Stochastic Variational Gradient Descent

Jan Povala

<2020-12-11 Fri>

Outline

The problem

- Classical Bayesian setup:

$$p := \pi(x|y) = \frac{\pi(y|x)\pi(x)}{\pi(y)} \tag{1}$$

- The objective is to approximate the posterior distribution p.

SVGD – high-level overview

[Liu and Wang, 2016]

- Leverages Stein's identity to construct an efficient optimisation procedure to approximate the posterior.
- The minimisation of the KL divergence between the posterior and its approximation.
- The optimisation bypasses the computation of the normalisation constant in the posterior in (1).
- The approximating distribution is represented by 'particles'.
- Particles are updated using a specific smooth transform that corresponds to the steepest descent direction of the KL divergence.

Why is this a good idea?

What does this bring to the Bayesian landscape?

- For variational inference, the variational families that we often consider are too restrictive and one ofted needs to choose them on a model by model basis
- MCMC is just too slow for certain applications.

Stein Characterisation

Definition (Stein Characterisation)

A measure P on $\mathcal{X} \subset \mathbb{R}^d$ with density p is characterised by the pair $(\mathcal{A}, \mathcal{F})$, consisting of a Stein Operator $\mathcal{A}$ and a Stein Class $\mathcal{F}$, if it holds

$$x \sim P \quad \text{iff} \quad \mathbb{E}_P[\mathcal{A}_P \phi(x)] = 0 \quad \forall \phi \in \mathcal{F},$$

where $\phi(x) = [\phi_1(x), \cdots, \phi_d(x)]^\top$.
All the papers I've seen use the following Stein operator

$$\mathcal{A}_p \phi(\cdot) = \nabla_x \cdot \phi(\cdot) + \phi(\cdot) \cdot \nabla_x \log p(\cdot).$$

Example

Example (Stein, 1972)

- $P = N\left(\mu, \sigma^2\right)$ with density function $p(x)$
- $\mathcal{A} : f \mapsto \frac{\nabla(fp)}{p}$
- $\mathcal{F} = \left\{f : \mathbb{R} \to \mathbb{R} \text{ s.t. } fp \in W^{1,1} \text{ and } \lim_{x \searrow -\infty} f(x)p(x) = \lim_{x \nearrow +\infty} f(x)p(x)\right\}$

Stein Discrepancy

By having two distributions p and q defined on $\mathcal{X}$, we can compute $\mathbb{E}_{x\sim q}\left[\mathcal{A}_p\phi(x)\right]$. This would no longer be 0, unless $p = q$. We can use this fact to define the discrepancy measure between those two distributions:

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

This discrepancy measure seeks to find the function ϕ from the Stein class $\mathcal{F}$ that 'violates' the Stein's identity the most.

Kernel Stein Discrepancy

To make this 'search' tenable, we restrict the Stein class to be a unit ball of an RKHS $\mathcal{H}^d$. In this case, the optimisation has a closed form solution:

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

where

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

for which we have

$$\mathbb{S}(q, p) = \left\| \phi^*_{q,p} \right\|^2_{\mathcal{H}^d}$$

The space of functions obtained by applying $\mathcal{A}$ to the unit ball of $\mathcal{H}^d$ with kernel k we obtain an RKHS with the kernel k_0 [Oates et al., 2017]:

$$\begin{aligned} k_0\left(x, x'\right) := \left(\nabla_x \cdot \nabla_{x'}\right) k\left(x, x'\right) + \left(\nabla_x \log p(x)\right) \cdot \left(\nabla_{x'} k\left(x, x'\right)\right) \\ + \left(\nabla_{x'} \log p\left(x'\right)\right) \cdot \left(\nabla_x k\left(x, x'\right)\right) \\ + \left(\nabla_x \log p(x)\right) \cdot \left(\nabla_{x'} \log p\left(x'\right)\right) k\left(x, x'\right) \end{aligned}$$

What can we use the results above for?

The general VI framework seeks a distribution q^* to approximate the target posterior p:

$$q^* = \arg\min_{q \in \mathcal{Q}} \left\{ \mathrm{KL}(q \| p) \equiv \mathbb{E}_q[\log q(x)] - \mathbb{E}_q[\log(\pi(y|x)\pi(x))] + \log p(y) \right\}$$

To perform the optimisation, the paper proposes using smooth one-to-one transforms $z = \boldsymbol{T}(x)$ to explore the space, where $\boldsymbol{T} \colon \mathcal{X} \to \mathcal{X}$ where x is drawn from the reference distribution $q_0(x)$.

Choosing the transform

If we let $\boldsymbol{T}(x) = x + \epsilon\phi(x)$, the paper shows that:

Theorem (Steepest Descent)

Let $q_{[\mathrm{T}]}(z)$ be the density of $z = \mathrm{T}(x)$ when $x \sim q(x)$, then we have

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

where $\mathcal{A}_p\phi(x)$ is the Stein operator.

But we've seen the RHS before and we know how to choose $\mathcal{A}_p$ and $\phi(x)$ to maximise it.

Finite sample approximation

To obtain the ϕ that maximises the discrepancy we approximate the expectation

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

using the empirical distribution of q, represented by n particles:

$$\hat{\phi}^*_{q,p}(x) = \frac{1}{n}\sum_{j=1}^{n}\left[k\left(x_j,x\right)\nabla_{x_j}\log p\left(x_j\right) + \nabla_{x_j} k\left(x_j,x\right)\right]$$

The 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**

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

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

Figure: The algorithm

Example 1

1000 particles to approximate q.

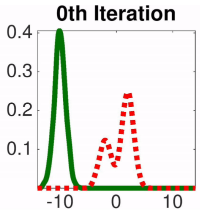

Figure: Initial q

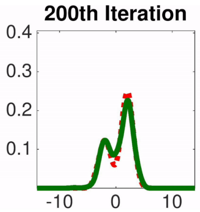

Figure: q after 200 iterations

Example 2

10 particles to approximate q

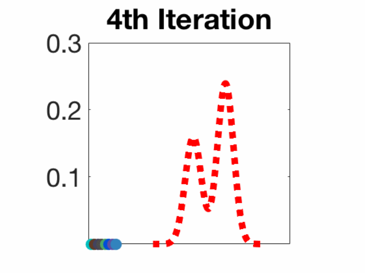

Figure: Initial q

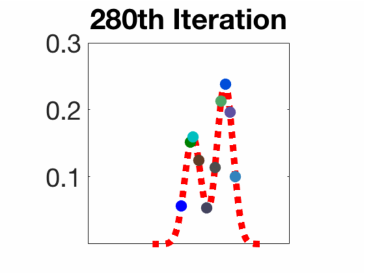

Figure: q after 280 iterations

Discussion

- SVGD as a gradient flow of the KL divergence functional in the space of probability measures metrized by a RKHS variant of Wasserstein distance.
- A follow-up work proves that as the number of particles and the number of steps go to infinity, the approximation converges weakly to the posterior measure.
- Intuition why ϕ^* does maximise Kernel Stein Discrepancy.

Bibliography I

On top of the below, I've made use of Chris Oates's presentation at MCQMC 2020.

Liu, Q. and Wang, D. (2016).
Stein variational gradient descent: a general purpose bayesian inference algorithm.
Alsow worth checking out this one: https://www.cs.utexas.edu/ qlearning/project.html?p=svgd Extra material: https://arxiv.org/pdf/2004.01822.pdf,.

Oates, C. J., Girolami, M., and Chopin, N. (2017).
Control functionals for monte carlo integration.
Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(3):695–718.