Stein Variational descent

Two objectives

- deterministic, particle-based approximation for Fokker-Planck and related eq.
- methods that are robust wrt dimension.

(connection to MMD - max. mean discrepancy)

μ in $\mathbb{R}^d$ $\quad d$- large

μ_N

$$d_W(\mu_N, \mu) \gtrsim \left(\frac{1}{N}\right)^{\frac{1}{d}}$$

$$d_{MMD}(\mu_N, \mu) \sim \frac{1}{\sqrt{N}}$$

F-P

$$\partial_t \rho = \Delta \rho - \operatorname{div}(\rho \nabla U)$$

$$d(\mu, \nu) = \sup_{f \in \mathrm{Lip}(1)} \int f \, d\mu - d\nu$$

$$\sup_{f \in \mathcal{H}_K} \int f \, d\mu - d\nu$$

Reproducing Kernel Hilbert Spaces (RKHS)

Let $\mathcal{H}$ be a Hilbert space of functions from $\Omega \subseteq \mathbb{R}^d$ to $\mathbb{R}$.

Def H is an RKHS if $\forall x \in \Omega$ the evaluation at x is a continuous function on H, that is $\delta_x : H \to \mathbb{R}$ defined by

$$f \xmapsto{\delta_x} f(x)$$

is a continuous functional.

Consequence: By Riesz Representation Theorem there exists $\phi_x \in H$ such that

$$\forall f \in H \qquad \boxed{f(x) = \delta_x(f) = \langle f, \phi_x \rangle_H}$$

The mapping from Ω to H, $x \to \phi_x$ is called the feature map.

Let $K(x,y) = \langle \phi_x, \phi_y \rangle = \phi_x(y) = \phi_y(x)$

K is called the kernel.

- K is positive definite. For μ- signed measure $|\mu| < \infty$.

$$\iint K(x,y)\, d\mu(x)\, d\mu(y) = \iint \langle \phi_x, \phi_y \rangle\, d\mu(x)\, d\mu(y) =$$

$$(\star) \quad = \left\langle \int \phi_x\, d\mu(x), \int \phi_y\, d\mu(y) \right\rangle \geq 0$$

$$= \int \langle E_\mu(\phi_x), \phi_y \rangle\, d\mu(y) = \langle E_{\mu_x}(\phi_x), E_{\mu_y}(\phi_y) \rangle \geq 0$$

A bit more precisely. For G probability measure $(\mu = aG_1 - bG_2)$ consider

$$f \mapsto \int f(x)\, dG(x)$$

we note $\left|\int f(x)\, dG(x)\right| = \left|\int \langle \phi_x, f \rangle\, dG(x)\right|$

$$\leq \int \|\phi_x\|_H \, \|f\|_H \, dG(x)$$

$$\leq \int \|\phi_x\|_H \, dG(x) \cdot \|f\|_H$$

$$\leq \underbrace{\sqrt{\int K(x,x)\, dG(x)}} \cdot \|f\|_H$$

consider K with $K(x,x)$ bounded.

So $f \mapsto \int f\, dG$ is a linear functional.

Therefore $\exists\ E_G \in H$

$$\langle E_G, f \rangle = \int f\, dG = \left\langle \int \phi_x \, dG(x), f \right\rangle$$

$G \mapsto E_G$ is the mean embedding

From $(\star)$ follows that if $g(x) = \int K(x,y)\, d\mu(y) = E_\mu(\phi_x)$ then

$$\|g\|_H^2 = \iint K(x,y)\, d\mu(x)\, d\mu(y)$$

<u>Theorem</u> (Moore-Aronszajn)

If $K : \Omega \times \Omega \to \mathbb{R}$ is symmetric, positive definite then $\mathcal{H}_0 = \operatorname{span}\{K(\cdot, y) : y \in \Omega\}$ is an inner product space with

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_i \sum_j \alpha_i \beta_j \, k(x_i, y_j)$$

where $f = \sum_i \alpha_i \, k(x_i, \cdot)$, $g = \sum_j \beta_j \, k(\cdot, y_j)$.

Let $\mathcal{H}$ be a completion of $\mathcal{H}_0$. Then $\mathcal{H}$ is an RKHS.

$\mathcal{H}$ is the set of pointwise limits of f_n in $\mathcal{H}_0$.

[Notes by Sejdinović and Gretton].

[Bob mentioned: Paulsen & Raghupathi "An Intro to the Theory of RKHS" available online via CMU Library.]

Examples of RKHS

① $K(x,y) = \eta(x-y)$ $\quad \eta$ - Gaussian

$$\mathcal{H} = \{ \eta * f : f \text{ - measure}, |f| < \infty \}$$

$$\langle \eta * f, \eta * g \rangle = \iint \eta(x-y) f(x) g(y) \, dx \, dy$$

② M - compact manifold of dimension d.

$\dot{H}^s$ - fractional Sobolev space.

If $s > \frac{d}{2}$ $\quad H^s \hookrightarrow C(M)$ and H^s is RKHS

Spectral representation: (λ_i, ψ_i) Laplacian eigenval. and eigenfunctions, orthonormal in $L^2(M)$.

$f \in H^s$ $\qquad f = \sum_{i=1}^{\infty} a_i \psi_i$

$$\|f\|_{H^s}^2 = \sum_{i=1}^{\infty} (1+\lambda_i)^s a_i^2 \qquad (1-\Delta)^s$$

Let us compute the kernel

$$\langle \phi_{\bar{x}}, f \rangle = f(\bar{x})$$

$$\phi_{\bar{x}} = \sum b_i \psi_i \qquad f = \sum a_j \psi_j$$

$$\langle \phi_{\bar{x}}, f \rangle = \sum_i (1+\lambda_i)^s a_i b_i = \sum_i a_i \psi_i(\bar{x})$$

thus $\qquad b_i = \frac{1}{(1+\lambda_i)^s} \psi_i(\bar{x})$

So

$$\Phi_{\bar{x}}(x) = \sum_i \frac{1}{(1+\lambda_i)^s} \psi_i(\bar{x}) \psi_i(x) .$$

[~ Mercer theorem]

Stein Variational Descent for Relative Entropy

[Liu, Wang NeurIPS '16, Liu NeurIPS '17, Lu, Lu, Nolen '18]

$$E(\rho) = \int_{\mathbb{R}^d} \rho \ln\left(\frac{\rho}{\sigma}\right) dx = \int \rho \log \rho + \rho\, U\, dx$$

if $\sigma \sim e^{-U}$

σ is a.c. wrt Lebesgue measure.

- Wasserstein gradient flow

$\frac{dE(\gamma)}{dt}\Big|_{t=0}$ $\gamma'(0) = V$

- on a manifold - grad E is a minimizer of

$$R(V) = \frac{1}{2} g(V,V) + \operatorname{diff} E[V] = dE(V)$$

For Wass: Tangent vector $\partial_t \rho = -\operatorname{div}(\rho V)$

$$R(V) = \frac{1}{2} \int |V|^2 \rho\, dx + \int V \cdot (\nabla \rho - \rho \nabla U)\, dx$$

$$\frac{dE}{dt} = \int \rho_t \ln \rho + \rho_t + \rho_t \ln \sigma\, dx$$

$$= \int \cancel{\rho} V \frac{\nabla \rho}{\cancel{\rho}} + \rho V \frac{\nabla \sigma}{\sigma} dx = \int V \cdot (\nabla \rho - \rho \nabla U)\, dx$$

Minimizing $R(V)$ gives $\rho V_{Wass} = -\nabla \rho + \rho \nabla U$

so $\partial_t \rho = -\operatorname{div}(\rho V) = \Delta \rho - \operatorname{div}(\rho \nabla U)$ (FP)

- Stein variational descent

Consider a different metric for velocity

$$R(v) = \frac{1}{2}\|v\|_H^2 + \int v\cdot(\nabla\rho - \rho\nabla U)\,dx$$

minimizing gives $\forall w$

$$\langle v, w\rangle_H = -\int (\nabla\rho - \rho\nabla U)\cdot w\,dx$$

By above intro we have that v is the mean embedding for measure $(\rho\nabla U - \nabla\rho)$.

So $$v(x) = \int k(x,y)(\rho\nabla U - \nabla\rho)\,dy$$

If $k(x,y) = k(x-y)$ then

$$v_s = k * (\rho\nabla U - \nabla\rho) = k*(\rho\nabla U) - \nabla(k*\rho)$$

Gradient flow

$$\text{(SVD)} \quad \partial_t\rho = \mathrm{div}(k*(\nabla\rho - \rho\nabla U))$$

So by penalizing the velocity in higher norm we get a smoother velocity.

<u>Particle methods</u>

Consider $\rho_N = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}$

v_{wass}, and relative entropy do not make sense.

But V_{stein} does

$$SVD_N \begin{cases} V_i = \frac{1}{N}\sum_j \left(k(x_i - x_j)\,\nabla U(x_j) - \nabla k(x_i - x_j) \right) \\ \frac{dx_i}{dt} = V_i \end{cases}$$

Lu, Lu, Nolen study well posedness and asymptotics for SVD. (if $K = J * J$)

Solutions of SVD_N are weak sol of SVD

In T 2.7 they show stability for SVD using a coupling technique. This implies discrete to continuum convergence.

Alternative functional (Carrillo, Craig, Patacchini)

$$E_k(\rho) = \int \ln\left(\frac{k*\rho}{\sigma}\right) d\rho = \int \ln(k*\rho)\, d\rho + \int U\, d\rho$$

$$\partial_t \rho = -\operatorname{div}(\rho v)$$

$$\operatorname{diff} E_k[v] = \frac{dE}{dt}$$

$$R(v) = \frac{1}{2}\int |v|^2\, d\rho + \operatorname{diff} E_k[v]$$

minimizer

$$v = -\nabla U - \nabla k * \left(\frac{\rho}{k*\rho}\right) - \frac{1}{k*\rho}\nabla k*\rho$$

[CCP] show that as $N \to \infty$ particle approximations converge, using Sandier-Serfaty framework.