Stein Operator $(\mathcal{T}_p f)(x) := \langle \nabla_x \log p(x), f(x) \rangle + \nabla_x \cdot f(x)$

we know $\mathbb{E}_p[\mathcal{T}_p f(x)] = \int \langle \nabla_x p, f \rangle + p \cdot \nabla_x^T f = \int \nabla(pf) = 0, \ \forall f$

$\min_{\{x_i\}} KL[\{x_i\} \| p]$, given p

$q(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{x - x_i\}$ — empirical dist. of x'

$x_i' \longleftarrow x_i + \varepsilon \phi(x_i)$, $\quad \phi = \arg\max_{\phi \in \mathcal{F}} \{ KL[q \| p] - KL[q_{[\varepsilon\phi]} \| p] \}$

$$= \arg\max_{\phi \in \mathcal{F}} \{ \underbrace{-\tfrac{d}{d\varepsilon} KL[q_{[\varepsilon\phi]} \| p] \big|_{\varepsilon=0}} \}$$

$$\overset{(*)}{=} \underline{\mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)]}$$

(*) **Proof:** $T(x) \triangleq x + \varepsilon \cdot \phi(x)$

$KL[q_{[T]} \| p] = KL[q \| p_{[T^{-1}]}] = \int q \cdot (\log q - \log p_{[T^{-1}]}(x))\, dx$

$\log p_{[T^{-1}]}(x) = \log\det(I + \varepsilon \nabla \phi(x)) + \log p(x + \varepsilon\phi(x))$

$\quad = \varepsilon \cdot Tr(\nabla\phi) + \log p(x) + \varepsilon \cdot \nabla \log p(x)^T \phi(x)$

$\Rightarrow \frac{d}{d\varepsilon} KL[q_{[T]} \| p] = \frac{d}{d\varepsilon}[-\int q(x) \log p_{[T^{-1}]}(x)\, dx]$

$\quad = -\int q(x) \cdot [Tr(\nabla\phi) + \nabla \log p(x)^T \phi(x)]\, dx$

Functional derivative in RKHS

$T(x) := x + f(x)$

$KL[q \| p_{[T^{-1}]}] = \int q(x) \cdot [\log q(x) - \log p(x) - (\mathcal{T}_p f)(x)]\, dx$ (when $f \approx 0$)

$(\mathcal{T}_p f)(x) = \langle \mathcal{T}_p k(x,\cdot), f(\cdot) \rangle$

$\Rightarrow \ \delta/\delta f\, KL[q_{[T]} \| p]\big|_{f=0} = -\int q(x)\, \mathcal{T}_p k(x,\cdot)\, dx = \phi^*_{q,p}(\cdot)$

(kernelized) Stein Discrepancy

$\mathbb{D}_{\mathcal{F}}(p \| q) := \max_{\phi\in\mathcal{F}} \mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] \overset{\mathcal{F}=\mathcal{H}}{=\!=} \max_{\phi\in\mathcal{H}} \{ \mathbb{E}_{x\sim q} \mathcal{T}_p \phi(x) \mid \|\phi\|_{\mathcal{H}} \le 1 \}$

$\mathbb{E}_{x\sim q}\mathcal{T}_p\phi(x) = \mathbb{E}_{x\sim q} \langle k(x,\cdot), \mathcal{T}_p \phi(\cdot) \rangle_{\mathcal{H}} = \langle \mathbb{E}_{x\sim q} \mathcal{T}_p k(x,\cdot), \phi(\cdot) \rangle_{\mathcal{H}}$

$(\langle k(x,\cdot), \phi(\cdot) \rangle = \phi(x))$

$\Rightarrow \phi^*(\cdot) \propto \mathbb{E}_{x\sim q}[\mathcal{T}_p k(x,\cdot)]$

$\Rightarrow \mathbb{D}^2(p \| q) = \mathbb{E}_{x,x'\sim q}[\mathcal{T}_p^x \mathcal{T}_p^{x'} k(x,x')]$

Goodness of fit test:

$\Rightarrow \mathbb{D}^2(\{x_i\} \| p) = \frac{1}{n(n-1)} \sum_{i\ne j} \mathcal{T}_p^x \mathcal{T}_p^{x'} k(x_i, x_j)$

whether $\mathbb{D}(\{x_i\} \| p) \ge \cdots$

SVGD

$x_j \longleftarrow x_j + \varepsilon \cdot \hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}[\mathcal{T}_p k(x, x_j)]$

$\quad = [\nabla \log p(x) \cdot k(x, x_j) + \nabla_x k(x, x_j)]$

repulsive force

move to high $p(x)$, general case of MAP

(MAP: $x_j \leftarrow x_j + \varepsilon \nabla \log p(x_j)$)

de-Brujin's Identity

If $\phi_{q,p}(x) := \nabla_x \log \frac{p(x)}{q(x)}$, $T(x) = x + \varepsilon \phi_{q,p}(x)$

$$\frac{d}{d\varepsilon} KL[q_{[T]} \| p]\Big|_{\varepsilon=0} = -\mathbb{E}_{x\sim q}[\mathcal{T}_p \phi(x)] = -\int \frac{q}{p}\nabla(p\phi) = \int \nabla\left(\frac{q}{p}\right)\cdot p\phi$$

$$= \int \frac{\nabla q\cdot p - q\nabla p}{p}\phi = \int q(\nabla\log q - \nabla\log p)\,\phi = -\underbrace{\mathbb{E}_{x\sim q}\left[\left\|\nabla\log\frac{p(x)}{q(x)}\right\|_2^2\right]}_{\text{Fisher divergence}}$$

Fokker-Planck Eq. derivation

(Continuity equation)

$$dx/dt = \phi(x) \Rightarrow \underbrace{\frac{dp(x)}{dt}}_{\neq\, dp_t(x_t)/dt} = -\nabla\cdot(p(x)\phi(x))$$

$$T(x) := x + \varepsilon\phi(x) \Rightarrow T^{-1}(x) = x - \varepsilon\phi(x) + o(\varepsilon)$$

$$\log\tilde{p}(x) = \log p(T^{-1}(x)) + \log\det(\nabla_x T^{-1}(x))$$

$$= \log p(x - \varepsilon\phi(x)) + \log\det(I - \varepsilon\nabla\phi(x)) + o(\varepsilon)$$

$$= \log p(x) \underbrace{- \varepsilon\nabla\log p^T\phi(x) - \varepsilon\, Tr(\nabla\phi)}_{-(\mathcal{T}_p\phi)(x)} + o(\varepsilon)$$

$$\Rightarrow \frac{d}{dt}\log p_t(x) = -(\mathcal{T}_p\phi)(x)$$

$$\Rightarrow \frac{d}{dt}p_t(x) = -p(x)\cdot\mathcal{T}_p\phi(x) = -\nabla\cdot(p(x)\phi(x))$$

Regularized Stein Discrepancy

$$RSD(p\|q; f) = \mathbb{D}(p\|q; f) - \frac{1}{2}\|f\|^2_{L^2(q)}$$

$$= \mathbb{E}_q[\nabla\log p^T f + div(f)] - \frac{1}{2}\mathbb{E}_{x\sim q}[\|f(x)\|^2]$$

$$= \mathbb{E}_q\left[\left(\nabla\log\frac{p}{q}\right)^T f\right] - \frac{1}{2}\mathbb{E}_q[\|f(x)\|^2]$$

$$\Rightarrow f^* = \nabla\log\frac{p}{q}$$

Gradient Flows

$$P_2(\mathcal{M}) = \{\rho : \mathcal{M} \to \mathbb{R}_{\geq 0} \mid \int_{\mathcal{M}} d\rho = 1,\ \int_{\mathcal{M}} |x|^2 \rho(x)\, dx < +\infty\}$$

$$D_{KL}(\rho, \pi) = \int_{\mathcal{M}} (\log \rho(x) - \log \pi(x)) \cdot \rho(x)\, dx$$

$$W_2^2(\mu, \nu) = \inf_{p \in \Pi(\mu,\nu)} \int |x-y|^2\, dp(x,y)$$

SVGD

Gradient Flows

mean-field limits

$$dx/dt = \int [k(x',x) \nabla_{x'} \log \pi(x') + \nabla_{x'} k(x',x)]\, \rho(x')\, dx'$$

$$= \int k(x',x) \nabla_{x'} \log \pi(x')\, \rho(x')\, dx' + \underbrace{\int \nabla_{x'} k(x',x)\, \rho(x')\, dx'}_{}$$

$$- \int k(x',x) \nabla_{x'} \rho(x')\, dx'$$

$$= \int k(x',x) \nabla_{x'} [\log \pi(x') - \log \rho(x')]\, \rho(x')\, dx'$$

$$dx/dt = \mathbb{E}_{x' \sim \rho}[k(x',x) \nabla_{x'} (\log \pi(x') - \log \rho(x'))] = \mathcal{K}_\rho \nabla_x (\log \pi(x) - \log \rho(x))$$

$$(\mathcal{K}_\rho \phi)(x) := \mathbb{E}_{x' \sim \rho}[k(x',x)\phi(x')]$$

Liouville Eq.

$$\partial \rho(x)/\partial t = -\nabla \cdot [\ \ ]$$

$$= \nabla \cdot [\int k(x',x) \nabla_{x'} \frac{\delta}{\delta \rho} D_{KL}(\rho,\pi)\, \rho(x')\, dx'] = \nabla \cdot (\rho(x)\, \mathcal{K}_\rho \nabla_{x'} \frac{\delta}{\delta \rho} D_{KL}(\rho, \pi))$$

似乎是KL为objective, 在Wasserstein metric下的梯度流

? // $\nabla_{W_2} D_{KL}(\rho, \pi)$

Liouville Eq.

$$\partial \rho(x)/\partial t = \nabla \cdot [\rho(x) \nabla \frac{\delta}{\delta \rho} D_{KL}(\rho, \pi)] = \nabla \cdot [\rho(x) \nabla (\log \rho(x) - \log \pi(x))]$$

$$= -\nabla \cdot [\rho(x) \nabla \log \pi(x)] + \nabla^2 \rho(x)$$

mean-field

Wasserstein

dynamics

$$dx/dt = -\nabla \cdot [\log \rho(x) - \log \pi(x)]$$

$$\tilde{x}_i \leftarrow x_i - \varepsilon \cdot [\nabla \log \rho(x_i) - \nabla \log \pi(x_i)]$$

$$\rho(x) \approx \frac{1}{n} \sum_{i=1}^n k(x, x_i) \Rightarrow \nabla \log \rho(x) = \nabla \rho(x) / \rho(x) \approx \sum_i \nabla_x k(x, x_i) \Big/ \sum_i k(x, x_i)$$

$\mathrm{div}(\cdot)$

W_2 grad flow $\quad \partial_t \rho_t = \nabla(\rho_t \nabla_{W_2} F[\rho_t]) \qquad \nabla_{W_2} F[\rho] \overset{?}{:=} \nabla \frac{\delta}{\delta\rho} F[\rho]$

continuity Eq. $\quad \partial_t \rho_t = \mathrm{div}(\rho_t \frac{dx}{dt}) \Rightarrow \frac{dx}{dt} = -\nabla_{W_2} F[\rho_t]$

分步积分符号对不上

not sure

$$\partial_t F[\rho_t] = \partial_t \rho_t \cdot \delta F[\rho_t] = \nabla(\rho_t \frac{dx}{dt}) \cdot \delta F[\rho_t] \overset{?}{=} \mathbb{E}_{\rho_t}[\langle \nabla_{W_2} F[\rho_t], \frac{dx}{dt} \rangle]$$
$$= -\mathbb{E}_{\rho_t}[\| \nabla_{W_2} F[\rho_t] \|^2]$$

f-divergence $\quad D_f[\rho \| \pi] = \mathbb{E}_\pi[f(\rho/\pi)]$

$$D_f[\rho + \varepsilon v \| \pi] - D_f[\rho \| \pi] = \int \pi \cdot [f(\frac{\rho}{\pi} + \varepsilon \frac{v}{\pi}) - f(\frac{\rho}{\pi})] \approx \int \pi \cdot f'(\frac{\rho}{\pi}) \cdot \frac{v}{\pi} \cdot \varepsilon$$

$$\Rightarrow \delta/\delta\rho \, D_f[\rho \| \pi] = f'(\rho/\pi)$$

$$\Rightarrow \nabla_{W_2} D_f[\rho \| \pi] = \nabla \, \delta/\delta\rho \, D_f[\rho \| \pi] = \nabla f'(\rho/\pi)$$

KL: $f(t) = t\log t, \quad f'(t) = \log t + 1 \qquad \chi^2: f(t) = t^2 - 1, \quad f'(t) = 2t,$

结论: $\quad \nabla_{W_2} KL[\rho \| \pi] = \nabla_x \log \frac{\rho}{\pi} \qquad \nabla_{W_2} \chi^2[\rho \| \pi] = 2\nabla(\frac{\rho}{\pi})$

We know SVGD is kernelized KL grad flow

kernelized χ^2 grad flow

$$-\frac{dx}{dt} = \mathcal{K}_\pi \nabla_{W_2} \chi^2[\rho \| \pi] = \mathbb{E}_{x' \sim \pi}[k(x,x') \, 2\nabla(\frac{\rho(x')}{\pi(x')})]$$
$$= 2\int k(x,x')[\nabla\rho(x') - \nabla\log\pi(x')\rho(x')]\,dx'$$
$$= 2\int k(x,x') \nabla\log\frac{\rho(x')}{\pi(x')} \rho(x')\,dx' = 2\mathcal{K}_\rho \nabla_{W_2} KL[\rho \| \pi]$$

本质上是 $\int \pi \nabla(\frac{\rho}{\pi}) = \int \rho \nabla(\log \frac{\rho}{\pi})$