On the Geometry of SVGD

Won E Hong

Carnegie Mellon University

November 17, 2020

Outline

Minimizing Movement

$$\left(x_{k+1} \in \arg\min_x F(x) + \frac{|x - x_k|^2}{2dt} \right) \Rightarrow \left(\frac{x_{k+1} - x_k}{dt} = -\nabla F(x_{k+1}) \right)$$

The minimizing movement approximates the gradient flow.

$$x' = -\nabla F(x)$$

Gradient flow provide the steepest direction in the space.

Direction=Angle=Inner product

Theorem (AC curve in $\mathcal{W}_p$)

Let $(\rho_t)_{t\in[0,1]}$ be an absolutely continuous curve in $\mathcal{W}_p(\Omega)$. Then for a.e. $t \in [0,1]$, there exists a vector field $v_t \in L^p_{\rho_t}(\mathbb{R}^d)$ such that

$$\partial_t \rho + \nabla \cdot (\rho v) = 0,$$

$$\|v_t\|_{L^p(\rho_t)} = |\partial_t \rho|(t).$$

For $p = 2$ we can bring Hilbert structure on the tangent manifold.

Riemannian structure

The tangent pair (s, u) is defined $s = -\nabla \cdot (\rho u)$

$$g_\rho(s_1, s_2) := \langle u_1, u_2 \rangle_\rho := \int u_1 u_2 d\rho \ .$$

$$\boxed{\rho_{k+1} \in \arg\min_{\rho} F(\rho) + \frac{W_2^2(\rho, \rho_k)}{2dt}}$$

$$\frac{1}{dt}\left(F(\rho) - F(\rho_k) + \frac{W_2^2(\rho, \rho_k)}{2dt}\right) = \frac{F(\rho) - F(\rho_k)}{dt} + \frac{W_2^2(\rho, \rho_k)}{2dt^2}$$

$$\approx g(\mathrm{grad} F(\rho_k), \partial_t \rho) + \frac{1}{2}|\partial_t \rho|^2$$

$$= \langle \nabla \frac{\delta F}{\delta \rho}(\rho), v \rangle_\rho + \frac{1}{2}\|v\|_\rho^2$$

Gradient of F

$$\boxed{\mathrm{grad} F(\rho_k) = -\nabla \cdot (\rho \nabla \frac{\delta F}{\delta \rho}(\rho))} \text{ and } \boxed{\partial_t \rho = -\nabla \cdot (\rho v)}$$

.

Thus we aim to get vector field from **the minimization problem**

$$\boxed{\min_{v} \langle \nabla \frac{\delta F}{\delta \rho}(\rho), v \rangle_{\rho} + \frac{1}{2} \|v\|_{\rho}^{2}}$$

So variation on v let us to conclude

$$\boxed{v_{\min} = -\nabla \frac{\delta F}{\delta \rho}(\rho)}$$

$(\partial_t \rho, v_{\min})$ and $(\mathrm{grad} F(\rho), \nabla \frac{\delta F}{\delta \rho}(\rho))$ implies that

$$\boxed{\partial_t \rho = -\mathrm{grad} F(\rho)}$$

KL divergence

$$KL(\rho||\nu) := \int \frac{d\rho}{d\nu} \log(\frac{d\rho}{d\nu}) d\nu$$

We initialize the particles with some simple distribution ρ, and update them via the map $T(x) = x + \epsilon\phi(x)$, where ϵ is a small step size

Functional optimization

$$\arg\max_{\phi \in \mathcal{H}} \left\{ -\frac{d}{d\epsilon} KL(T\rho||\nu)|_{\epsilon=0} \quad \text{s.t.} \quad \|\phi\|_{\mathcal{H}} \leq 1 \right\}$$

$-\frac{d}{d\epsilon} KL(T\rho||\nu)|_{\epsilon=0} = \mathbb{E}_\rho[\nabla \log \nu(x)^\top \phi(x) + \nabla \cdot \phi(x)] =: \mathbb{E}_\rho[\mathcal{S}_\nu \phi]$ where $\mathcal{S}_\nu$ is called the Stein operator. Note that $\mathbb{E}_\nu[\mathcal{S}_\nu \phi] = 0$

The Stein discepancy

$$\mathbb{D}(\rho||\nu) := \max_{\phi \in \mathcal{H}} \{\mathbb{E}_\rho[\mathcal{S}_\nu \phi] \quad \text{s.t.} \quad \|\phi\|_{\mathcal{H}} \leq 1\}$$

If we let $\mathcal{H} = \mathcal{H}_0^d$ where $\mathcal{H}_0$ is a RKHS for a p.d. $k(x, x')$ then

$$\boxed{\phi^*_{\rho,\nu}(\cdot) \propto \mathbb{E}_{x\sim\rho}[\mathcal{S}_\nu \otimes k(x, \cdot)],}$$

where $\mathcal{S}_\nu \otimes k(x, \cdot) := \nabla \log \nu(x) k(x, \cdot) + \nabla k(x, \cdot)$.
Let $\nu = \frac{d\nu}{dx} = e^{-V}$ yields

$$\boxed{X^i_{n+1} = X^i_n - \frac{\epsilon}{N}\left(\sum_{j=1}^N \nabla k(X^i_n, X^j_n) + \sum_{j=1}^N k(X^i_n, X^j_n)\nabla V(X^j_n)\right)}$$

$$\boxed{\partial_t \rho_t(x) = \nabla \cdot \left(\rho_t(x) \int_{\mathbb{R}^d} k(x, y)[\nabla \rho_t(y) + \rho_t(y)\nabla V(y)]dy\right)}$$

Definition

The kernel k is integrally strictly positive definite if for all finite nonzero signed Borel measure ρ, $\int \int k(x,y)d\rho(x)d\rho(y) > 0$.

Let $(\mathcal{H}_k, \langle \cdot, \cdot \rangle)$ be the RKHS associated with the kernel k so that $f(x) = \langle f, k(x,\cdot)\rangle_{\mathcal{H}_k}$,
Define the linear operator $T_{k,\rho} : L^2(\rho) \to \mathcal{H}_k$ via

$$T_{k,\rho}\phi = \int k(\cdot, y)\phi(y)d\rho(y), \quad \phi \in L^2(\rho).$$

For $\rho \in \mathcal{P}_k(\mathbb{R}^d)$, $T_{k,\rho}$ is compact, self-adjoint and p.s.d..

Definition

$\mathcal{P}_k(\mathbb{R}^d) = \{\rho \in \mathcal{P}(\mathbb{R}^d) : \rho$ admits a smooth Lebesgue density,
$\text{supp}\rho = \mathbb{R}^d, \quad \int_{\mathbb{R}^d} k(x,x)d\rho(x) < \infty\}$

Definition (Tangent plane)

For $\rho \in \mathcal{P}_k(\mathbb{R}^d)$, we define the tangent space

$$T_\rho M = \Big\{ \xi \in \mathcal{D}'(\mathbb{R}^d) : \text{there exists } v \in \overline{T_{k,\rho} \nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_d^k} \text{ such that } \xi + \nabla \cdot (\rho v) = 0 \text{ in the sense of distributions} \Big\}$$

and the Riemannian metric $g_\rho : T_\rho M \times T_\rho M \to \mathbb{R}$ by

$$g_\rho(\xi, \chi) = \langle u, v \rangle_{\mathcal{H}_k^d},$$

where $\xi + \nabla \cdot (\rho u) = 0$ and $\chi + \nabla \cdot (\rho v) = 0$

Lemma

1. $(T_\rho M, g_\rho)$ *is a Hilbert space.*
2. *For all* $\xi \in T_\rho M$ *there exists a unique* $v \in \overline{T_{k,\rho}\nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_d^k}$ *such that* $\xi + \nabla \cdot (\rho v) = 0$ *in the sense of distributions. Also* $(\overline{T_{k,\rho}\nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_d^k}, \langle \cdot, \cdot \rangle_{\mathcal{H}_k^d}) \approx (T_\rho M, g_\rho)$.

Definition (L^2 functional derivative)

$$\frac{d}{d\epsilon}|_{\epsilon=0}\mathcal{F}(\rho + \epsilon\phi) =: \int \frac{\delta \mathcal{F}}{\delta \rho}(\rho)\phi dx$$

for $\phi \in C_c^\infty(\mathbb{R}^d)$ with $\int \phi dx = 0$.

Lemma (Stein gradient)

Assume that $T_{k,\rho}\nabla\frac{\delta\mathcal{F}}{\delta\rho}(\rho) \in \overline{T_{k,\rho}\nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_d^k}$. *Then the Riemannian gradient associated to* $T_\rho M, g_\rho)$ *is given by*

$$(grad_k\mathcal{F})(\rho) = -\nabla\cdot\left(\rho T_{k,\rho}\nabla\frac{\delta\mathcal{F}}{\delta\rho}(\rho)\right).$$

Proof

$$g_\rho\left(\text{grad}_\rho\mathcal{F}, \partial_t\mu_t|_{t=0}\right) = \frac{d}{dt}\mathcal{F}(\mu_t)|_{t=0}$$

for all sufficiently regular curves $(\mu_t)_{t\in(-\epsilon,\epsilon)} \subset M$ with $\mu_0 = \rho$ and $\partial_t\mu_t|_{t=0} \in T_\rho M$. Then we can find the corresponding vector fields $(w_t)_{t\in(-\epsilon,\epsilon)}$ satisfying $\partial_t\mu + \nabla\cdot(\mu w) = 0$.

$$
\begin{aligned}
(\text{RHS}) &= \int \frac{\delta \mathcal{F}}{\delta \rho}(\mu_t)\partial_t \mu_t dx\Big|_{t=0} = \int \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\mu_t)(\rho), w_0 d\rho \\
&= \int \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\mu_t)(\rho), \langle k_x, w_0 \rangle_{\mathcal{H}_k^d} d\rho \\
&= \langle T_{k,\rho} \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho), w_0 \rangle_{\mathcal{H}_k^d}
\end{aligned}
$$

Since $w_0 \in \mathcal{H}_k^d$ arbitrary. Thus

$$
g_\rho \left(\text{grad}_\rho \mathcal{F}, \partial_t \mu_t|_{t=0} \right) = \langle T_{k,\rho} \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho), w_0 \rangle_{\mathcal{H}_k^d}
$$

in other words, $(\text{grad}_\rho \mathcal{F}, T_{k,\rho} \nabla \frac{\delta \mathcal{F}}{\delta \rho}(\rho)) = (\xi, u)$ such that $\xi + \nabla \cdot (\rho u) = 0$. $\square$

Corollary

The gradient flow dynamics of the KL-divergence with respect to the Stein geometry is given by the Stein PDE.

$$\partial_t \rho = -grad_\rho \mathcal{F}$$

Furthermore, for solutions $(\rho_t)_{t\geq 0}$ *to the Stein PDE it holds that*

$$\frac{d}{dt} KL(\rho_t || \pi) \leq 0.$$

Definition (Stein distance)

For $\mu, \nu \in M$ we define the Stein distance

$$d_k^2(\mu,\nu) = \inf_{(\rho,\nu)\in\mathcal{A}(\mu,\nu)} \left\{ \int_0^1 \|v_t\|_{\mathcal{H}_k^d}^2 dt, \quad v_t \in \overline{T_{k,\rho}\nabla C_c^\infty(\mathbb{R}^d)}^{\mathcal{H}_d^k} \right\}$$
$$= \inf_{\rho} \left\{ \int_0^1 g_{\rho_t}(\partial_t\rho_t, \partial_t\rho_t)dt : \quad \rho_0 = \mu, \rho_1 = \nu \right\}.$$

Lemma

The following hold:

1. *The Stein distance d_k is an extended metric on M.*
2. *If k is continuous and bounded, then there exists a constant $C > 0$ such that, $\mathcal{W}_2(\mu,\nu) \leq C d_k(\mu,\nu)$, for $\mu, \nu \in M$.*

Proposition (Geodesic equations)

Let $(\rho_t, v_t)_{0\le t\le 1}$ be a critical point the Stein distance. Then

$$\partial_t \rho + \nabla \cdot (\rho T_{k,\rho} \nabla \Psi) = 0$$
$$\partial_t \Psi + \nabla \Psi \cdot T_{k,\rho} \nabla \Psi = 0,$$

for some function $\Psi : [0,1] \times \mathbb{R}^d \to \mathbb{R}$, and $v_t = T_{k,\rho} \nabla \Psi$

comparison with the decoupled equation $\mathcal{W}^2$

$$\partial_t \rho + \nabla \cdot (\rho \nabla \Psi) = 0$$
$$\partial_t \Psi + \frac{1}{2} |\nabla \Psi|^2 = 0,$$

The stein geometry explains the mean-field limit of an interacting particle system.

Proof Let's substitute $v_t = T_{k,\rho_t}\nabla\Phi_t$ with $\Phi_t \in C_c^\infty(\mathbb{R}^d)$, $t \in [0,1]$, to obtain

$$d_k^2(\mu,\nu) = \inf_{(\rho,\Phi)} \left\{ \int_0^1 \| T_{k,\rho}\nabla\Phi\|_{\mathcal{H}_k^d}^2 dt : \partial_t\rho + \nabla\cdot(\rho T_{k,\rho}\nabla\Psi) = 0, \right\}$$

in the sense of distribution. So (ρ,Φ) satisfies

$$\underbrace{-\int_0^1 \int \partial_t\Psi d\rho - \langle\nabla\Psi, T_{k,\rho_t}\nabla\Phi_t,\rangle_\rho dt + \int \Psi_1 d\nu - \int \Psi_0 d\mu}_{\textbf{(c)}} = 0$$

for all test functions Ψ.

Constraint functional

$$\mathcal{E}(\rho,\Phi) := \sup_\Psi \{\textbf{(c)} = \begin{cases} 0 & \text{if } (\rho,\Phi) \text{ solves the condition,} \\ \infty & \text{otherwise.} \end{cases} \}$$

by the linearity in Ψ.

Therfore

$$\frac{1}{2}d_k^2(\mu,\nu) = \inf_{(\rho,\Phi)} \sup_{\Psi} \left\{ \frac{1}{2}\int_0^1 \| T_{k,\rho_t}\nabla\Phi_t\|^2_{\mathcal{H}_k^d} dt + \mathcal{E}(\rho,\Phi) \right\}$$

Note the convexity in Φ and concativity (in fact, linearity) in Ψ. Thus by exchanging inf and sup to obtain

$$\frac{1}{2}d_k^2(\mu,\nu) = \inf_{(\rho,\Phi)} \sup_{\Psi} \left\{ -\int_0^1\int_\Omega \partial_t\Psi d\rho dt + \int_\Omega \Psi_1 d\nu - \int_\Omega \Psi_0 d\mu \right.$$
$$\left. + \inf_{\Phi} \left\{ \frac{1}{2}\int_0^1 \| T_{k,\rho_t}\nabla\Phi_t\|^2_{\mathcal{H}_k^d} - \langle \nabla\Psi, T_{k,\rho_t}\nabla\Phi_t\rangle_{L^2(\rho)} dt \right\} \right\}$$

$$\inf_{\Phi} \left\{ \frac{1}{2} \int_0^1 \| T_{k,\rho_t} \nabla \Phi_t \|_{\mathcal{H}_k^d}^2 - \langle T_{k,\rho_t} \nabla \Psi_t, T_{k,\rho_t} \nabla \Phi_t, \rangle_{\mathcal{H}_k^d} dt \right\}$$
$$= -\frac{1}{2} \int_0^1 \| T_{k,\rho_t} \nabla \Psi_t \|_{\mathcal{H}_k^d}^2 dt$$

So let $\Phi = \Psi$, formally

$$\frac{\delta}{\delta \rho} \left(\frac{1}{2} \| T_{k,\rho} \nabla \Psi_t \|_{\mathcal{H}_k^d}^2 \right)(x) = \nabla \Psi(x) \cdot (T_{k,\rho} \nabla \Psi)(x),$$

$$\frac{\delta}{\delta \Psi} \left(\frac{1}{2} \| T_{k,\rho} \nabla \Psi_t \|_{\mathcal{H}_k^d}^2 \right)(x) = \nabla \cdot (\rho T_{k,\rho} \nabla \Psi)(x).$$

□

Lemma (Computing the Hessian)

Let $(\rho_t, \Psi_t)_{t\in(-\epsilon,\epsilon)}$ *be a Stein geodesic and* $\rho_0 = \rho$, $\Psi_0 = \Psi$. *Then*

$$\frac{d^2}{dt^2} KL(\rho_t||\nu)|_{t=0} = Hess_\rho(\Psi, \Psi),$$

where

$$Hess_\rho(\Phi, \Psi) = \sum_{i,j=1}^{d} \int \int \partial_i \Phi(y) q_{ij}[\rho](y, z) \partial_j \Psi(z) d\rho(y) d\rho(z),$$

Lemma

where

$$q_{ij}[\rho](y,z) = \delta_{ij}\sum_{l=1}^{d}\int \partial_{x_l}\left(e^{-V(x)}k(x,y)\right)e^{V(x)}d\rho(x)\partial_{y_l}k(y,z)$$
$$-\int \partial_{y_j}\partial_{x_i}\left(e^{-V(x)}k(x,y)\right)e^{V(x)}d\rho(x)k(y,z)$$
$$-\int \partial_{x_j}\left(e^{V(x)}\partial_{x_i}(e^{-V(x)}k(x,y))\right)k(x,z)d\rho(x).$$

Theorem (Informal)

Assume that there exists $\lambda > 0$ *such that*

$$\frac{d^2}{dt^2} KL(\rho_t||\pi) > \lambda \int \int \nabla\Psi(y) \cdot k(y,z) \nabla\Psi(z) d\rho(y) d\rho(z),$$
$$Hess_\rho(\Psi,\Psi) > \lambda g_\rho(v,v),$$

for all $(\rho)_{t\in(-\epsilon,\epsilon)}$ (ρ,Ψ) *satisfying the geodesic equation. Then*

$$KL(\rho_t||\pi) \leq e^{-2\lambda t} KL(\rho_0||\pi)$$

along solutions ρ_t *of the Stein equation.*

Remark

or

$$\partial_t KL(\rho_t||\pi) = \int\int \nabla(\frac{\rho_t}{\pi})\cdot k(y,z)\nabla(\frac{\rho_t}{\pi})d\pi(y)d\pi(z) =: I_{Stein}(\rho_t|\pi),$$

I_{Stein} is called 'Stein-Fisher information'.
Assuming a 'Stein-log-Sobolev inequality'

$$KL(\rho_t||\pi) \leq \frac{1}{2\lambda} I_{Stein}(\rho_t|\pi)$$

the exponential decay estimate also would follow.

Remark

- *The exponential convergence to equilibrium does not hold if k is too regular i.e. $k \in C^{1,1}(\mathbb{R}^d \times \mathbb{R}^d)$ then the inequality only hold for $\lambda = 0$*
- *'Measuring Sample Quality with Kernels' (Gorham, Lester Mackey 20') deduced that if $k \in C^{1,1}$ the Kernelized Stein Discrepancy does not imply weak convergence of measure.*

Thank you!