Stein variatonal GD
vs
black-box variational inference

И. М. Латыпов

Кафедра Интеллектуальных систем МФТИ

2024

About paper

Motivation: There are two popular methods for Bayesian inference: Stein variational gradient descent (SVGD)[1] and black-box variational inference (BBVI). Are they equivalent on some meanings?
PLAN:

1. Stein variational gradient descent (SVGD).
2. Black-box variational inference (BBVI).
3. Equivalence demonstration.

Results: BBVI corresponds precisely to SVGD when the kernel is the neural tangent kernel.
Interpretation of SVGD and BBVI as kernel gradient flows and their connectivity with GANs.

SVGD

Notations:

1. Let $p(x), q(x)$ be a continuously differentiable density, supported on $\mathcal{X} \subseteq \mathbb{R}^d$.
2. $\phi(x) : \mathbb{R}^d \to \mathbb{R}^d$ – smooth vector function.

Then **Stein‘s Identity** is satisfied:

$$\mathbb{E}_{x\sim p}\mathcal{A}_p(x) = 0, \tag{1}$$

$$\mathcal{A}_p(x) = \phi(x)\nabla_x \log p(x)^T + \nabla_x \phi(x). \tag{2}$$

HINT: take the derivative of the mathematical expectation.
Define **Stein discrepancy**:

$$\mathbb{S}(q, p) = \max_{f\in\mathcal{F}} \left\{ [\mathbb{E}_{x\sim q}\text{trace}(\mathcal{A}_p\phi(x))]^2 \right\} \tag{3}$$

Kernelized Stein discrepancy on reproducing kernel Hilbert space $\mathcal{H}^d$ by Liu et al. [2]:

$$\mathbb{S}(q,p) = \max_{f \in \mathcal{H}^d} \left\{ [\mathbb{E}_{x \sim q} \text{trace}(\mathcal{A}_p \phi(x))]^2 \quad s.t. \|\phi\|_{\mathcal{H}^d} \leq 1 \right\}. \tag{4}$$

The point: there is *kernel* $k(x, x‘)$ in $\mathcal{H}^d$, and we can find optimal solution.

$$\phi(x) = \phi^*_{q,p}(x) / \|\phi^*_{q,p}(x)\|_{\mathcal{H}^d}, \tag{5}$$

$$\phi^*_{q,p}(\cdot) = \mathbb{E}_{x \sim q}[\mathcal{A}_p k(x, \cdot)] \tag{6}$$

$$\mathbb{S}(q,p) = \|\phi^*_{q,p}(x)\|_{\mathcal{H}^d}. \tag{7}$$

Var inference with Smooth Transforms

$$q^* = \arg\min_{q \in \mathcal{Q}} \{\mathrm{KL}(q||p) \equiv \mathbb{E}_q[\log q(x) - p(x)p(D|x)] + C)\} \quad (8)$$

Consider $\mathcal{Q}$ as a small evolutions:

$$x \sim q(x) \quad (9)$$

$$z = T(x) = x + \epsilon\phi(x) \quad (10)$$

Theorem 3.1. *Let* $\boldsymbol{T}(x) = x + \epsilon\boldsymbol{\phi}(x)$ *and* $q_{[\boldsymbol{T}]}(z)$ *the density of* $z = \boldsymbol{T}(x)$ *when* $x \sim q(x)$*, we have*

$$\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]} \,||\, p)\,\big|_{\epsilon=0} = -\mathbb{E}_{x\sim q}[\mathrm{trace}(\mathcal{A}_p\boldsymbol{\phi}(x))], \quad (5)$$

where $\mathcal{A}_p\boldsymbol{\phi}(x) = \nabla_x \log p(x)\boldsymbol{\phi}(x)^\top + \nabla_x\boldsymbol{\phi}(x)$ *is the Stein operator.*

Theorem 3.1. *Let* $\boldsymbol{T}(x) = x + \epsilon\boldsymbol{\phi}(x)$ *and* $q_{[\boldsymbol{T}]}(z)$ *the density of* $z = \boldsymbol{T}(x)$ *when* $x \sim q(x)$*, we have*

$$\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]} \,||\, p) \,\big|_{\epsilon=0} = -\mathbb{E}_{x\sim q}[\mathrm{trace}(\mathcal{A}_p \boldsymbol{\phi}(x))], \tag{5}$$

where $\mathcal{A}_p\boldsymbol{\phi}(x) = \nabla_x \log p(x)\boldsymbol{\phi}(x)^\top + \nabla_x\boldsymbol{\phi}(x)$ *is the Stein operator.*

Lemma 3.2. *Assume the conditions in Theorem 3.1. Consider all the perturbation directions* $\boldsymbol{\phi}$ *in the ball* $\mathcal{B} = \{\boldsymbol{\phi} \in \mathcal{H}^d \colon ||\boldsymbol{\phi}||^2_{\mathcal{H}^d} \leq \mathbb{S}(q,\ p)\}$ *of vector-valued RKHS* $\mathcal{H}^d$*, the direction of steepest descent that maximizes the negative gradient in* (5) *is the* $\boldsymbol{\phi}^*_{q,p}$ *in* (3)*, i.e.,*

$$\boldsymbol{\phi}^*_{q,p}(\cdot) = \mathbb{E}_{x\sim q}[k(x,\cdot)\nabla_x \log p(x) + \nabla_x k(x,\cdot)], \tag{6}$$

for which the negative gradient in (5) *equals KSD, that is,* $\nabla_\epsilon \mathrm{KL}(q_{[\boldsymbol{T}]} \,||\, p) \,\big|_{\epsilon=0} = -\mathbb{S}(q,\ p)$.

$$T^*(x)_l = x + \epsilon_l \cdot \phi^*_{q_l,p}(x) \tag{11}$$

$$q_{l+1} = T^*_l(q_l) \tag{12}$$

SVGD algorithm

Algorithm 1 Bayesian Inference via Variational Gradient Descent

Input: A target distribution with density function $p(x)$ and a set of initial particles $\{x_i^0\}_{i=1}^n$.
Output: A set of particles $\{x_i\}_{i=1}^n$ that approximates the target distribution.
for iteration ℓ **do**

$$x_i^{\ell+1} \leftarrow x_i^\ell + \epsilon_\ell \hat{\boldsymbol{\phi}}^*(x_i^\ell) \quad \text{where} \quad \hat{\boldsymbol{\phi}}^*(x) = \frac{1}{n}\sum_{j=1}^n \left[k(x_j^\ell, x)\nabla_{x_j^\ell} \log p(x_j^\ell) + \nabla_{x_j^\ell} k(x_j^\ell, x)\right], \quad (8)$$

where ϵ_ℓ is the step size at the ℓ-th iteration.
end for

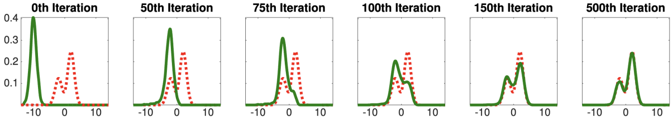

Рис.: The red dashed lines are the target density function and the solid green lines are the densities of the particles at different iterations of algorithm.

SVGD integral form

$$\frac{dx_i}{dt} = \mathbb{E}_{y\sim q_t}[k(x_i, y)\nabla_y \log p(y) + \nabla_y k(x_i, y)] \tag{13}$$

$$q_t = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)} \tag{14}$$

In limit it is equivalent to [3]:

$$\frac{dx}{dt} = \mathbb{E}_{y\sim q_t}[k(x, y)\nabla_y(\log p(y) - \log q_t(y))] \tag{15}$$

$$\tag{16}$$

BBox variational inference

ELBO maximization:

$$L(\phi) := \mathbb{E}_{x\sim q_\phi}\left[\log\frac{P(D|x)P(x)}{q_\phi(x)}\right], \tag{17}$$

$$\mathrm{KL}(q_\phi(x)||p(x)) = P(z) - L(\phi) \to \min. \tag{18}$$

ϕ dynamics:

$$\frac{d\phi}{dt} = \nabla_\phi L(\phi). \tag{19}$$

To get derivative we use reparametrization trick by Kingma:

$$x \sim q_\phi \Longleftrightarrow \varepsilon \sim \omega \text{ and } x = f_\phi(\varepsilon). \tag{20}$$

According to [2]:

$$\nabla_\phi L(\phi) = \mathbb{E}_{w\sim\omega}\nabla_\phi f_\phi(w)\cdot\nabla_y(\log(p(y) - \log(q_\phi(y))|_{y=f_\phi(w)} \tag{21}$$

BBox variational inference

We can get derivative dx/dt:

$$\frac{dx}{dt} = (\nabla_\phi f_\phi(\varepsilon))^T \frac{d\phi}{dt} = \tag{22}$$

$$\mathbb{E}_{w\sim\omega} \nabla_\phi f_\phi(\varepsilon)^T \nabla_\phi f_\phi(w) \cdot \nabla_y (\log(p(y) - \log(q_\phi(y))|_{y=f_\phi(w)} \tag{23}$$

Let's introduce *neural tangent kernel* [4]:

$$\Theta_\phi(\varepsilon, w) := \nabla_\phi f_\phi(\varepsilon)^T \nabla_\phi f_\phi(w) \tag{24}$$

$$k_\phi(x, y) := \Theta_\phi(f_\phi^{-1}(\varepsilon), f_\phi^{-1}(w)) \tag{25}$$

Finalle::

$$\frac{dx}{dt} = \mathbb{E}_{y\sim q_t}[k(x, y)\nabla_y(\log p(y) - \log q_t(y))]$$

Summary

1. Reviewed SVGD method.
2. Repeated what BBVI does and found its derivatives.
3. Show, that SVGD distribution evolution with *neural tangent kenel* is equivalent to BBVI.

finalle

[1] Qiang Liu и Dilin Wang. "Stein variational gradient descent: A general purpose bayesian inference algorithm". В: *Advances in neural information processing systems* 29 (2016).

[2] Qiang Liu, Jason Lee и Michael Jordan. "A kernelized Stein discrepancy for goodness-of-fit tests". В: *International conference on machine learning*. PMLR. 2016, с. 276—284.

[3] Jianfeng Lu, Yulong Lu и James Nolen. "Scaling limit of the Stein variational gradient descent: The mean field regime". В: *SIAM Journal on Mathematical Analysis* 51.2 (2019), с. 648—671.

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