Info Geom & Grad Flow Reading Group

19/09/2024

Paper: Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm.
Liu Qiang, Wang Dilin (NeurIPS 2016)

Motivation

Compute $\int f(x) \underbrace{p(x)}_{\text{probability density function}} dx$.

① Intractable integral → no (easy) analytic solution

② (Potentially) high-dimensional x → grid method fails, curse of dimensionality

Example: posterior mean in Bayesian inference.

Notice that $\int f(x) p(x) dx = \mathbb{E}_{X\sim p}[f(X)]$.

exact:

$$\approx \frac{1}{N} \sum_{i=1}^{N} f(X_i^e)$$

with $X_1^e, X_2^e, \ldots, X_N^e \sim p$.

(not necessarily independent)

e.g. MCMC, SMC

approximate:

$$\approx \frac{1}{N} \sum_{i=1}^{N} f(X_i^a)$$

with $X_1^a, X_2^a, \ldots X_N^a \sim \tilde{p} \approx p$

e.g. ABC, VI

SVGD ⊂ Particle VI ⊂ VI.

Get some particles. Move them around. Use their empirical distribution $\tilde{p}$ to approximate p.

↳ Key Question.

Kernel Stein Discrepancy

Sometimes, we wish to measure the similarity of two distributions p, q.

e.g. $KL(p \| q)$, $TV(p \| q)$, $W^2(p, q)$.

① couldn't be computed easily. → not practical

② require knowing the densities → we often only have samples ...

Applications: Goodness of fit, Variational inference.

One computable measure of distribution distance is the Kernel Stein Discrepancy (KSD)

$$KSD^2(p \| q) = \sup_{\substack{f \in \mathcal{H}_k \\ \|f\|_{\mathcal{H}_k} \le 1}} \left| \mathbb{E}_{X \sim q}[S_p f(X)] \right|^2 = \mathbb{E}_{X, Y \sim q}[k_p(X, Y)] \approx \frac{1}{n} \sum_{i,j=1}^{n} k_p(x_i, x_j).$$

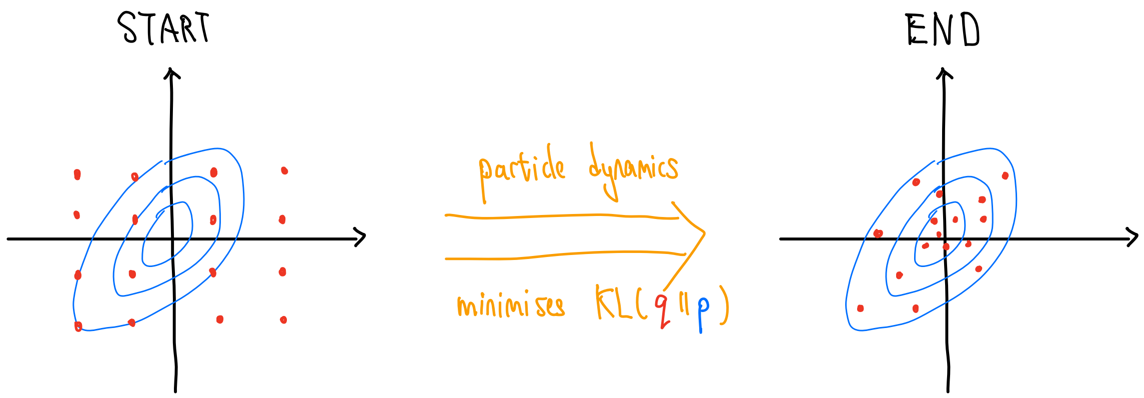
unit ball in RKHS

(Langevin) Stein operator $S_p f(x) = \langle \nabla_x \log p(x), f(x) \rangle + \langle \nabla, f(x) \rangle$

$\nabla_x \log p(x)$, $\nabla_y \log p(y)$ (Stein) score function

$$k_p(x, y) = S_p(x) S_p(y) k(x, y) + S_p(x) \nabla_y k(x, y) + S_p(y) \nabla_x k(x, y) + tr(\nabla_x \nabla_y k(x, y)).$$

SVGD



<u>Key Result</u> (THM 3.1 & Lemma 3.2 of Liu & Wang (2016)).

Consider updates of the form $x_i \leftarrow T(x_i) = x_i + \varepsilon \phi(x_i)$. q denote the empirical distribution of the particles $\{x_i\}$. $q_{[T]}$ denote the empirical distribution after moving the particles by T. p is the target.

Then,
$$\nabla_\varepsilon KL(q_{[T]} \| p) \Big|_{\varepsilon=0} = -\mathbb{E}_{x\sim q}\left[tr\left(S_p \phi(x) \right) \right].$$

If we force $\phi \in \mathcal{H}_k$ & $\|\phi\|_{\mathcal{H}_k} \le 1$, i.e. ϕ is in the unit ball of RKHS, then we have

$$\underset{\substack{\phi \in \mathcal{H}_k \\ \|\phi\|_{\mathcal{H}_k} \le 1}}{\arg\max} \nabla_\varepsilon KL(q_{[T]} \| p) \Big|_{\varepsilon=0} = \phi^*(x') = \mathbb{E}_{x'\sim q}\left[S_p k(x, x') \right].$$
$$= \mathbb{E}_{x'\sim q}\left[k(x, x')^T \nabla_{x'} \log p(x') + \nabla_{x'} k(x', x) \right].$$

Algorithm 1 SVGD

Require: Target distribution p. Initial particles $\{x_i^0\}_{i=1}^n$. Kernel k. Step size ε..

for $l = 0, 1, \ldots$ **do**

  for $i = 1, 2, \ldots, n$ **do**

$$x_i^{l+1} \leftarrow x_i^l + \frac{\varepsilon}{n} \sum_{j=1}^n \left[k(x_j^l, x_i^l) \nabla_{x_j^l} \log p(x_j^l) + \nabla_{x_j^l} k(x_j^l, x_i^l) \right].$$

  end for

end for