Stein Variational Gradient Descent (SVGD)

Jaewoo Park

Department of Applied Statistics, Yonsei University

Remind: Variational Inference (VI)

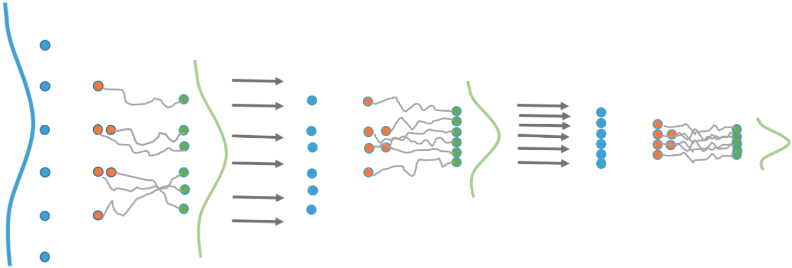
- Approximates the posterior through $q_{\xi}(\boldsymbol{W}, \boldsymbol{b})$ (variational distribution)
- Find $q_{\xi}(\boldsymbol{W}, \boldsymbol{b})$ that minimizes the Kullback–Leibler (KL) divergence between $q_{\xi}(\boldsymbol{W}, \boldsymbol{b})$ and $p(\boldsymbol{W}, \boldsymbol{b}|\boldsymbol{X}, \boldsymbol{Y})$
  - We need to set a class of distributions (e.g., Gaussian)
  - A practical option for BNN
- How can we choose a class of distributions?
  - This can be a model-by model

Main Idea

(Liu an Wang, 2016)

- Iteratively transports a set of particles to approximate the posterior
  - Perturbation direction is determined through a gradient descent
  - Functional gradient descent that minimizes the KL divergence
- Goal: Find a set of particles that represent the posterior well
- Can be applied to general Bayesian inference (not just Bayes NN)

Particle Perturbation



- Smooth transforms from a tractable reference distribution

Notation

- Posterior distribution: $\pi(\boldsymbol{\theta}) \propto p(\boldsymbol{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$
  - $\boldsymbol{\theta}$ can be neural network parameters $(\boldsymbol{W}, \boldsymbol{b})$
- Variational distribution: $q(\boldsymbol{\theta})$
- Kernel: $k(\boldsymbol{\theta}, \boldsymbol{\theta}')$: $\Theta \times \Theta \to \mathbb{R}$
  - Example: RBF kernel, $\exp(-\frac{1}{h}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|^2)$
- Arbitrary smooth function: $\phi(\boldsymbol{\theta})$

Stein's Identity

- Stein's identity (1-D case):

$$\mathbb{E}_\pi[\mathcal{A}_\pi \phi(\boldsymbol{\theta})] = \int_\Theta \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi(\boldsymbol{\theta})\phi(\boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}}\phi(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$= \int_\Theta \frac{\partial}{\partial \boldsymbol{\theta}}\Big[\pi(\boldsymbol{\theta})\phi(\boldsymbol{\theta})\Big]d\boldsymbol{\theta} = 0$$

where $\mathcal{A}_\pi$ is a stein operator (function)

- if Stein's identity holds, ϕ is in the Stein class of π

Stein's Identity (contd.)

- Stein's identity (1-D case):

$$\mathbb{E}_\pi[\mathcal{A}_\pi \phi(\boldsymbol{\theta})] = \int_\Theta \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi(\boldsymbol{\theta})\phi(\boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}}\phi(\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$= \int_\Theta \frac{\partial}{\partial \boldsymbol{\theta}}\Big[\pi(\boldsymbol{\theta})\phi(\boldsymbol{\theta})\Big]d\boldsymbol{\theta} = 0$$

- When the above identity holds?
  - $\pi(\boldsymbol{\theta})\phi(\boldsymbol{\theta}) \to 0$ as $\boldsymbol{\theta} \to \infty$ or $\pi(\boldsymbol{\theta})\phi(\boldsymbol{\theta}) = 0$ at the boundary of Θ
  - Example: $\pi(\boldsymbol{\theta})$ is a Gaussian density, $\phi(\boldsymbol{\theta}) = a\boldsymbol{\theta} + b$
  - Most situations, it holds

Stein's Discrepancy

- Maximum violation of Stein's identity for function ϕ in a set $\mathcal{F}$

$$\mathbb{S}(q, \pi) = \max_{\phi \in \mathcal{F}} \{\mathbb{E}_q[\text{trace}(\mathcal{A}_\pi \phi(\boldsymbol{\theta}))]\}$$

  where, $\mathcal{F}$ is a set of function
- Intuitively, stein's discrepancy can measure differences between π, q

Kernerlized Stein Discrepancy (KSD)

- KSD assumes that the set $\mathcal{F}$ is defined through the unit ball

$$\mathbb{S}(q,\pi) = \max_{\phi \in \mathcal{F}} \{\mathbb{E}_q[\text{trace}(\mathcal{A}_\pi \phi(\boldsymbol{\theta}))], \text{ s.t. } \|\phi\| \leq 1\}$$

- The optimal solution of KSD is

$$\phi(\boldsymbol{\theta}) = \frac{\phi^*_{q,\pi}(\boldsymbol{\theta})}{\|\phi^*_{q,\pi}\|}$$

where $\phi^*_{q,\pi}(\cdot) = E_q[(\mathcal{A}_\pi k(\boldsymbol{\theta}, \cdot))]$ and the optimal discrepancy is

$$\mathbb{S}(q,\pi) = \|\phi^*_{q,\pi}\|$$

- Note:

$$\mathcal{A}_\pi k(\boldsymbol{\theta}, \cdot) = \frac{\partial}{\partial \boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}) k(\boldsymbol{\theta}, \cdot) + \frac{\partial}{\partial \boldsymbol{\theta}} k(\boldsymbol{\theta}, \cdot)$$

  and

$$\phi^*_{q,\pi}(\cdot) = E_q[(\mathcal{A}_\pi k(\boldsymbol{\theta}, \cdot))]$$

- Then we can write down the KSD as

$$\mathbb{S}(q, \pi) = \Big\| \int_\Theta \Big[\frac{\partial}{\partial \boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}) k(\boldsymbol{\theta}, \cdot) + \frac{\partial}{\partial \boldsymbol{\theta}} k(\boldsymbol{\theta}, \cdot) \Big] q(\boldsymbol{\theta}) d\boldsymbol{\theta} \Big\|$$

Variational inference using Smooth Transforms

- VI: Find a variational dist that minimizes KL divergence with $\pi(\boldsymbol{\theta})$
- SVGD: Find set of particles through smooth transformations
  - Start with a tractable reference distribution $\boldsymbol{\theta} \sim q(\boldsymbol{\theta})$
  - $\boldsymbol{\xi} = \mathbb{T}(\boldsymbol{\theta})$, where $\mathbb{T} : \Theta \to \Theta$ is a smooth one-to-one transformation
  - By using the change of variables formula we have

$$q_{\mathbb{T}}(\boldsymbol{\xi}) = q_0(\mathbb{T}^{-1}(\boldsymbol{\xi}))\left|\frac{\mathbb{T}^{-1}(\boldsymbol{\xi})}{\partial \boldsymbol{\xi}}\right|$$

- How can we find $\mathbb{T}$ that makes $q_{\mathbb{T}}(\boldsymbol{\xi})$ close to $\pi(\boldsymbol{\theta})$?

Stein Operator as the Derivative of KL Divergence

- Let $\boldsymbol{\xi} = \mathbb{T}(\boldsymbol{\theta}) = \boldsymbol{\theta} + \epsilon\phi(\boldsymbol{\theta})$ and $q_{\mathbb{T}}(\boldsymbol{\xi})$ is it's density when $\boldsymbol{\theta} \sim q(\boldsymbol{\theta})$, we have
$$\frac{\partial}{\partial \epsilon}\mathrm{KL}(q_{\mathbb{T}}||\pi)_{\epsilon=0} = -\mathbb{E}_q[\mathrm{trace}(\mathcal{A}_\pi \phi(\boldsymbol{\theta}))]$$
and the optimal direction is
$$\phi^*_{q,p}(\cdot) = \mathbb{E}_q\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log \pi(\boldsymbol{\theta}) k(\boldsymbol{\theta}, \cdot) + \frac{\partial}{\partial \boldsymbol{\theta}} k(\boldsymbol{\theta}, \cdot)\right]$$
- Implications: $\phi^*_{q,p}(\cdot)$ is the optimal direction to minimize $\mathrm{KL}(q_{\mathbb{T}}||\pi)$

Algorithm

1. Draw a set of particles $\{\boldsymbol{\theta}_i^0\}_{i=1}^p \sim q_0(\boldsymbol{\theta})$ (e.g., prior)
2. At tth iteration, each particle $\boldsymbol{\theta}_i$ is updated as

$$\boldsymbol{\theta}_i^{t+1} \leftarrow \boldsymbol{\theta}_i^t + \epsilon \widehat{\phi^*}(\boldsymbol{\theta}_i^t)$$

   where

$$\widehat{\phi^*}(\boldsymbol{\theta}) = \frac{1}{p} \sum_{j=1}^{p} \left[\frac{\partial}{\partial \boldsymbol{\theta}_j^t} \log \pi(\boldsymbol{\theta}_j^t) k(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}_j^t} k(\boldsymbol{\theta}_j^t, \boldsymbol{\theta}) \right]$$

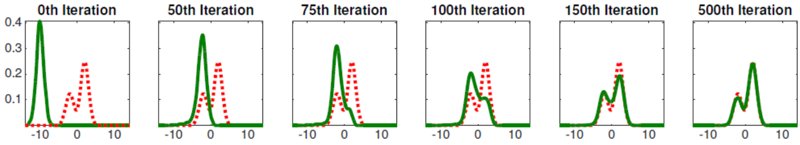
3. Repeat the above until converges

Interpretations of the Update Rule

$$\widehat{\phi^*}(\boldsymbol{\theta}) = \frac{1}{p}\sum_{j=1}^{p}\left[\frac{\partial}{\partial \boldsymbol{\theta}_j^t}\log \pi(\boldsymbol{\theta}_j^t)k(\boldsymbol{\theta}_j^t,\boldsymbol{\theta}) + \frac{\partial}{\partial \boldsymbol{\theta}_j^t}k(\boldsymbol{\theta}_j^t,\boldsymbol{\theta})\right]$$

- The first term (blue) moves particles towards the high mass of $\pi(\boldsymbol{\theta})$
  - Following a smoothed (weighted sum) gradient direction of all particles
  - Similarity is measured through $k(\boldsymbol{\theta}_j^t,\boldsymbol{\theta})$
- The second term (red) acts as a repulsive force
  - Avoid all particles to collapse together
- If we set $p = 1$, collapse to the gradient descent algorithm for MAP

A Toy Example



- Red lines indicate $\pi(\boldsymbol{\theta})$ and green lines indicate densities of the particles
- As iteration goes on, particles can approximate $\pi(\boldsymbol{\theta})$ well
- Demonstrates the ability of escaping the local mode