MATHÉMATIQUES
VISION
APPRENTISSAGE

Stein Variational Gradient Descent main ideas

Gaëtan Serré

École Normale Supérieure Paris-Saclay
Master Mathématiques, Vision, Apprentissage
gaetan.serre@ens-paris-saclay.fr

1 Goal

Given a smooth density π supported on $\mathcal{X} \subseteq \mathbb{R}^d$, find μ on $\mathcal{X}$ as close as possible to π.

2 Stein framework

Stein identity: Let $\mathcal{A}_\pi$ a Stein operator s.t.

$$\mathcal{A}_\pi \phi = \nabla \log \pi(\cdot)^\top \phi(\cdot) + \nabla \cdot \phi(\cdot)$$

with $\phi(x) = [\phi_1(x), ..., \phi_d(x)]^\top$. Then, if ϕ is in the Stein class f π i.e. $\phi(x)\pi(x) = 0$ for all $x \in \partial\mathcal{X}$ if $\mathcal{X}$ is compact or $\lim_{||x|| \to \infty} \phi(x)\pi(x) = 0$ if $\mathcal{X} = \mathbb{R}^d$, we have:

$$\mathbb{E}_{x\sim\pi}[\mathcal{A}_\pi \phi(x)] = 0 \tag{1}$$

Proof.

$$\begin{aligned}
\mathbb{E}_{x\sim\pi}[\mathcal{A}_\pi \phi(x)] &= \int_{\mathcal{X}} \big(\nabla \log \pi(\cdot)^\top \phi(\cdot) + \nabla \cdot \phi(\cdot)\big)\pi(x)dx \\
&= \int_{\mathcal{X}} \nabla \log \pi(\cdot)^\top \phi(\cdot)\pi(x)dx + \int_{\mathcal{X}} \nabla \cdot \phi(\cdot)\pi(x)dx \\
&= \int_{\mathcal{X}} \nabla \log \pi(\cdot)^\top \phi(\cdot)\pi(x)dx + \int_{\mathcal{X}} \sum_{k=1}^{d} \frac{\partial \phi_k}{\partial x_k}\pi(x)dx \\
&= \int_{\mathcal{X}} \nabla \log \pi(\cdot)^\top \phi(\cdot)\pi(x)dx + \sum_{k=1}^{d} \left([\pi(x)\phi_k(x)]_{\mathcal{X}} - \int_{\mathcal{X}} \frac{\partial \pi(x)}{\partial x_k}\phi_k(x)dx \right) \\
&= \int_{\mathcal{X}} \nabla \log \pi(\cdot)^\top \phi(\cdot)\pi(x)dx - \int_{\mathcal{X}} \sum_{k=1}^{d} \frac{\partial \pi(x)}{\partial x_k}\phi_k(x)dx \\
&= \int_{\mathcal{X}} \pi(x) \sum_{k=1}^{d} \frac{\partial \log \pi(x)}{\partial x_k}\phi_k(x) - \pi(x) \sum_{k=1}^{d} \frac{\partial \log \pi(x)}{\partial x_k}\phi_k(x)dx \quad \text{(log trick)} \\
&= 0
\end{aligned}$$

■

Now, let μ a smooth density supported on $\mathcal{X}$ different from π. Now, Eq. 1 do not hold anymore with $x \sim \mu$. However, we can use $\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi \phi(x)]$ as a discrepancy measure between μ and π, as its magnitude relates to how different μ and π are (see Liu and Wang [2016] & Liu [2017]). The objective becomes:

$$\mu^* = \arg\min_{\mu} \mathbb{S}(\mu, \pi) = \arg\min_{\mu} \max_{\phi \in \mathcal{H}} \{\mathbb{E}_{x\sim\mu}[\mathcal{A}_\pi \phi(x)]\} \tag{2}$$

As $\mathbb{S}(\mu, \pi) = 0$ iff $\mu = \pi$ and $\mathbb{S}(\mu, \pi) > 0$ otherwise with $\mathcal{H}$ sufficiently large. The choice of $\mathcal{H}$ is therefore crucial. One way to ensure it is both rich enough and computationally tractable is to let $\mathcal{H}$ be a RKHS.

Bibliography

Qiang Liu. Stein variational gradient descent as gradient flow, 2017. URL https://arxiv.org/abs/1704.07520.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm, 2016. URL https://arxiv.org/abs/1608.04471.