Stein Variational Gradient Descent (Qiang Liu, Dartmouth)

Given a set of particles $\{x_i\}_{i=1}^n$, iteratively transport it to match target p by minimizing KL divergence.

Stein Variational Gradient Descent

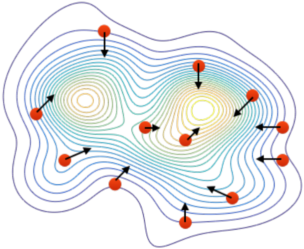
- Idea: Iteratively move $\{x_i\}_{i=1}^n$ towards the target p by updates of form

$$x_i' \leftarrow x_i + \epsilon \phi(x_i),$$

where ϕ is a "particle gradient direction" chosen to maximally decrease the KL divergence with p, that is,

$$\phi = \underset{\phi \in \mathcal{F}}{\arg\max} \left\{ -\frac{\partial}{\partial \epsilon} \mathrm{KL}(q' \parallel p)\big|_{\epsilon=0} \right\},$$

where q' is the density of $x' = x + \epsilon \phi(x)$ when the density of x is q.



Stein Variational Gradient Descent

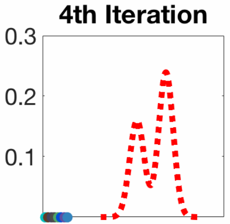
Our algorithm:

$$x_i \leftarrow x_i + \epsilon \hat{\mathbb{E}}_{x \sim \{x_i\}_{i=1}^n} [\underbrace{\nabla_x logp(x)}_{\text{gradient}} k(x, x_i) + \underbrace{\nabla_x k(x, x_i)}_{\text{repulsive force}}], \quad \forall i = 1, \ldots, n.$$

Two terms:

- $\nabla_x logp(x)$: moves the particles $\{x_i\}$ towards high probability regions of $p(x)$.
- $\nabla_x k(x, x')$: enforce diversity in $\{x_i\}$ (otherwise all x_i collapse to modes of $p(x)$).
- Reduces to **gradient ascent for MAP** when using a single particle ($n = 1$):

$$x \leftarrow x + \epsilon \nabla_x \log p(x).$$



“Learning to Sample”

- Given $p(x)$ and a neural network $f(\eta,\ \xi)$ with parameter η and random input ξ.
- Find η to match the density of random output $x = f(\eta,\ \xi)$ with $p(x)$.
- Idea: Iteratively adjust η to make the output move along the Stein variational gradient direction.
- Applied for MLE training of deep energy model for GAN-style image generation (arxiv:1611.01722).