Bayesian inference: variational forms + gradient flows

Standard Bayes:

prior - $\pi(\theta)$

likelihood - $L(X|\theta)$

posterior - $\pi_n(\theta|X) = \frac{L(X|\theta)\pi(\theta)}{Z}$

$Z = \int_{\theta \in \Theta} L(X|\theta)\pi(\theta)\, d\theta$

A classic variation formulation

$q(\theta) \in \mathcal{P}(\mathcal{M})$ - measures on a manifold

$\pi(\theta) \in \mathcal{P}(\mathcal{M})$

$\pi_n(\theta|X) \in \mathcal{P}(\mathcal{M})$

$\ell(X|\theta) = -\text{Log}[L(X|\theta)]$

$$\pi_n(\theta|x) = \arg\min_q \left[D_{KL}(q(\theta)||\pi(\theta)) + \int l(x|\theta) q(\theta) d\theta \right]$$

$$D_{KL}(q||\pi) = \sum_\theta Q(\theta) \log \frac{Q(\theta)}{\pi(\theta)}$$

$$= \int_\theta q(\theta) \log \frac{q(\theta)}{\pi(\theta)} d\theta$$

Maxent (Jaynes)

$$\pi_n(\theta|x) = \arg\min_q \left[-\lambda \, Ent(q) + \int l(x|\theta) q(\theta) d\theta \right]$$

What we will talk about today:

$$\arg\min_q \left[\int l(x|\theta) q(\theta) d\theta + D(q||\pi) \right]$$

here D is a general variational functional

Questions we will ask:

1) What do different variational functionals imply in terms of the solution

2) What do gradient flows look like for these functionals, what information do they provide?

3) An application Riemannian Markov chain Monte Carlo

Notation and defs.

(M, g) - m-dim Riemannian manifold with metric g on Ω

Vol_g - volume form

$\nabla_g, div_g, Hess_g, \Delta_g$ - gradient, divergence, Hessian, Laplace-Beltrami

$\mathcal{P}(\mathcal{M})$ - space of measures on $\mathcal{M}$

$$\pi(\theta) = e^{-\varphi} \mathrm{Vol}_g$$

$$\pi_n(\theta|x) \propto e^{-\Phi} \pi, \quad \Phi(\cdot) = \Phi(\cdot; x) = -L(x|\theta)$$

$$\mathcal{P}_2(\mathcal{M}) := \left\{ \nu \in \mathcal{P}(\mathcal{M}) : \int_{\mathcal{M}} d^2(x, x_0)\, d\nu(x) < \infty, \text{ for some } x_0 \in \mathcal{M} \right\}$$

Functionals:

a) KL: $\nu_1 << \nu_2$

$$D_{KL}(\nu_1 \| \nu_2) = \begin{cases} \int_{\mathcal{M}} \frac{d\nu_1}{d\nu_2} \log\left(\frac{d\nu_1}{d\nu_2}(\theta)\right) d\nu_2(\theta) \\ \infty \quad \text{o.w.} \end{cases}$$

b) χ^2: $\nu_1 << \nu_2$

$$D_{\chi^2}(\nu_1 \| \nu_2) = \begin{cases} \int_{\mathcal{M}} \left[\frac{d\nu_1}{d\nu_2}(\theta) - 1\right]^2 d\nu_2(\theta) \\ \infty \quad \text{o.w.} \end{cases}$$

c) Dirichlet energy:

$$\mathcal{D}^{\mu}(f) = \begin{cases} \int_{\mathcal{M}} \|\nabla_g f(\theta)\|^2 \, d\mu(\theta), \\ +\infty \qquad \text{o.w.} \end{cases}$$

$$f \in L^2(\mathcal{M}, \mu) \cap H^1(\mathcal{M})$$

d) W_2^2:

$$W_2^2(\nu_1, \nu_2) = \inf_{\alpha} \int_{\mathcal{M} \times \mathcal{M}} d(x,y)^2 \, d\alpha(x,y)$$

$$\nu_1, \nu_2 \in \mathcal{P}_2(\mathcal{M})$$

α = transportation plan,
$\alpha \in \mathcal{P}(\mathcal{M} \times \mathcal{M})$

Geodesic space: $f \in L^2(\mathcal{M}, \mu)$ with a constant speed geodesic connecting f_0 & f_1 is $t \in [0,1] \mapsto (1-t) f_0 + t f_1$

Convexity properties:

1) A functional $E : \mathcal{X} \to \mathbb{R} \cup \{\infty\}$ is λ-geodesically convex if $\forall x_1, x_2 \in \mathcal{X}$ $\exists$ a constant speed geodesic $t \in [0,1] \to \gamma(t) \in \mathcal{X}$, $\gamma(0) = x_1$, $\gamma(1) = x_2$

$$E(\gamma(t)) \le (1-t)\,E(x_0) + t\,E(x_1) - \lambda \frac{t(1-t)}{2} d_x^2(x_0, x_1),$$

$$\forall t \in [0,1],$$

2) Let $\psi \in C^2(M)$. The following are equivalent

(i) ψ is λ-geodesically convex

(ii) $\mathrm{Hess}_g \psi_x(v,v) \ge \lambda \quad \forall x \in M, \; v \in T_x M$

If (M, d) is Euclidean (i) $\Leftrightarrow$ (ii) $\Leftrightarrow$ (iii)

(iii) $\psi - \frac{\lambda}{2} |\cdot|^2$ is convex.

Strong convexity in optimization

Gradient flows:

Metric space (X, d_X)

1) $X = \mathbb{R}^d$, $d_X(u,v) = \|u - v\|$

$E : \mathbb{R}^d \to \mathbb{R}$ is differentiable

(gf) $$\dot{x}(t) = -\nabla E(x(t)), \quad t \geq 0$$
$$x(0) = x_0$$

The solution of (gf) is the gradient flow.

More general in integral form

$$E(x_0) = E(x(t)) + \frac{1}{2}\int_0^t |\dot{x}(r)|^2 dr + \frac{1}{2}\int_0^t |\nabla E(x(r))|^2 dr, \quad t \geq 0$$

Energy dissipation equality.

$$|\dot{x}(t)| := \lim_{s \to t} \frac{d_X(x(t), x(s))}{|s-t|}$$

$$|\nabla E(x)| := \limsup_{y \to x} \frac{(E(x) - E(y))_+}{d_X(x,y)}$$

Variational forms for Bayesian updating:
from stat phys/ prob.

$u(d\theta) = \frac{1}{Z} \exp(-\phi(\theta)) \pi(d\theta)$ is
the minimizer of

$$J_{KL}(\nu) := D_{KL}(u \| \pi) + F_{KL}(\nu; \phi)$$

(1)
$$F_{KL}(\nu; \phi) = \int_M \phi(\theta)\, d\nu(\theta)$$

$$J_{\chi^2}(\nu) := D_{\chi^2}(\nu \| \pi) + F_{\chi^2}(\nu; \phi)$$

(2)
$$F_{\chi^2}(\nu; \phi) := \int_M \tilde{\phi}(\theta)\, d\nu(\theta),$$

$$\tilde{\theta} = g(\exp(\phi(\theta)), \quad g(t) = t^{-1}, \quad t > 0$$

(3) Given posterior u

(a) $D^u(f) = \int_M \|\nabla_g f(\theta)\|^2 \, du(\theta)$

$$\min \int_M \|\nabla_g f(\theta)\|^2 \, du(\theta)$$
$$\rho_u \in L^2(M, u) \text{ s.t. } \int_M \rho \, du = 1$$

Geodesic convexity & functional inequalities

$\pi = e^{-\psi} \mathrm{vol}_g$, $\mu \propto e^{-\phi} \pi$

Prop. 3.1: $\psi, \phi \in C^2(M)$, J_{KL} is λ-geodesically convex iff

$$\mathrm{Ric}_g(v,v) + \mathrm{Hess}_g \psi(v,v) + \mathrm{Hess}_g \phi(v,v) \geq \lambda$$

$\forall x \in M$, $\forall v \in T_x M$

Prop. 3.2: J_{χ^2} is λ-geodesically iff the following hold

1. $\mathrm{Ric}_g(v,v) + \mathrm{Hess}_g \psi(v,v) + \frac{1}{m+1} \langle \nabla_g \psi, v \rangle^2 \geq 0$

$\forall x \in M$, $\forall v \in T_x M$

2. ϕ is a λ-geodesically convex fcn.

Convexity notions:

Convexity implies convergence / flows of

$\pi \to \mu$ (posterior)

$$\pi \to \mu_t \to \mu.$$

$$D_{KL}(\mu_t \| \mu) \le e^{-\lambda t} D_{KL}(\pi \| \mu)$$

$$W_2(\mu_t \| \mu) \le \sqrt{\lambda} e^{-\lambda t} D_{KL}(\pi \| \mu)$$

A measure μ has Poincaré inequality λ if $\forall f \in L^2(\pi\lambda, \mu)$ & $\int_m f \, d\mu = 0$

$$\|f\|_\mu^2 \le \frac{1}{\lambda} D^\mu(f)$$

and D^μ is 2λ-geodesically convex $\Leftrightarrow$ Poincaré constant λ.

Spectral gap:

Poincaré constant is the smallest non-trivial eigenvalue of

$$-\Delta_g^{\mu} f := -\frac{1}{2} \operatorname{div}_g \left(e^{-\mathcal{V}-\chi} \nabla_g f \right)$$

$$\lambda_2 := \min_{f \in L^2(m,\mu)} \frac{D^{\mu}(f)}{\|f - f_m\|_{\mu}^2}$$

$$f_{\mu} := \int_m f \, d\mu$$

PDEs & Diffusions

1) J_{KL} + Wasserstein flow

gradient flow $t \in (0,\infty) \mapsto \mu_t$ of $D_{KL}(\cdot \| \mu)$ starting at $\mu_0 = \pi \in \mathcal{P}(M)$

$$\rho_t := \frac{d\mu_t}{d\mu}, \quad \theta_t := \frac{d\mu_t}{dvol_g}$$

satisfy Fokker Planck

$$\frac{\partial \rho}{\partial t} = \Delta_g^{\mu} \rho$$

$$\frac{\partial \theta}{\partial t} = \Delta_g \theta + div_g(\theta(\nabla_g \varphi + \nabla_g \psi))$$

$$dX_t = -\nabla_g(\psi(X_t) + \varphi(X_t))\,dt + \sqrt{2}\, dB_t'$$

2) J_{χ^2} & Wasserstein flow

$$\tilde{\rho}_t := \frac{d\mu_t}{d\pi}$$

J_{χ^2} - Wasserstein flow $t \in (0,\infty) \to \mu_t$

$$\frac{\partial \tilde{\rho}}{\partial t} = \Delta_g^{\pi} \tilde{\rho}^2 + \operatorname{div}(\tilde{\rho} \nabla_g \varphi)$$

$$\Delta_g^{\pi} := \Delta_g f - \langle \nabla_g f, \nabla_g \psi \rangle$$

$$\operatorname{div}_g^{\pi} F := \operatorname{div}_g F - \langle F, \nabla_g \psi \rangle$$

$$dX_t = -(\tilde{\rho}(t, X_t) \nabla_g \psi(X_t) + \nabla_g \varphi(X_t))dt + \sqrt{2\tilde{\rho}(t,X_t)}\, dB_t^g, \quad u_0 \sim \rho_0$$

$$\frac{\partial \theta}{\partial t} = -\operatorname{div}_g(\theta(-\tilde{\rho}\nabla_g \psi - \nabla_g \varphi)) + \Delta_g(\tilde{\rho}\theta)$$

$$\beta = \frac{1}{Z} \exp(-\psi)\theta$$

$$\Delta_g^{\pi} \beta^2 = e^{\psi}\left(\Delta_g(\beta^2 e^{-\psi}) + \operatorname{div}_g((\beta^2 e^{-\psi})\nabla_g \psi)\right)$$

Stop and talk about Bayes updating vs. PDE idea.

Semi supervised learning -

$$\{(X_i, y_i)\}_{i=1}^{n}, \quad \{(\hat{x}_i)\}_{i=n+1}^{n+m}$$

Use a graph diffusion to propogate labels from labeled points to unlabeled points

$G = (X, W)$, $W_{ij} = k\left(\frac{|x_i - x_j|}{r}\right)$

k: compact Kernel fcn.

$L = D - W$, $D_{ii} = \sum_j w_{ij}$

Probit model

$$G(u) := \frac{1}{2}\langle L^{\alpha} u, u\rangle - \sum_j \log\left(H(y_j u_j)_{j,\gamma}\right)$$

$$H(w;\gamma) = \int_{-\infty}^{w} \exp\left(-t^2/2\gamma^2\right) dt$$

$$\hat{u}_{in} = \underset{\substack{u \in \mathbb{R}^n \\ \sum_i u_i = 0}}{\arg\min} \; G(u)$$

$$\hat{u} = \operatorname{sign}(\hat{u}_{in}) .$$

Prior - $U := \{u \in \mathbb{R}^n : \sum u_i = 0\}$

$$\frac{d\pi}{du}(u) \propto \exp\left(-\tfrac{1}{2}\langle L^{\alpha} u, u\rangle\right)$$

$$:= \exp(-\psi_u)$$

Likelihood:

$$y_i = \operatorname{Sign}(u_i + \eta_i)$$

$$\eta_j \overset{iid}{\sim} N(0, \gamma^2)$$

$$\Phi(u;\gamma) := -\sum_j \log\left(H(y_j u_j;\gamma)\right)$$

$$u \in U$$

Posterior :

$$\frac{d\mu^y}{d\pi}(v) = \exp(-\phi(v;y))$$

$$\frac{d\mu^y}{dv}(v) \propto \exp(-\psi(v) - \phi(v;y))$$

$$J_{KL}(v) := D_{KL}(v \| \pi) + \int_{\mathbb{R}^n} \phi(u;y)\, dv(u)$$

$$v \in P(\mathbb{R}^n)$$

1) π is Gaussian with $\Sigma = L^{-\alpha}$

$\Lambda(L)$ is the smallest n.t. eigenvalue

2) $\phi(\cdot;y)$ is convex

$F_{KL}(v) = \int_{\mathbb{R}^n} \phi(u;y)\, dv(u)$ is

θ-geodesically convex.

diffusion equation

$$dX_t = \left(-L^{\alpha} X_t - \nabla \phi(X_t)\right) dt + \sqrt{2}\, dB_t$$

as n gets large L^{α} has issues

$$dX_t = -\left(X_t + L^{-\alpha} \nabla \phi(X_t)\right) dt + \sqrt{2L^{-\alpha}}\, dB_t$$

$$\nabla_g = G^{-1} \nabla \,, \quad B^g = \sqrt{G^{-1}}\, B$$

$$dX_t = -\nabla_g\left(\phi(X_t) + \tfrac{1}{2} \langle L^{\alpha} X_t, X_t \rangle\right) dt + \sqrt{2}\, dB_t^g$$