepancy useful for evaluating generative models such as energy-based models. Although regularized Stein discrepancy has been considered in prior work, the impact of regularization strength parameters on neural network optimization and training performance was overlooked in previous studies.

The work is supported by NSF DMS-2134037. M.R. and Y.X. are partially supported by an NSF CAREER CCF-1650913, and NSF DMS-2134037, CMMI-2015787, CMMI-2112533, DMS-1938106, and DMS-1830210. X.C. is also partially supported by NSF DMS-1818945, NIH R01GM131642 and the Alfred P. Sloan Foundation.

The concept of goodness-of-fit (GoF) is closely related to estimating such discrepancies. Integral probability metrics are widely used for such problems. Particularly, Stein discrepancy methods have been developed for GoF testing, including kernel methods [4–6] and, more recently, deep neural network-aided techniques [7].

Energy-based models (EBMs) are a particularly useful subset of generative models, and can be described by an *energy function* describing a probability density up to a normalizing constant [8]. While such models provide flexibility in representing a probability density, the normalizing constant is required to compute the likelihood of data given the model. The Stein discrepancy provides a metric for evaluating EBMs without knowledge of this normalization constant [7]. In Section 3.3, we outline our approach for evaluating generative EBM models using neural Stein critics.

We present a novel staging of regularization for learning the Stein discrepancy in training neural network Stein critics. We consider the L^2-regularization of the neural Stein critic, which has been adopted in past studies [7, 9]. However, our method focuses on the role of the regularization strength in L^2 neural Stein methods, exploiting its impact on neural network optimization, which was overlooked in previous studies. Such trained neural Stein critics provide model comparison capabilities to assess the accuracy of a model's approximation of the true distribution, allowing for identification of locations of distribution departure in observed data. This naturally leads to applications for GoF testing and evaluation of generative models. The contributions of the current work are as follows:

1. We introduce a staging of the regularization weight in training neural Stein critics, annealing from strong to weak regularization throughout training.
2. We propose a generic scheme (log-linear staging) which obtains GoF tests of improved power compared to using fixed regularization weight in training.
3. The ability to localize discrepancy between distributions is highlighted in experiments distinguishing generative models of image data.

2. BACKGROUND

Consider probability densities on $\mathcal{X} \subseteq \mathbb{R}^d$. Given such a density q, for a sufficiently regular vector field $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$, the Stein operator T_q [10, 11] applied to $\mathbf{f}$ is defined as $T_q\mathbf{f}(x) := \mathbf{s}_q(x) \cdot \mathbf{f}(x) + \nabla \cdot \mathbf{f}(x)$, where $\mathbf{s}_q := \nabla q/q = \nabla \log q$ is the *score function* of q. Given another probability density p on $\mathcal{X}$ and sufficiently regular, bounded function class $\mathcal{F}$, the Stein discrepancy [3] measuring the difference between p and q is defined as

$$\mathrm{SD}_{\mathcal{F}}(p, q) := \sup_{\mathbf{f} \in \mathcal{F}} \mathrm{SD}[\mathbf{f}], \quad \mathrm{SD}[\mathbf{f}] := \mathbb{E}_{x \sim p} T_q \mathbf{f}(x). \quad (1)$$

We call $\mathbf{f}$ the "critic" and $\mathrm{SD}[\mathbf{f}]$ the Stein discrepancy evaluated at the critic $\mathbf{f}$. Under mild boundary and regularity conditions on $\mathbf{f}$ and when $p = q$, Stein's identity states that $\mathrm{SD}[\mathbf{f}] = 0$. The other direction, namely that zero Stein discrepancy implies $p = q$, is also established under certain conditions [12].

2.1. L^2 Stein critic

For $\mathbf{v}, \mathbf{w} : \mathcal{X} \to \mathbb{R}^d$, define the L^2 inner-product of vector fields on $(\mathcal{X}, p(x)dx)$ as $\langle \mathbf{v}, \mathbf{w} \rangle_p := \int_{\mathcal{X}} \mathbf{v}(x) \cdot \mathbf{w}(x) p(x) dx$ with L^2 norm defined as $\|\mathbf{v}\|_p^2 := \langle \mathbf{v}, \mathbf{v} \rangle_p$. We denote all critics $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$ such that $\|\mathbf{f}\|_p^2 < \infty$ the space of $L^2(p) := L^2(\mathcal{X}, p(x)dx)$.

The Stein discrepancy over L^2 critics which vanish at the boundary of $\mathcal{X}$ is defined, for some $r > 0$, as

$$\mathrm{SD}_r(p, q) = \sup_{\|\mathbf{f}\|_p \leq r} \mathrm{SD}[\mathbf{f}], \quad (2)$$

where $\mathrm{SD}[\mathbf{f}]$ is defined as in (1). Define

$$\mathbf{f}^* := \mathbf{s}_q - \mathbf{s}_p. \quad (3)$$

The supremum of (2) is achieved at $\mathbf{f} = (r/\|\mathbf{f}^*\|_p)\mathbf{f}^*$ under the assumption that the score functions are in $L^2(p)$ and vanish at the boundary of $\mathcal{X}$, implying that if $\|\mathbf{f}^*\|_p = 0$, then $\mathrm{SD}_r(p, q) = 0$. There exists a closed-form solution of the Stein discrepancy under these assumptions:

$$\mathrm{SD}[\mathbf{f}] = \mathbb{E}_{x \sim p} T_q \mathbf{f}(x) = \langle \mathbf{f}^*, \mathbf{f} \rangle_p, \quad (4)$$

see [4, Lemma 2.3] and the proof of [13, Lemma II.1].

2.2. Goodness-of-Fit (GoF) tests

To assess whether samples $X = \{x_i\}$ drawn from an unknown distribution p come from some *model distribution* q, define the null hypothesis as $H_0 : p = q$ and the alternative as $H_1 : p \neq q$. A GoF test is conducted using a test statistic $\hat{T} = \hat{T}(X)$. After specifying a "test threshold" t_{thresh}, H_0 is rejected if $\hat{T} > t_{\mathrm{thresh}}$. The selection of t_{thresh} controls the Type-I error, defined as $\Pr[\hat{T} > t_{\mathrm{thresh}}]$ under H_0. The goal is to guarantee that $\Pr[\hat{T} > t_{\mathrm{thresh}}] \leq \alpha$, which is called the "significance level". The Type-II error is defined as $\Pr[\hat{T} \leq t_{\mathrm{thresh}}]$ under H_1. Finally, the "test power" is defined as one minus the Type-II error. See Section 3.2 for details of the test statistic computed using a neural Stein critic.

3. NEURAL L^2 STEIN CRITIC

Replacing the L^2-norm constraint in (2) to be a regularization term leads to the following minimization:

$$\mathcal{L}_\lambda[\mathbf{f}] := -\mathrm{SD}[\mathbf{f}] + \frac{\lambda}{2}\|\mathbf{f}\|_p^2 = \frac{1}{2\lambda}(\|\lambda\mathbf{f} - \mathbf{f}^*\|_p^2 - \|\mathbf{f}^*\|_p^2), \quad (5)$$

where $\lambda > 0$ is the penalty weight of the L^2 regularization. The final expression is by (4). Thus, (5) immediately gives that $\mathcal{L}_\lambda[\mathbf{f}]$ is minimized at

$$\mathbf{f}_\lambda^* := \frac{\mathbf{f}^*}{\lambda} = \frac{1}{\lambda}(\mathbf{s}_q - \mathbf{s}_p), \quad (6)$$

see Theorem 4.1 of [9]. The expression (6) reveals that λ only acts to scale the optimal critic, indicating that the choice of λ may play a role only in the optimization of neural Stein critics. See Section 3.1 for details as to how this can be incorporated in the training of neural Stein critics.

Following [7,9], we parameterize the critic by a neural network mapping $\mathbf{f}(x, \theta)$ parameterized by θ, assuming $\mathbf{f}(\cdot, \theta) \in L^2(p)$ for any θ being considered. We denote this $\mathbf{f}(\cdot, \theta)$ the "neural Stein critic". The population loss of θ follows from (5) as $L_\lambda(\theta) = \mathcal{L}_\lambda[\mathbf{f}(\cdot, \theta)]$. The neural Stein critic is trained by minimizing the empirical version of $L_\lambda(\theta)$ given n_{tr} training samples: $\hat{L}_\lambda(\theta) := \frac{1}{n_{\mathrm{tr}}} \sum_{i=1}^{n_{\mathrm{tr}}} (-T_q\mathbf{f}(x_i, \theta) + \frac{\lambda}{2}\|\mathbf{f}(x_i, \theta)\|^2)$.

Suppose a neural Stein critic trained by minimizing $\hat{L}_\lambda(\theta)$ approximates the minimizer of $\mathcal{L}_\lambda$, namely the "optimal critic" $\mathbf{f}_\lambda^*$, then both $\mathbf{f}(\cdot, \theta)$ and $\mathrm{SD}[\mathbf{f}(\cdot, \theta)]$ would scale like $1/\lambda$. We call $\mathbf{f}^*$ defined in (3) the "scaleless optimal critic function". The expression (6) also suggests that, if the neural Stein critic successfully approximates the optimal, we would expect $\lambda\mathbf{f}(\cdot, \theta) \approx \mathbf{f}^*$, which is theoretically supported in [13]. We thus call $\lambda\mathbf{f}(\cdot, \theta)$ the "scaleless neural Stein critic".

3.1. Staged-λ regularization in training

A generic choice for staging the weight of regularization is to decrease λ by a multiplicative factor $\beta \in (0, 1)$ every B_w number of batches, beginning with λ_{init} and finishing with λ_{term}. The discrete-time staging is defined as:

$$\Lambda\left(B_i; \lambda_{\mathrm{init}}, \lambda_{\mathrm{term}}, \beta\right) = \max\left\{\lambda_{\mathrm{init}} \cdot \beta^i, \lambda_{\mathrm{term}}\right\},$$

where $B_i = i \cdot B_w$ for $i \in \mathbb{N}$. That is, λ will be set to $\Lambda(B_i)$ when i increments of B_w number of batches have occurred.

The staging of λ in training is supported by the analysis of [13]. Using large λ early in training of the neural Stein critic can be approximately related to Neural Tangent Kernel (NTK) optimization [14], rapidly reaching its best approximation in

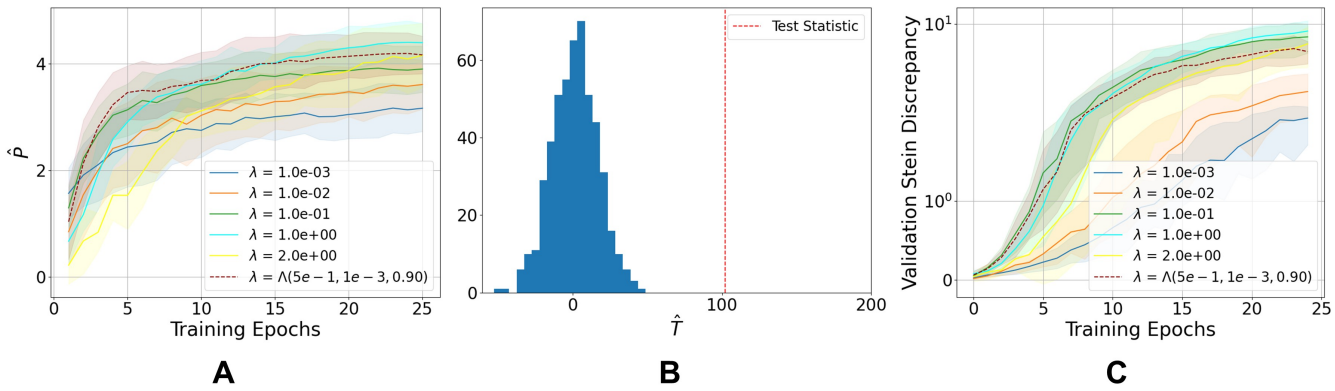

Fig. 1: In (**A**), $\hat{P}$ (8) is visualized through training for each regularization strategy applied to the MNIST mixture dataset. To compute $\hat{P}$, the model is applied to a validation dataset of 1,000 samples. For each regularization strategy, 10 models are trained for 25 epochs; the mean and standard deviation are visualized over these 10 models. In (**B**), a distribution of null statistics (7) are plotted alongside a statistic calculated over an $n_{\text{GoF}} = 100$ sample testing set from p, computed using a staged-regularization critic. Visualized in (**C**), the $\text{SD}[\lambda \mathbf{f}(\cdot, \theta)]$ is computed through training using the same validation datasets.

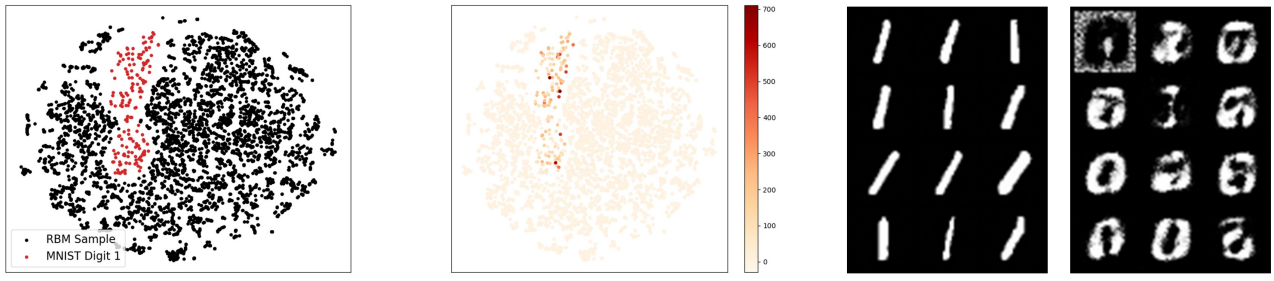

Fig. 2: Embedding via t-SNE of $w(\cdot, \theta)$ scaled by λ using a staged-regularization neural Stein critic applied to a validation dataset of 6,000 samples from p. In (**A**), the black points represent validation samples coming from the RBM while the red points represent true digits 1 from MNIST. In (**B**), the points with high critic witness value are more darkly colored. In (**C**), the 12 images with the highest critic witness value are shown. Similarly, (**D**) shows the 12 images with the lowest critic witness value.

$\sim 1/\lambda$ time, see [13, Theorem IV.6]. However, the proper λ may be much smaller. Therefore, a staging of λ aims to fully exploit the kernel learning (large λ) in the early training phase and annealing to small λ in the later phase of training, going beyond the NTK regime. The log-linear staging outlined in this section is just one method to achieve this.

3.2. Goodness-of-Fit testing

Assume we are given data samples $x_i \sim p$ and that we can sample from the model q and access $\mathbf{s}_q$. We first train a neural Stein critic $\mathbf{f}(x, \theta)$ from a training split of sampled $\{x_i^{\text{tr}}\}_{i=1}^{n_{\text{tr}}}$ followed by computing the following test statistic on the test split $\{x_i\}_{i=1}^{n_{\text{GoF}}}$:

$$\hat{T} = \frac{1}{n_{\text{GoF}}} \sum_{i=1}^{n_{\text{GoF}}} T_q \mathbf{f}(x_i, \theta), \tag{7}$$

which can be viewed as a sample-average estimator of $\text{SD}[\mathbf{f}(\cdot, \theta)]$ as defined in (1).

3.3. Evaluation of EBM generative models

Trained Stein critics can reveal where p and q locally differ on their support. In the case of EBMs, the model probability

density is $q(x) = \exp(-E_\phi(x))/Z$ given normalizing constant Z and real-valued energy function $E_\phi(x)$ (parameterized by ϕ, e.g., a neural network). The score function of q, therefore, is the gradient of the energy function, $\mathbf{s}_q = -\nabla E_\phi(x)$, which can be straightforward to compute. For example, the energy function for Gaussian-Bernoulli Restricted Boltzmann Machines (RBMs) [7], a specific type of EBM, is defined as $E(x, h|B, b, c) = -\frac{1}{2}x^{\mathrm{T}}Bh - b^{\mathrm{T}}x - c^{\mathrm{T}}h + \frac{1}{2}\|x\|^2$ for latent Bernoulli variable h. The score function is thus $\mathbf{s}_q(x) = b - x + B \cdot \tanh(B^{\mathrm{T}}x + c)$. In Section 4, we evaluate RBMs using neural Stein critic functions.

4. EXPERIMENT

We compare fixed and staged regularization strategies to train critics that differentiate a data distribution p and a model distribution q. The data are sampled from the MNIST handwritten digits dataset [15]. The model q is a Gaussian-Bernoulli RBM following the approach of [7]. The data distribution p is a mixture model composed of 97% the RBM and 3% true digits "1" from MNIST. Codes to reproduce the results in this section can be found at the following repository: `https://github.com/mrepasky3/Staged_L2_Neural_Stein_Critics`.

We introduce a validation metric to evaluate the fit of the neural Stein critic $\mathbf{f}(\cdot, \theta)$ in this setting. First, denote the quantity computed by applying the Stein operator with respect to q on neural Stein critic $\mathbf{f}(\cdot, \theta)$ evaluated at a sample $x \in \mathcal{X}$ as the "critic witness" of the sample x: $w(x, \theta) = T_q\mathbf{f}(x, \theta)$. This represents the magnitude of the difference between distributions p and q at $x \in \mathcal{X}$. Evaluating the critic witness at n_{GoF} samples $x_i \sim q$, under the central limit theorem assumption, random variables $w(x_i, \theta)$ have a (centered) normal distribution with standard deviation $\sigma(w)/\sqrt{n_{\mathrm{GoF}}}$ when n_{GoF} is large, where $\sigma(w)$ is the standard deviation of $w(x_i, \theta)$.

Note that the test statistic (7) is the mean of $w(x_i, \theta)$ computed over testing data $x_i \sim p$. To assess the GoF testing power for the neural Stein critic $\mathbf{f}(\cdot, \theta)$, we compare the mean and variance of $w(\cdot, \theta)$ applied to a testing dataset sampled from p and to a "null" dataset drawn from q, both of size n_{GoF}:

$$\hat{P} = \frac{\bar{w}_p}{\sigma(w_p) + \sigma(w_q)} \tag{8}$$

where $\bar{w}_p$ and $\sigma(w_p)$ are the empirical mean and standard deviation of $w(x_i, \theta)$, respectively. This quantity reflects the capability of the critic to differentiate between the distributions in the GoF hypothesis testing setting described in Section 3.2. In addition to the $\hat{P}$ metric from Equation (8), we may also apply the Stein discrepancy evaluated at the scaleless neural Stein critic, i.e., the $\mathrm{SD}[\lambda\mathbf{f}(\cdot, \theta)]$ (1), to the holdout dataset from the data distribution as an evaluation metric for the models as described in Section 3.3.

We train 2-layer (512 hidden units) networks with Swish activation [16]. The critics observe 2,000 training samples from p, training on mini-batches of size 100 with a learning rate 10^{-3} using the default PyTorch Adam optimzer. Each model is trained for 25 epochs. We consider fixed $\lambda \in \{1 \times 10^{-3}, 1 \times 10^{-2}, 1 \times 10^{-1}, 1 \times 10^{0}, 2 \times 10^{0}\}$ and a $\Lambda(5 \times 10^{-1}, 1 \times 10^{-3}, 0.90)$ staging scheme with $B_w = 20$. We fit 10 critics per regularization strategy. For each critic, we compute the validation SD and the power metric (8) throughout training using $n_{\mathrm{GoF}} = 1,000$ samples from p and the same number of "null" samples from q.

The power approximation using Equation (8) is plotted in Figure 1(**A**) for each regularization strategy. Staging provides a more rapid increase in the validation metric in the early training period than fixed-λ strategies while yielding a final model of comparable performance. In Figure 1(**B**), we observe that the test statistic (7) exhibits clear separation from its bootstrapped null distribution (estimated using 500 samples from q). This holds for staged- and fixed-λ regularization strategies. Finally, $\mathrm{SD}[\lambda\mathbf{f}(\cdot, \theta)]$ applied to the holdout dataset from p is visualized throughout training for each model in Figure 1(**C**). Again, the staged approach performs comparably to the best fixed-λ regularization strategies.

We also examine the interpretability of the critic as a diagnostic tool for anomalous observations. By Equation (3), the scaleless optimal critic captures the difference in the score of the model and data distributions. Therefore, a trained neural Stein critic can indicate which samples in a validation dataset represent the largest departure from the distribution q. We isolate such samples by identifying samples with high critic witness value $w(\cdot, \theta)$.

Using a staged regularization model, we visualize the critic witness applied to a validation set of 6,000 samples from p by reducing the images to a two-dimensional embedding via t-SNE [17]. In Figure 2(**A**), the embedding of the validation data is visualized, where the true MNIST points are highlighted in red. In Figure 2(**B**), the true MNIST digits have a larger critic witness value than RBM samples. Visualizing the images which have highest critic witness value in Figure 2(**C**) and those which have the lowest critic witness value in Figure 2(**D**), we find that this approach correctly identifies true digits one from MNIST as anomalous while accepting those generated by the RBM as normal.

5. CONCLUSION

We outline a novel staged regularization method for learning neural Stein critics by starting with strong L^2 regularization and progressively decreasing the regularization weight over the course of training. Critics trained using staged regularization yield more rapid approximations of the target than those using fixed regularization and can, for example, detect distribution differences between authentic and synthetic MNIST data. Further applications can be conducted on other modern generative modeling approaches, such as the Gaussian-Bernoulli RBM, including normalizing flow architectures.

6. REFERENCES

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[2] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet, "On integral probability metrics,\phi-divergences and binary classification," *arXiv preprint arXiv:0901.2698*, 2009.

[3] Jackson Gorham and Lester Mackey, "Measuring sample quality with stein's method," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[4] Qiang Liu, Jason Lee, and Michael Jordan, "A kernelized stein discrepancy for goodness-of-fit tests," in *International conference on machine learning*. PMLR, 2016, pp. 276–284.

[5] Qiang Liu and Dilin Wang, "Stein variational gradient descent: A general purpose bayesian inference algorithm," *Advances in neural information processing systems*, vol. 29, 2016.

[6] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton, "A kernel test of goodness of fit," in *International conference on machine learning*. PMLR, 2016, pp. 2606–2615.

[7] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, and Richard Zemel, "Learning the stein discrepancy for training and evaluating energy-based models without sampling," in *International Conference on Machine Learning*. PMLR, 2020, pp. 3732–3747.

[8] Yee Whye Teh, Max Welling, Simon Osindero, and Geoffrey E Hinton, "Energy-based models for sparse overcomplete representations," *Journal of Machine Learning Research*, vol. 4, no. Dec, pp. 1235–1260, 2003.

[9] Tianyang Hu, Zixiang Chen, Hanxi Sun, Jincheng Bai, Mao Ye, and Guang Cheng, "Stein neural sampler," *arXiv preprint arXiv:1810.03545*, 2018.

[10] Charles Stein, "A bound for the error in the normal approximation to the distribution of a sum of dependent random variables," in *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*. University of California Press, 1972, vol. 6, pp. 583–603.

[11] Andreas Anastasiou, Alessandro Barp, François-Xavier Briol, Bruno Ebner, Robert E Gaunt, Fatemeh Ghaderinezhad, Jackson Gorham, Arthur Gretton, Christophe Ley, Qiang Liu, et al., "Stein's method meets computational statistics: A review of some recent developments," *Statistical Science*, vol. 38, no. 1, pp. 120–139, 2023.

[12] Charles Stein, Persi Diaconis, Susan Holmes, and Gesine Reinert, "Use of exchangeable pairs in the analysis of simulations," *Lecture Notes-Monograph Series*, pp. 1–26, 2004.

[13] Matthew Repasky, Xiuyuan Cheng, and Yao Xie, "Neural stein critics with staged L2-regularization," *IEEE Transactions on Information Theory*, 2023.

[14] Arthur Jacot, Franck Gabriel, and Clément Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," *Advances in neural information processing systems*, vol. 31, 2018.

[15] Yann LeCun, Corinna Cortes, and CJ Burges, "MNIST handwritten digit database," 2010.

[16] Prajit Ramachandran, Barret Zoph, and Quoc V Le, "Searching for activation functions," in *International Conference on Learning Representations*, 2018.

[17] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.