

# A Spectral Approach to Gradient Estimation for Implicit Distributions

Jiaxin Shi<sup>1</sup>, Shengyang Sun<sup>2</sup> and Jun Zhu<sup>1</sup>

<sup>1</sup>Tsinghua University, <sup>2</sup>University of Toronto.

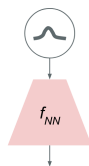
July 11, 2018

## Implicit Distributions

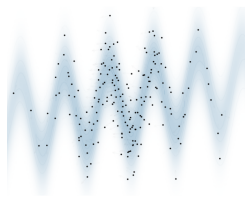
- **Implicit distributions:** Distributions defined by a sampling process but without tractable densities.
- **Examples:**
  - (a) Marginal distributions defined by a **non-conjugate** hierarchical model;
  - (b) Distributions transformed by non-invertible mappings (e.g., **neural networks**);
  - (c) **Particles** generated from a sampling algorithm (e.g., MCMC) or other nonparametric inference algorithms.



(a)  $p(x) = \int p(x|z)p(z)dz.$



(b)  $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{x} = f_{NN}(z).$



(c) Particles.

## Dealing with Intractable Densities...

- A fundamental question: Can we estimate the gradient function

$$g(\mathbf{x}) = \nabla_{\mathbf{x}} \log q(\mathbf{x}) \quad (1)$$

for any implicit distribution  $q(\mathbf{x})$ ?

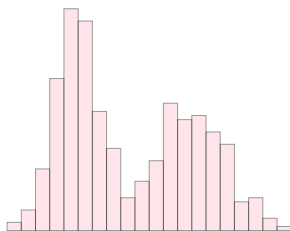
## Dealing with Intractable Densities...

- A fundamental question: Can we estimate the gradient function

$$g(\mathbf{x}) = \nabla_{\mathbf{x}} \log q(\mathbf{x}) \quad (1)$$

for any implicit distribution  $q(\mathbf{x})$ ?

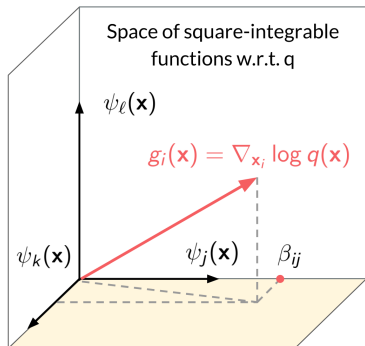
- What we have:



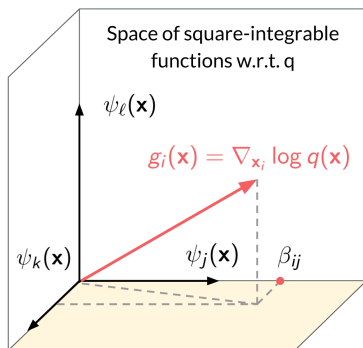
$$x^{1:M} \sim q$$

## Main Idea

- Construct an orthonormal basis  $\{\psi_j(\mathbf{x})\}_{j \geq 1}$  for the function space.



## Main Idea

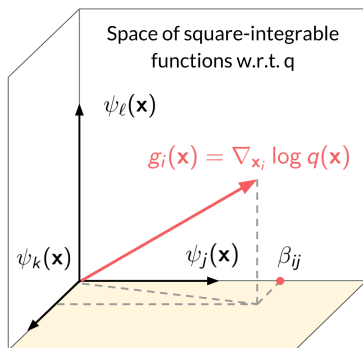


- Construct an orthonormal basis  $\{\psi_j(\mathbf{x})\}_{j \geq 1}$  for the function space.
- Expand  $g(\mathbf{x}) = \nabla_{\mathbf{x}} \log q(\mathbf{x})$  onto this basis:

$$g_i(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_{ij} \psi_j(\mathbf{x}), \quad (2)$$

where  $g_i(\mathbf{x}) = \nabla_{x_i} \log q(\mathbf{x})$  is the  $i$ th component of the gradient.

## Main Idea



- Construct an orthonormal basis  $\{\psi_j(\mathbf{x})\}_{j \geq 1}$  for the function space.
- Expand  $g(\mathbf{x}) = \nabla_{\mathbf{x}} \log q(\mathbf{x})$  onto this basis:

$$g_i(\mathbf{x}) = \sum_{j=1}^{\infty} \beta_{ij} \psi_j(\mathbf{x}), \quad (2)$$

where  $g_i(\mathbf{x}) = \nabla_{x_i} \log q(\mathbf{x})$  is the  $i$ th component of the gradient.

- Estimate the coefficients.

## An orthonormal basis by spectral decomposition of a kernel operator

Consider a p.d. kernel  $k(\mathbf{x}, \mathbf{y})$  and its spectral decomposition:

$$\int k(\mathbf{x}, \mathbf{y})\psi_j(\mathbf{y})q(\mathbf{y})d\mathbf{y} = \mu_j\psi_j(\mathbf{x}). \quad (3)$$

The eigenfunctions  $\{\psi_j\}_{j \geq 1}$  are an orthonormal basis of  $L^2(\mathcal{X}, q)$ :

$$\int \psi_i(\mathbf{x})\psi_j(\mathbf{x})q(\mathbf{x})d\mathbf{x} = \mathbb{1}(i = j). \quad (4)$$

## An orthonormal basis by spectral decomposition of a kernel operator

Consider a p.d. kernel  $k(\mathbf{x}, \mathbf{y})$  and its spectral decomposition:

$$\int k(\mathbf{x}, \mathbf{y})\psi_j(\mathbf{y})q(\mathbf{y})d\mathbf{y} = \mu_j\psi_j(\mathbf{x}). \quad (3)$$

The eigenfunctions  $\{\psi_j\}_{j \geq 1}$  are an orthonormal basis of  $L^2(\mathcal{X}, q)$ :

$$\int \psi_i(\mathbf{x})\psi_j(\mathbf{x})q(\mathbf{x})d\mathbf{x} = \mathbb{1}(i = j). \quad (4)$$

The **Nyström formula** for approximating the  $j$ th eigenfunction:

$$\hat{\psi}_j(\mathbf{x}) \approx \frac{\sqrt{M}}{\lambda_j} \sum_{m=1}^M u_{jm}k(\mathbf{x}, \mathbf{x}^m), \quad \mathbf{x}^{1:M} \sim q. \quad (5)$$

$\mathbf{u}_1, \dots, \mathbf{u}_J$  : eigenvectors of  $\mathbf{K}$  :  $\mathbf{K}_{ij} = k(\mathbf{x}^i, \mathbf{x}^j)$  with the  $J$  largest eigenvalues  $\lambda_1 \geq \dots \geq \lambda_J$ .

## Estimate the coefficients

### Generalized Stein's Identity [Gorham et al., 2015; Liu et al., 2016]

Assume that  $q(\mathbf{x})$  is a continuous differentiable density supported on  $\mathcal{X} \subset \mathbb{R}^d$ .  $\mathbf{h} : \mathcal{X} \rightarrow \mathbb{R}^{d'}$  is a smooth vector-valued function, and  $h_i$  is in the **Stein class** of  $q$ , i.e.,

$$\int_{\mathbf{x} \in \mathcal{X}} \nabla_{\mathbf{x}} (h_i(\mathbf{x})q(\mathbf{x})) d\mathbf{x} = 0. \quad (6)$$

Then the following identity holds:

$$\mathbb{E}_q[\mathbf{h}(\mathbf{x})\nabla_{\mathbf{x}} \log q(\mathbf{x})^\top + \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x})] = \mathbf{0}. \quad (7)$$

## Estimate the coefficients

### Generalized Stein's Identity [Gorham et al., 2015; Liu et al., 2016]

Assume that  $q(\mathbf{x})$  is a continuous differentiable density supported on  $\mathcal{X} \subset \mathbb{R}^d$ .  $\mathbf{h} : \mathcal{X} \rightarrow \mathbb{R}^{d'}$  is a smooth vector-valued function, and  $h_i$  is in the **Stein class** of  $q$ , i.e.,

$$\int_{\mathbf{x} \in \mathcal{X}} \nabla_{\mathbf{x}} (h_i(\mathbf{x})q(\mathbf{x})) d\mathbf{x} = 0. \quad (6)$$

Then the following identity holds:

$$\mathbb{E}_q[\mathbf{h}(\mathbf{x})\nabla_{\mathbf{x}} \log q(\mathbf{x})^\top + \nabla_{\mathbf{x}} \mathbf{h}(\mathbf{x})] = \mathbf{0}. \quad (7)$$

### Proposition

If  $k(\cdot, \cdot)$  has continuous second order partial derivatives, and both  $k(\mathbf{x}, \cdot)$  and  $k(\cdot, \mathbf{x})$  are in the Stein class of  $q$ , then:

$$\mathbb{E}_q[\psi_j(\mathbf{x})g(\mathbf{x}) + \nabla_{\mathbf{x}} \psi_j(\mathbf{x})] = \mathbf{0}, \quad j = 1, 2, \dots, \infty. \quad (8)$$

## Estimate the coefficients

$$\mathbb{E}_q[\psi_j(\mathbf{x})\mathbf{g}_i(\mathbf{x}) + \nabla_{\mathbf{x}}\psi_j(\mathbf{x})] = 0, \quad j = 1, 2, \dots, \infty.$$

## Estimate the coefficients

$$\mathbb{E}_q[\psi_j(\mathbf{x}) \sum_{\ell=1}^{\infty} \beta_{\ell} \psi_{\ell}(\mathbf{x}) + \nabla_{\mathbf{x}} \psi_j(\mathbf{x})] = 0, \quad j = 1, 2, \dots, \infty.$$

## Estimate the coefficients

$$\mathbb{E}_q[\psi_j(\mathbf{x}) \sum_{\ell=1}^{\infty} \beta_{i\ell} \psi_{\ell}(\mathbf{x}) + \nabla_{\mathbf{x}} \psi_j(\mathbf{x})] = 0, \quad j = 1, 2, \dots, \infty.$$
$$\implies \beta_{ij} = -\mathbb{E}_q \nabla_{x_i} \psi_j(\mathbf{x}).$$

## Estimate the coefficients

$$\mathbb{E}_q[\psi_j(\mathbf{x}) \sum_{\ell=1}^{\infty} \beta_{i\ell} \psi_{\ell}(\mathbf{x}) + \nabla_{\mathbf{x}} \psi_j(\mathbf{x})] = 0, \quad j = 1, 2, \dots, \infty.$$
$$\implies \beta_{ij} = -\mathbb{E}_q \nabla_{x_i} \psi_j(\mathbf{x}).$$

How to approximate  $\nabla_{x_i} \psi_j(\mathbf{x})$ ?

## Estimate the coefficients

$$\mathbb{E}_q[\psi_j(\mathbf{x}) \sum_{\ell=1}^{\infty} \beta_{i\ell} \psi_{\ell}(\mathbf{x}) + \nabla_{\mathbf{x}} \psi_j(\mathbf{x})] = 0, \quad j = 1, 2, \dots, \infty.$$

$$\implies \beta_{ij} = -\mathbb{E}_q \nabla_{x_i} \psi_j(\mathbf{x}).$$

How to approximate  $\nabla_{x_i} \psi_j(\mathbf{x})$ ?

$$\mu_j \nabla_{x_i} \psi_j(\mathbf{x}) = \nabla_{x_i} \int k(\mathbf{x}, \mathbf{y}) \psi_j(\mathbf{y}) q(\mathbf{y}) d\mathbf{y} = \int \nabla_{x_i} k(\mathbf{x}, \mathbf{y}) \psi_j(\mathbf{y}) q(\mathbf{y}) d\mathbf{y}.$$

## Estimate the coefficients

$$\mathbb{E}_q[\psi_j(\mathbf{x}) \sum_{\ell=1}^{\infty} \beta_{i\ell} \psi_{\ell}(\mathbf{x}) + \nabla_{\mathbf{x}} \psi_j(\mathbf{x})] = 0, \quad j = 1, 2, \dots, \infty.$$

$$\implies \beta_{ij} = -\mathbb{E}_q \nabla_{x_i} \psi_j(\mathbf{x}).$$

How to approximate  $\nabla_{x_i} \psi_j(\mathbf{x})$ ?

$$\mu_j \nabla_{x_i} \psi_j(\mathbf{x}) = \nabla_{x_i} \int k(\mathbf{x}, \mathbf{y}) \psi_j(\mathbf{y}) q(\mathbf{y}) d\mathbf{y} = \int \nabla_{x_i} k(\mathbf{x}, \mathbf{y}) \psi_j(\mathbf{y}) q(\mathbf{y}) d\mathbf{y}.$$

By Monte-Carlo we have an estimate of  $\nabla_{x_i} \psi_j(\mathbf{x})$ :

$$\hat{\nabla}_{x_i} \psi_j(\mathbf{x}) = \frac{1}{\mu_j M} \sum_{m=1}^M \nabla_{x_i} k(\mathbf{x}, \mathbf{x}^m) \psi_j(\mathbf{x}^m) \approx \nabla_{x_i} \hat{\psi}_j(\mathbf{x}). \quad (9)$$

Eq. (9) indicates that  $\nabla_{x_i} \hat{\psi}_j(\mathbf{x})$  is a good approximation to  $\nabla_{x_i} \psi_j(\mathbf{x})$ .

## Spectral Stein Gradient Estimator (SSGE)

Now truncating the series expansion to the first  $J$  terms and plugging in the Nyström approximations  $\{\hat{\psi}_j\}_{j=1}^J$  for eigenfunctions  $\{\psi_j\}_{j=1}^J$ :

$$\hat{\mathbf{g}}_i(\mathbf{x}) = \sum_{j=1}^J \hat{\beta}_{ij} \hat{\psi}_j(\mathbf{x}), \quad (10)$$

$$\hat{\beta}_{ij} = -\frac{1}{M} \sum_{m=1}^M \nabla_{x_i} \hat{\psi}_j(\mathbf{x}^m), \quad (11)$$

## Spectral Stein Gradient Estimator (SSGE)

Now truncating the series expansion to the first  $J$  terms and plugging in the Nyström approximations  $\{\hat{\psi}_j\}_{j=1}^J$  for eigenfunctions  $\{\psi_j\}_{j=1}^J$ :

$$\hat{g}_i(\mathbf{x}) = \sum_{j=1}^J \hat{\beta}_{ij} \hat{\psi}_j(\mathbf{x}), \quad (10)$$

$$\hat{\beta}_{ij} = -\frac{1}{M} \sum_{m=1}^M \nabla_{x_i} \hat{\psi}_j(\mathbf{x}^m), \quad (11)$$

### Theorem (Error Bound)

Given mild assumptions, the error  $\int |\hat{g}_i(\mathbf{x}) - g_i(\mathbf{x})|^2 q(\mathbf{x}) d\mathbf{x}$  is bounded by

$$J^2 \left( O_p \left( \frac{1}{M} \right) + O_p \left( \frac{C}{\mu_J \Delta_J^2 M} \right) \right) + JO_p \left( \frac{C}{\mu_J \Delta_J^2 M} \right) + \|g_i\|_{\mathcal{H}}^2 O(\mu_J),$$

where  $\Delta_J = \min_{1 \leq j \leq J} |\mu_j - \mu_{j+1}|$ ,  $O_p$  is the Big O notation in probability.

## Toy Example

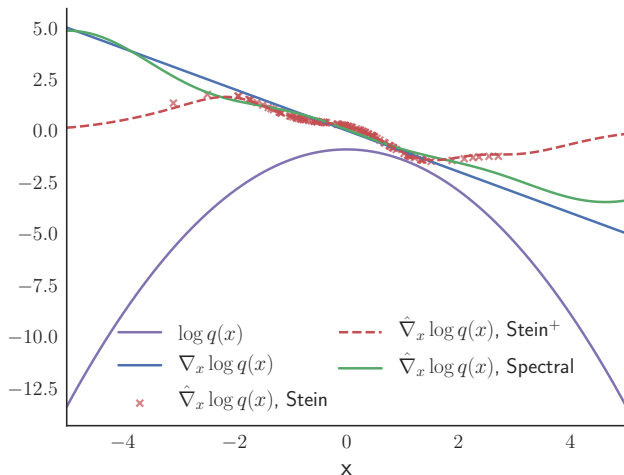
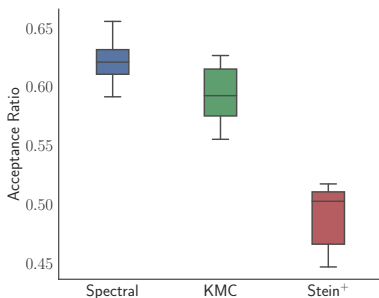


Figure: Gradient estimates of  $q(x) = \mathcal{N}(x|0, 1)$ :  $\log q(x) = -\frac{1}{2} \log 2\pi - \frac{1}{2}x^2$ .

## Gradient-free Hamiltonian Monte Carlo

**Problem:** Parameter inference for **non-conjugate** latent-variable models (e.g. Gaussian Process classification)

$$p(\theta|\mathbf{y}) \propto p(\theta) \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\theta)d\mathbf{f} \quad (12)$$



**Figure:** The average acceptance ratios of gradient-free HMC using SSGE, KMC [Strathmann et al., 2015], and Stein<sup>+</sup> [Li and Turner, 2017].

# Variational Inference with Implicit Distributions



Figure: VAE

$$\mathcal{L}(\mathbf{x}; \phi) = \mathbb{E}_{q_{\phi}(\mathbf{z})} \log p(\mathbf{z}, \mathbf{x}) + \mathbb{H}(q_{\phi}), \text{ } q \text{ is a Normal.}$$

# Variational Inference with Implicit Distributions

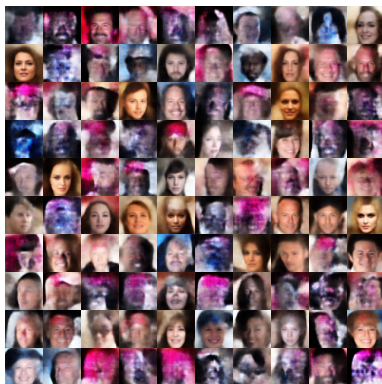


Figure: Implicit VAE, w/o entropy

$$\mathcal{L}^*(\mathbf{x}; \phi) = \mathbb{E}_{q_\phi(\mathbf{z})} \log p(\mathbf{z}, \mathbf{x}), \text{ } q \text{ is implicit.}$$

# Variational Inference with Implicit Distributions



Figure: Implicit VAE, entropy gradients estimated by SSGE.

$$\nabla_{\phi} \mathcal{L}(\mathbf{x}; \phi) \approx \nabla_{\phi} \mathbb{E}_{q_{\phi}(\mathbf{z})} \log p(\mathbf{z}, \mathbf{x}) + \nabla_{\phi}^{\text{SSGE}} \mathbb{H}(q_{\phi}), \text{ } q \text{ is implicit.}$$

## Summary

- When applying Bayesian methods to modern probabilistic models, it is often the case that we have to deal with **implicit distributions**, due to

## Summary

- When applying Bayesian methods to modern probabilistic models, it is often the case that we have to deal with **implicit distributions**, due to
  - **non-conjugate** latent structures, e.g., GP classification, deep generative models;

## Summary

- When applying Bayesian methods to modern probabilistic models, it is often the case that we have to deal with **implicit distributions**, due to
  - **non-conjugate** latent structures, e.g., GP classification, deep generative models;
  - distributions transformed by **neural networks**, e.g., Generative adversarial networks;

## Summary

- When applying Bayesian methods to modern probabilistic models, it is often the case that we have to deal with **implicit distributions**, due to
  - **non-conjugate** latent structures, e.g., GP classification, deep generative models;
  - distributions transformed by **neural networks**, e.g., Generative adversarial networks;
  - **particle-based** inference algorithms, e.g., MCMC.

## Summary

- When applying Bayesian methods to modern probabilistic models, it is often the case that we have to deal with **implicit distributions**, due to
  - **non-conjugate** latent structures, e.g., GP classification, deep generative models;
  - distributions transformed by **neural networks**, e.g., Generative adversarial networks;
  - **particle-based** inference algorithms, e.g., MCMC.
- We developed a spectral estimator for the **log-derivative function** of an implicit density.

## Summary

- When applying Bayesian methods to modern probabilistic models, it is often the case that we have to deal with **implicit distributions**, due to
  - **non-conjugate** latent structures, e.g., GP classification, deep generative models;
  - distributions transformed by **neural networks**, e.g., Generative adversarial networks;
  - **particle-based** inference algorithms, e.g., MCMC.
- We developed a spectral estimator for the **log-derivative function** of an implicit density.
- **Code is available** at

`github.com/thjashin/spectral-stein-grad`

Thanks

Poster tonight at #53