

# 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  $\pi$  supported on  $\mathcal{X} \subseteq \mathbb{R}^d$ , find  $\mu$  on  $\mathcal{X}$  as close as possible to  $\pi$ .

## 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  $\phi(x) = [\phi_1(x), \dots, \phi_d(x)]^\top$ . Then, if  $\phi$  is in the Stein class of  $\pi$  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  $\mu$  a smooth density supported on  $\mathcal{X}$  different from  $\pi$ . 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  $\mu$  and  $\pi$ , as its magnitude relates to how different  $\mu$  and  $\pi$  are (see Liu and Wang [2016] & Liu [2017]). Indeed, if we assume  $\phi$  to be in the Stein class of  $\mu$  as well (this is mild condition as  $\pi$  and  $\mu$  are two densities on  $\mathcal{X}$ , one can choose  $\phi$  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  $\mu$  and  $\pi$ .

Therefore, one can define an objective to find a density  $\mu^*$  close to  $\pi$ :

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

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  $\pi$  and  $\mu$ . Let  $\mathcal{H} = (\mathcal{H}_0^{(1)}, \dots, \mathcal{H}_0^{(d)})$ . The KSD maximizes  $\phi$  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) (see [Liu et al., 2016, Oates et al., 2014, Chwialkowski et al., 2016]) is <sup>1</sup>:

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

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

## 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):

$$\nabla_{\gamma} KL(T_{\#} \mu || \pi) = -\mathbb{E}_{x \sim \mu}[\mathcal{A}_\pi \phi(x)]. \tag{6}$$

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

$$\phi^*(\cdot) = \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) + \nabla k(x, \cdot) d\mu(x) \tag{7}$$

<sup>1</sup> J'ai fait la preuve sur papier, je l'écrirai plus tard. (définir  $\mathcal{B}$  et utiliser ineq triangulaire sur  $\langle \mathcal{B}, f \rangle_{\mathcal{H}}$ )

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

$$\begin{aligned}
P_\mu \nabla \log \frac{\mu}{\pi}(\cdot) &= \int_{\mathcal{X}} k(x, \cdot) \nabla \log \mu(x) d\mu(x) - \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) d\mu(x) \\
&= \int_{\mathcal{X}} k(x, \cdot) \nabla \mu(x) dx - \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) d\mu(x) \\
&= - \int_{\mathcal{X}} \nabla k(x, \cdot) d\mu(x) - \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) d\mu(x) \\
&= - \int_{\mathcal{X}} k(x, \cdot) \nabla \log \pi(x) + \nabla k(x, \cdot) d\mu(x) \\
&= -\phi^*(\cdot)
\end{aligned} \tag{8}$$

The Stein Variational Gradient Descent (SVGD) algorithm consists in an iterative procedure where one apply successive transformations to an initial density  $\mu_0$  towards the "direction"  $\phi^*$  that minimizes the gradient of the Kullback-Leibler divergence:

$$\mu_{n+1} = (I + \gamma \phi^*)_{\#} \mu_n = \left( I - \gamma P_\mu \nabla \log \frac{\mu}{\pi} \right)_{\#} \mu_n. \tag{9}$$

#### 4 Not understood yet

- $k(x, \cdot)$  in the Stein class of  $\pi$  and  $\mu$ ?
- Link with Wasserstein distance?
- Why did they defined so much about their RKHS?

## Bibliography

- Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness of fit, 2016. URL <https://arxiv.org/abs/1602.02964>.
- 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>.
- Qiang Liu, Jason D. Lee, and Michael I. Jordan. A kernelized stein discrepancy for goodness-of-fit tests and model evaluation, 2016. URL <https://arxiv.org/abs/1602.03253>.
- Chris J. Oates, Mark Girolami, and Nicolas Chopin. Control functionals for monte carlo integration, 2014. URL <https://arxiv.org/abs/1410.2392>.