

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}_π a Stein operator s.t.

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

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

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

Proof.

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

■

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]). Indeed, if we assume ϕ to be in the Stein class of μ as well (this is mild condition as π and μ are two densities on \mathcal{X} , one can choose ϕ to be in the Stein class of all distribution on \mathcal{X} . E.g. if $\mathcal{X} = \mathbb{R}^d$, one can pick

$\phi(x) = \exp[-\|x - y\|^2]$, we have:

$$\begin{aligned}
\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)] &= \int_{\mathcal{X}} (\nabla \log \pi(x)^\top \phi(x) + \nabla \cdot \phi(x)) \mu(x) dx \\
&= \int_{\mathcal{X}} \nabla \log \pi(x)^\top \phi(x) \mu(x) dx + \sum_{k=1}^d \left([\mu(x) \phi_k(x)]_{\mathcal{X}} - \int_{\mathcal{X}} \frac{\partial \mu(x)}{\partial x_k} \phi_k(x) dx \right) \\
&= \int_{\mathcal{X}} \mu(x) \sum_{k=1}^d \frac{\partial \log \pi(x)}{\partial x_k} \phi_k(x) - \mu(x) \sum_{k=1}^d \frac{\partial \log \mu(x)}{\partial x_k} \phi_k(x) dx \quad (\text{log trick}) \\
&= \int_{\mathcal{X}} \mu(x) \left[\sum_{k=1}^d \phi_k(x) \left(\frac{\partial \log \pi(x)}{\partial x_k} - \frac{\partial \log \mu(x)}{\partial x_k} \right) \right] dx \\
&= \int_{\mathcal{X}} \mu(x) \left[\sum_{k=1}^d \phi_k(x) \left(\frac{\partial \log \frac{\pi(x)}{\mu(x)}}{\partial x_k} \right) \right] dx.
\end{aligned} \tag{2}$$

As expected, the scale of $\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)]$ increases w.r.t. the difference between μ and π .

Therefore, one can define an objective to find a density μ^* close to π :

$$\mu^* = \arg \min_{\mu} \mathbb{S}(\mu, \pi) = \arg \min_{\mu} \max_{\phi \in \mathcal{H}} \{\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)]\}, \tag{3}$$

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.

2.1 Kernelized Stein Discrepancy

Let \mathcal{H}_0 be a RKHS with a kernel $k(x, x')$ in the Stein class of π and μ . Let $\mathcal{H} = (\mathcal{H}_0^{(1)}, \dots, \mathcal{H}_0^{(d)})$. The KSD maximizes ϕ in the unit ball of \mathcal{H} . The objective in (3) is then:

$$\mathbb{S}(\mu, \pi) = \max_{\phi \in \mathcal{H}} \{\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)], \text{ s.t. } \|\phi\|_{\mathcal{H}} \leq 1\}. \tag{4}$$

Within this framework, one can show that the optimal solution of (4) is given by:

$$\phi(x) = \frac{\phi^*(x)}{\|\phi^*\|_{\mathcal{H}}}, \text{ where } \phi^*(\cdot) = \mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \otimes k(x, \cdot)] = \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) + \nabla k(x, \cdot) d\mu(x), \tag{5}$$

where $\mathcal{A}_\pi \otimes f(x) = f(x) \nabla \log \pi(x) + \nabla f(x)$, is a variant of Stein operator. Moreover, $\mathbb{S}(\mu, \pi) = \|\phi^*\|_{\mathcal{H}}$.

Proof. We first need to prove that

$$\begin{aligned}
\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi f(x)] &= \langle f, \phi^* \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H} : \\
\langle f, \phi^* \rangle_{\mathcal{H}} &= \sum_{l=1}^d \left\langle f^{(l)}, \mathbb{E}_{x \sim \mu} \left[k(x, \cdot) \nabla \log \pi(x)^{(l)} + \nabla k(x, \cdot)^{(l)} \right] \right\rangle_{\mathcal{H}^0} \\
&= \mathbb{E}_{x \sim \mu} \left[\sum_{l=1}^d \left\langle f^{(l)}, k(x, \cdot) \nabla \log \pi(x)^{(l)} + \nabla k(x, \cdot)^{(l)} \right\rangle_{\mathcal{H}^0} \right] \\
&= \mathbb{E}_{x \sim \mu} \left[\sum_{l=1}^d \nabla \log \pi(x)^{(l)} \left\langle f^{(l)}, k(x, \cdot) \right\rangle_{\mathcal{H}^0} + \left\langle f^{(l)}, \nabla k(x, \cdot)^{(l)} \right\rangle_{\mathcal{H}^0} \right] \\
&= \mathbb{E}_{x \sim \mu} \left[\sum_{l=1}^d \nabla \log \pi(x)^{(l)} f^{(l)}(x) + \nabla_{x_l} f(x)^{(l)} \right] \quad (\text{see Zhou [2008]}) \\
&= \mathbb{E}_{x \sim \mu} [\nabla \log \pi(x)^\top f(x) + \nabla \cdot f(x)] \\
&= \mathbb{E}_{x \sim \mu} [\mathcal{A}_\pi f(x)].
\end{aligned} \tag{6}$$

Moreover, $\langle f, \phi^* \rangle_{\mathcal{H}} \leq \|f\|_{\mathcal{H}} \|\phi^*\|_{\mathcal{H}}$. Thus,

$$\mathbb{S}(\mu, \pi) = \max_{f \in \mathcal{H}} \{\mathbb{E}_{x \sim \mu}[\mathcal{A}_{\pi} f(x)] = \langle f, \phi^* \rangle_{\mathcal{H}}, \text{ s.t. } \|f\|_{\mathcal{H}} \leq 1\} \leq \|\phi^*\|_{\mathcal{H}}.$$

Let $f = \frac{\phi^*}{\|\phi^*\|_{\mathcal{H}}}$, then $\|f\|_{\mathcal{H}} = 1$ and

$$\mathbb{E}_{x \sim \mu}[\mathcal{A}_{\pi} \phi(x)] = \langle f, \phi^* \rangle_{\mathcal{H}} = \|\phi^*\|_{\mathcal{H}},$$

ending the proof. ■

3 Link with Kullback-Leibler Divergence

Let $T : \mathcal{X} \rightarrow \mathcal{X}$, $x \mapsto (I + \gamma\phi)(x)$. One can show that (see Liu and Wang [2016] Theorem 3.1):

$$\nabla_{\gamma} KL(T_{\#}\mu || \pi) = -\mathbb{E}_{x \sim \mu}[\mathcal{A}_{\pi} \phi(x)]. \quad (7)$$

Therefore, assuming $\phi \in \mathcal{H}$ with \mathcal{H} as defined as in Section 2.1, using (5), we know that:

$$\phi^*(\cdot) = \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) + \nabla k(x, \cdot) d\mu(x) \quad (8)$$

minimizes $\nabla_{\gamma} KL(T_{\#}\mu || \pi)$. Furthermore, one can show that:

$$\begin{aligned} P_{\mu} \nabla \log \frac{\mu}{\pi}(\cdot) &= \int_{\mathcal{X}} k(x, \cdot) \nabla \log \mu(x) d\mu(x) - \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) d\mu(x) \\ &= \int_{\mathcal{X}} k(x, \cdot) \nabla \mu(x) dx - \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) d\mu(x) \\ &= - \int_{\mathcal{X}} \nabla k(x, \cdot) d\mu(x) - \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) d\mu(x) \\ &= - \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) + \nabla k(x, \cdot) d\mu(x) \\ &= -\phi^*(\cdot) \end{aligned} \quad (9)$$

The Stein Variational Gradient Descent (SVGD) algorithm consists in an iterative procedure where one apply successive transformations to an initial density μ_0 towards the "direction" ϕ^* that minimizes the gradient of the Kullback-Leibler divergence:

$$\mu_{n+1} = (I + \gamma\phi^*)_{\#} \mu_n = \left(I - \gamma P_{\mu} \nabla \log \frac{\mu}{\pi} \right)_{\#} \mu_n. \quad (10)$$

4 Not understood yet

- $k(x, \cdot)$ in the Stein class of π and μ ?
- Link with Wasserstein distance?
- Why did they defined so much about their 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>.
- Ding-Xuan Zhou. Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, 220(1-2):456–463, October 2008. doi: 10.1016/j.cam.2007.08.023. URL <https://doi.org/10.1016/j.cam.2007.08.023>.

A Lemmas

Lemma 1. *Let two distributions μ and π on $\mathcal{X} \subseteq \mathbb{R}^d$. Let $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$ be in the Stein class of μ and $\mathcal{A}_\pi \phi(x) = \nabla \log \pi(x)^\top \phi(x) + \nabla \cdot \phi(x)$. Then,*

$$\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)] = \mathbb{E}_{x \sim \mu}[(\nabla \log \pi(x) - \nabla \log \mu(x))^\top \phi(x)]$$

Proof.

$$\begin{aligned} \mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)] &= \mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x) - \mathcal{A}_\mu \phi(x)] \\ &= \mathbb{E}_{x \sim \mu}[\nabla \log \pi(x)^\top \phi(x) - \nabla \cdot \phi(x) - \nabla \log \mu(x)^\top \phi(x) + \nabla \cdot \phi(x)] \\ &= \mathbb{E}_{x \sim \mu}[(\nabla \log \pi(x) - \nabla \log \mu(x))^\top \phi(x)] \end{aligned}$$

■

Lemma 2. *Let two distributions μ and π on $\mathcal{X} \subseteq \mathbb{R}^d$. Let $\phi : \mathcal{X} \rightarrow \mathbb{R}$ be in the Stein class of μ and $\mathcal{A}_\pi \otimes \phi(x) = \phi(x) \nabla \log \pi(x) + \nabla \phi(x)$. Then,*

$$\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \otimes \phi(x)] = \mathbb{E}_{x \sim \mu}[(\nabla \log \pi(x) - \nabla \log \mu(x)) \phi(x)]$$

Proof. Same as Lemma 1.

■