Estimation and Sampling of Unnormalized Statistical Models with Stein Score Matching

Jiachang Liu[*] Department of ECE Duke University

Nikhil Mehta[*] Department of ECE Duke University

Chenyang Tao Department of ECE Duke University

Lawrence Carin Department of ECE Duke University

Abstract

Stein's method provides a powerful tool for handling probabilistic distributions with differential operators. Recently there has been a surge of research interest in kernelizing Stein methods, covering a wide range of important applications such as Bayesian inference [5], generative modeling [6], and goodness-of-fit tests [4], among many others. This study intends to fill a gap in the kernelized Stein literature, by addressing the problem of estimation and sampling of unnormalized statistical models based on empirical examples, a challenge shared by both the machine learning and statistics community. Experimental evidence shows encouraging results.

1 Introduction

Unnormalized statistical models provide a flexible characterization of probability distributions using energy potentials. However, their estimation is particularly challenging since the normalizing constant of the energy potential typically does not enjoy a closed form expression wrt model parameters, which prohibits the use of conventional *maximal likelihood estimation* (MLE). While most existing estimation procedures crucially rely on a tractable proposal sampler that approximates the data distribution, Hyvärinen's score matching [2] bypasses the difficulty of specifying a good proposal by (implicitly) minimizing the l^2 discrepancy between the (unknown) data score function and the model score function[2] using only data examples.

Our work considers a kernelized *Stein score matching* (SSM) formulation that avoids the costly Laplacian operator employed by Hyvärinen's original construction. Notably, our SSM differs significantly from existing kernel implicit gradient estimators [3, 7], which cannot be applied to parameter estimation tasks.

The SSM we develop can be readily adapted to learn a sampler that is faithful to the empirical examples. More explicitly, to cope with complex datasets, we model score functions with expressive non-parametric estimators such as deep neural nets. By implicitly minimizing the kernelized Stein discrepancy, SSM unveils the local geometry of data distribution, which defines the infinitesmal transition flow of an Ito diffusion process. As such, samples can be easily drawn from simulated diffusion trajectories (*e.g.*, using a Langevin or HMC simulator).

[*]Equal Contribution.

[2]In this work, we consider the score function defined by $\nabla_x \log p_\theta(x)$, where x is the data variable and θ is the parameter.

Accepted at ICML 2019 Workshop on Stein's Method for Machine Learning and Statistics

2 Background

2.1 Hyvärinen's Score Matching

Score matching was proposed in [2] for density estimation of unnormalized models. In unnormlized models, the partition function (normalizing constant) is difficult to compute, which makes the estimation task intractable using the commonly used maximum likelihood estimation (MLE) methods. In score matching (SM), the (data) score function $\boldsymbol{s}_p = \nabla_{\boldsymbol{x}} \log p(x)$ is used to define a score discrepancy metric which is minimized for the parameter estimaiton task. In particular, the score discrepancy metric for two probability distributions P (target) and Q (estimate) that possess, respectively, density functions p and q is defined as:

$$\mathcal{D}(P, Q) \triangleq \frac{1}{2} \int_{\boldsymbol{\chi}} p(x) \, ||\boldsymbol{s}_p(x) - \boldsymbol{s}_q(x)||_2^2 \, dx \tag{1}$$

$$= \mathbb{E}_{x \sim p} \left[\Delta_x \log q(x) + \frac{1}{2} \, ||\nabla_x \log q(x)||_2^2 \right] + C \tag{2}$$

where C is a constant term. In Eq. 1, we have considered P as our target distribution and Q as our estimate. Note that Eq. 2 does not involve the partition function, and other than the constant it only depends on the target distribution through the expectation. A major drawback of this approach is that, the computation involves taking the second-order derivatives, which is costly in practice.

2.2 Kernelized Stein's Discrepancy

Stein's method [8] provides a general theory on obtaining bounds on distances between two probability distributions. For a distribution P with a smooth density p, we can define a set of smooth functions (with proper boundary conditions) $\mathcal{F}$, also referred to as the Stein class of P, satisfying $\mathbb{E}_p[\boldsymbol{s}_p f(x) + \nabla_x f(x)] = 0$ where $f \in \mathcal{F}$. The stein discrepancy can then be defined as:

$$\mathbb{S}(P, Q) = \max_{f \in \mathcal{F}} (\mathbb{E}_p[\boldsymbol{s}_q(x) f(x) + \nabla_x f(x)])^2 \tag{3}$$

where q is a smooth density function of the probability distributions Q. The two distributions are equal if and only if $\mathbb{S}(P, Q) = 0$. A major drawback of this definition is that it is often computationally intractable.

There has been recent work that combines the theory of Stein discrepancies with Reproducing Kernel Hilbert Space (RKHS), introducing kernelized Stein discrepancies (KSDs) [1, 4]. In particular, KSDs restrict the Stein class of functions to an RKHS consisting of functions within a unit ball making them computationally tractable. The closed-form of KSD is defined as:

$$\mathbb{S}(p, q) = \mathbb{E}_{x, x' \sim P}[\boldsymbol{\delta}_{q,p}(x)^T k(x, x') \boldsymbol{\delta}_{q,p}(x')] \tag{4}$$

where $\boldsymbol{\delta}_{q,p}(x) = \boldsymbol{s}_q(x) - \boldsymbol{s}_p(x)$ and k(x, x') is an integrally strictly positive definite kernel.

2.3 Langevin Flow

Let X be a random variable that follows a distribution P. The Langevin Ito Diffusion equation defines a stochastic process $\{X_t\}$ as follows:

$$dX_t = \nabla \log p(X_t) + \sqrt{2} dW_t \tag{5}$$

where $\{W_t\}$ is the standard Brownian motion. The $\nabla_x \log p(x)$ is the drifting term of Langevin dynamics and $\sqrt{2} dW_t$ is the diffusion term. Together they define the infinitesimal transition dynamics of a Markov chain that characterizes the target density $p(x)$. As the system evolves, the probability distribution of $\{X_t\}$ converges to its stationary distribution p(x).

3 Stein Score Matching

For the kernelized Stein discrepancy, by iteratively applying the Stein identity, we can reformulate the KSD as:

$$\mathbb{S}(p, q) = \mathbb{E}_{x, x' \sim p}[u_q(x, x')] \tag{6}$$

where

$$u_q(x,x') = \boldsymbol{s}_q(x)^T k(x,x') \boldsymbol{s}_q(x') + \boldsymbol{s}_q(x)^T \nabla_{x'} k(x,x') + \nabla_x k(x,x')^T \boldsymbol{s}_q(x') + \text{trace}(\nabla_{x,x'} k(x,x'))$$

In practice, the kernelized Stein discrepancy is estimated using samples drawn from the distribution $p(x)$. Thus, this discrepancy can be computed using samples from the target distribution and score function from the proposed distribution. As we adjust the proposed distribution q so that $\mathbb{S}(p,q)$ becomes smaller and smaller, the proposed distribution q converges to the target distribution p. When $\mathbb{S}(p,q) = 0$, $p = q$ and $\boldsymbol{s}_p(x) = \boldsymbol{s}_q(x)$. We call this method Stein score Mmatching (SSM).

Compared with Hyvärinen's score matching, Stein score matching avoids computing the second order derivative of the density function, which makes it simpler and computationally more efficient.

We apply SSM to the task of parameter estimation with a number of representative toy models. The exact parametric form of the potential function is given, and the task is to estimate the parameter values of the model. We use 1,000 samples for training and 5,000 samples for evaluation.

The exact mathematical forms of the potential functions are summarized below.

- kidney: $\frac{1}{2}(\frac{||x||-\mu_1}{\sigma_1})^2 - \log(e^{-\frac{1}{2}(\frac{x_1-\mu_2}{\sigma_2})^2} + e^{-\frac{1}{2}(\frac{x_1+\mu_2}{\sigma_2})^2})$
- river: $-\ln(e^{-\frac{1}{2}[\frac{x_2-w_1(x;\sigma_3)}{\sigma_1}]^2} + e^{-\frac{1}{2}[\frac{x_2-w_1(x;\sigma_3)+w_3(x;\sigma_4,\mu_1)}{\sigma_2}]^2})$
- banana: $\frac{1}{2}[(\frac{(x_1-(x_2/\kappa)^2)}{\sigma_1})^2 + (\frac{(x_2-\mu_2)}{\sigma_2})^2]$
- wave: $\frac{1}{2}[\frac{x_2-w_1(x;\sigma_3)}{\sigma_1}]^2$

where $w_1(x;\sigma_3) = \sin(\frac{2\pi x_1}{\sigma_3})$ and $w_3(x;\sigma_4,\mu_1) = 3 * sigmoid(\frac{x_1-\mu_1}{\sigma_4})$

Table 1 is the mean squared error comparison of between Stein score matching and Hyvärinen's score matching. Table 2 is the runtime comparison between Stein score matching and Hyvärinen's score matching.

Table 1: MSE Comparison for Parameter Estimation (Scaled to $1e-3$)

Dataset	Kidney	River	Banana	Wave
SSM (Ours)	6.926 ± .525	1.662 ± .268	2.681 ± .563	0.436 ± .085
SM [2]	5.011 ± .455	1.759 ± .264	2.695 ± .532	0.328 ± .078

Table 2: Runtime Comparison for Parameter Estimation (Seconds)

Dataset	Kidney	River	Banana	Wave
SSM (Ours)	0.82 ± .04	1.02 ± .03	10.90 ± 1.19	3.91 ± .07
SM [2]	1.45 ± .05	1.85 ± .05	10.21 ± 1.14	1.35 ± .02

4 Stein Langevin Network

We apply SSM and Langevin dynamics on the MNIST dataset to generate new samples. To do this, we first project images $\{x_i\}_{i=1}^N$ to a low dimensional space $\{z_i\}_{i=1}^N$ (using autoencoder for dimension reduction). Then we use SSM to train a neural network to learn the score function $\hat{s_q}(z_t)$ based on samples from this low dimensional space. Finally, we generate new samples on the low dimensional space using the Langevin stochastic differential equation: $dz_t = \hat{s_q}(z_t) + \sqrt{2}dW_t$. After drawing new samples z_t's, we convert them back to the high dimensional image space through a decoder.

We first test whether the neural network would be able to learn the score function by estimating the gradient of log-density on a toy model of a 2D Gaussian distribution. Figure 1 shows the histogram of 2D Gaussian samples and the estimated and ground truth gradient field of log-density. The estimated gradient matches well with the ground truth gradient.

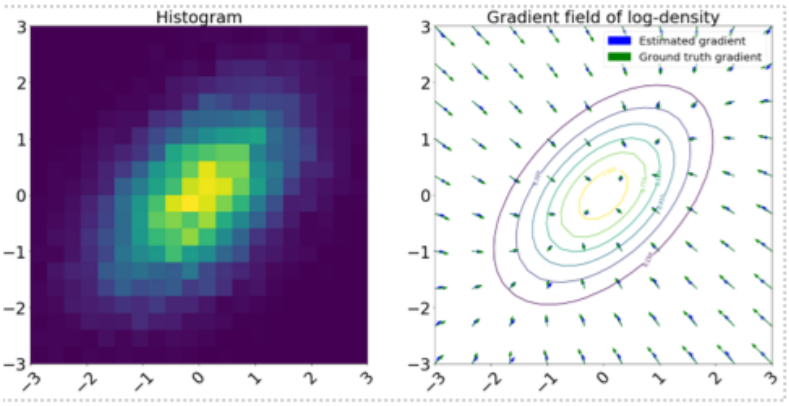

Figure 1: (Left) Histogram of 2D Gaussian samples. (Right): Estimated and ground truth gradient fields of log-density of a 2D Gaussian distribution.

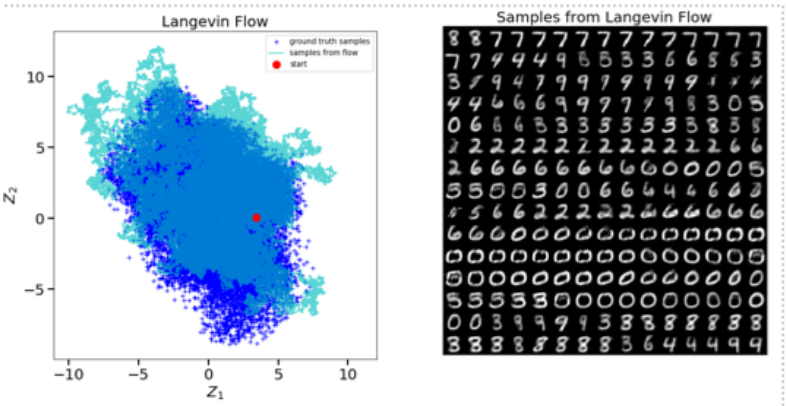

Figure 2: (Left) Samples in the latent space drawn by the Langevin flow. (Right): Image samples generated by transforming the Langevin trajectory through the decoder.

Next, we try generating new images based on the MNIST data set. Figure 2 shows the sampled trajectory of the Langevin flow and the corresponding images. We can see that the model is able to generate digits with high fidelity and diversity.

5 Conclusion and Discussion

In this work, we explore two novel applications of Stein's method and formulate a kernelized Stein score matching method. In the first part, we apply this method to estimate the parameters of a distribution based on some samples. In the second part, we use this method to estimate the score function of an empirical distribution and propose a new generative model by combining Stein score matching and the Langevin flow. The generated digits are diverse yet remains realistic.

References

[1] Kacper P. Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit. In *ICML*, 2016.

[2] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, December 2005.

[3] Yingzhen Li and Richard E. Turner. Gradient estimators for implicit models. In *International Conference on Learning Representations*, 2018.

[4] Qiang Liu, Jason D. Lee, and Michael I. Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation. *arXiv e-prints*, page arXiv:1602.03253, Feb 2016.

[5] Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2378–2386. Curran Associates, Inc., 2016.

[6] Yuchen Pu, Zhe Gan, Ricardo Henao, Chunyuan Li, Shaobo Han, and Lawrence Carin. VAE Learning via Stein Variational Gradient Descent. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4236–4245. Curran Associates, Inc., 2017.

[7] Jiaxin Shi, Shengyang Sun, and Jun Zhu. A spectral approach to gradient estimation for implicit distributions. In *ICML 2018*, 2018.

[8] 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*, pages 583–602, Berkeley, Calif., 1972. University of California Press.