A Spectral Approach to Gradient Estimation for Implicit Distributions

Jiaxin Shi\textsuperscript{1}, Shengyang Sun\textsuperscript{2} and Jun Zhu\textsuperscript{1}

\textsuperscript{1}Tsinghua University, \textsuperscript{2}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.

\[
(a) \quad p(x) = \int p(x|z)p(z)dz.
\\(b) \quad z \sim \mathcal{N}(0, I), \quad x = f_{NN}(z).\\(c) \quad \text{Particles.}
\]
Dealing with Intractable Densities...

- A fundamental question: Can we estimate the gradient function

\[ g(x) = \nabla_x \log q(x) \]  

for any implicit distribution \( q(x) \)?
Dealing with Intractable Densities...

- A fundamental question: Can we estimate the gradient function

\[ g(x) = \nabla_x \log q(x) \]  

for any implicit distribution \( q(x) \)?

- What we have:

\[ x^{1:M} \sim q \]
Construct an orthonormal basis \( \{ \psi_j(x) \}_{j \geq 1} \) for the function space.

\[
\begin{align*}
g_i(x) &= \nabla_x \log q(x) \\
g_i(x) &= \beta_{ij} \psi_j(x)
\end{align*}
\]
Main Idea

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

- Expand \( g(x) = \nabla_x \log q(x) \) onto this basis:

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

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

Main Idea

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

\[
g_i(x) = \sum_{j=1}^{\infty} \beta_{ij} \psi_j(x), \tag{2}
\]

where \( g_i(x) = \nabla_{x_i} \log q(x) \) 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 $k(x, y)$ and its spectral decomposition:

$$\int k(x, y) \psi_j(y) q(y) dy = \mu_j \psi_j(x). \quad (3)$$

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

$$\int \psi_i(x) \psi_j(x) q(x) dx = 1(i = j). \quad (4)$$
An orthonormal basis by spectral decomposition of a kernel operator

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

\[
\int k(x, y)\psi_j(y)q(y)\,dy = \mu_j\psi_j(x). \tag{3}
\]

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

\[
\int \psi_i(x)\psi_j(x)q(x)\,dx = \mathbb{1}(i = j). \tag{4}
\]

The **Nyström formula** for approximating the \( j \)th eigenfunction:

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

\( u_1, \ldots, u_J \) : eigenvectors of \( K : K_{ij} = k(x^i, x^j) \) with the \( J \) largest eigenvalues \( \lambda_1 \geq \cdots \geq \lambda_J \).
Estimate the coefficients

**Generalized Stein’s Identity** [Gorham et al., 2015; Liu et al., 2016]

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

$$
\int_{x \in \mathcal{X}} \nabla_x (h_i(x)q(x)) \, dx = 0.
$$

(6)

Then the following identity holds:

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

(7)
Estimate the coefficients

**Generalized Stein’s Identity [Gorham et al., 2015; Liu et al., 2016]**

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

\[
\int_{x \in \mathcal{X}} \nabla_x (h_i(x)q(x)) \, dx = 0. \tag{6}
\]

Then the following identity holds:

\[
\mathbb{E}_q[h(x)\nabla_x \log q(x)^\top + \nabla_x h(x)] = 0. \tag{7}
\]

**Proposition**

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

\[
\mathbb{E}_q[\psi_j(x)g(x) + \nabla_x \psi_j(x)] = 0, \quad j = 1, 2, \ldots, \infty. \tag{8}
\]
Estimate the coefficients

\[ \mathbb{E}_q[\psi_j(x)g_i(x) + \nabla_x \psi_j(x)] = 0, \quad j = 1, 2 \ldots, \infty. \]
Estimate the coefficients

\[ \mathbb{E}_q[\psi_j(x) \sum_{\ell=1}^{\infty} \beta_{i\ell} \psi_\ell(x) + \nabla_x \psi_j(x)] = 0, \quad j = 1, 2 \ldots, \infty. \]
Estimate the coefficients

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

\[ \implies \beta_{ij} = -\mathbb{E}_q \nabla_x \psi_j(x). \]
Estimate the coefficients

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

\[ \implies \beta_{ij} = -\mathbb{E}_q \nabla_x \psi_j(x). \]

How to approximate \( \nabla_x \psi_j(x) \)?
Estimate the coefficients

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

\[ \implies \beta_{ij} = -\mathbb{E}_q \nabla_x \psi_j(x). \]

How to approximate \( \nabla_x \psi_j(x) \)?

\[ \mu_j \nabla_x \psi_j(x) = \nabla_x \int k(x, y) \psi_j(y) q(y) dy = \int \nabla_x k(x, y) \psi_j(y) q(y) dy. \]
Estimate the coefficients

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

\[ \Rightarrow \beta_{ij} = -\mathbb{E}_q \nabla_x \psi_j(x). \]

How to approximate \( \nabla_x \psi_j(x) \)?

\[ \mu_j \nabla_x \psi_j(x) = \nabla_x \int k(x, y) \psi_j(y) q(y) dy = \int \nabla_x k(x, y) \psi_j(y) q(y) dy. \]

By Monte-Carlo we have an estimate of \( \nabla_x \psi_j(x) \):

\[ \hat{\nabla}_x \psi_j(x) = \frac{1}{\mu_j M} \sum_{m=1}^{M} \nabla_x k(x, x^m) \psi_j(x^m) \approx \nabla_x \hat{\psi}_j(x). \quad (9) \]

Eq. (9) indicates that \( \nabla_x \hat{\psi}_j(x) \) is a good approximation to \( \nabla_x \psi_j(x) \).
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 $\{\psi_j\}_{j=1}^J$:

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

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

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

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

Theorem (Error Bound)

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

$$J^2 \left( O_p \left( \frac{1}{M} \right) + O_p \left( \frac{C}{\mu_j \Delta_j^2 M} \right) \right) + J O_p \left( \frac{C}{\mu_j \Delta_j^2 M} \right) + \|g_i\|_H^2 O(\mu_j),$$

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

\[
p(\theta|y) \propto p(\theta) \int p(y|f)p(f|\theta)\,df
\]  

(12)

**Figure**: The average acceptance ratios of gradient-free HMC using SSGE, KMC [Strathmann et al., 2015], and Stein\(^++\) [Li and Turner, 2017].
Variational Inference with Implicit Distributions

\[ \mathcal{L}(\mathbf{x}; \phi) = \mathbb{E}_{q_{\phi}(\mathbf{z})} \log p(\mathbf{z}, \mathbf{x}) + \mathbb{H}(q_{\phi}), \text{ q is a Normal.} \]
Variational Inference with Implicit Distributions

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

Figure: Implicit VAE, w/o entropy
Variational Inference with Implicit Distributions

\[ \nabla_\phi \mathcal{L}(x; \phi) \approx \nabla_\phi \mathbb{E}_{q_\phi}(z) \log p(z, x) + \nabla_\phi^{SSGE} \mathbb{H}(q_\phi), \text{ } q \text{ is implicit.} \]
When applying Bayesian methods to modern probabilistic models, it is often the case that we have to deal with implicit distributions, due to
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;
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;
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.
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.
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