Variational formulations of Bayesian Inference

Overview of materials

Lec 1
1) Review of Bayesian (finite dimensional) parametric inference

2) Infinite dimensional Bayesian inference - Gaussian process model

Lec 2
3) Variational formulations of gradient updating and gradient flows

Lec 3
4) Posterior consistency
   a) Classic test fcn / sieve approach
   b) Variational approach

based on large deviations

Lec 4 5) Bayesian inverse problems

Bayes : Two manuscripts

1) Bayes' rule

2) A response to :
The analyst: or, a discourse addressed to an infidel mathematician

Given sets A & B

$$P(B|A) = \frac{P(A|B)\,P(B)}{P(A)}$$

there is a symmetry

Bayesian inference:

Likelihood or data generating process $f(x_1, \dots, x_n | \theta)$, $\theta \in \Theta$

$$f(x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i | \theta) \quad \text{iid}$$

Prior: $\pi(\theta)$ - belief you see data

Posterior:

$$\pi(\theta | x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n | \theta) \pi(\theta)}{\underbrace{\int_\Theta f(x_1, \dots, x_n | \theta) \pi(\theta) d\theta}_{\text{marginal likelihood or } P(D)}}$$

There is an assymetry.

a) Quantifies uncertainty

b) Need to specify a likelihood

c) Need a prior

d) The normalization constant

Contrast to frequentist/proceduralist approach.

Ex. 1. Binomial

$$f(x \mid n, p) = \binom{n}{x} p^{x} (1-p)^{n-x}$$

$$\pi(p) \sim Beta(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1} (1-p)^{\beta-1}$$

$$\alpha, \beta > 0$$

$$\pi(p \mid x, \alpha, \beta) = \frac{p^{x+\alpha-1}(1-p)^{n-x+\beta+1}}{\int_0^1 p^{x+\alpha-1}(1-p)^{n-x+\beta+1}\, dp}$$

$= \text{Beta}(x+\alpha, n-x+\beta)$

The Beta is conjugate prior for the binomial likelihood.

Posterior consistency: The posterior $\Pi_n(\theta \mid x_1, \dots, x_n)$ is consistent at θ_0 if for every neighborhood U of θ_0

$$\Pi_n(U) \to 1 \quad \text{a.s. under } \theta_0.$$

For observations with finite # of values at any point θ_0 which belongs in the support of Π.

Counterexample:

Infinite multinomial - Infer a pmf on the set of integers.

Let θ_0: $\Pr(X=k) = ((1-p)^{k} p$ be geometric.

One can construct a prior π that gives positive mass to every neighborhood of θ_0 but the posterior concentrates around

$$\theta_n : \Pr(X=k) = (1-p)^{9k} p$$

Bayesian hierarchical model:

Often count data is modeled as Poisson

$$f(m|\lambda) = \frac{m!\, e^{-\lambda}}{\lambda^m}$$

$$\mathbb{E}(m) = \mathrm{Var}(m) = \lambda$$

Often $\mathrm{Var}(m) > \mathbb{E}(m)$

$$\lambda \sim \mathrm{Gamma}\left(r, \frac{p}{1-p}\right)$$

$$f(\lambda \mid r, p) = \frac{\left(\frac{p}{1-p}\right)^r}{\Gamma(r)} \lambda^{r-1} e^{-p/1-p}$$

$$m \mid \lambda \sim \mathrm{Pois}(\lambda)$$

$$f(m \mid \lambda) = \frac{e^{-\lambda}\, m!}{\lambda^m}$$

$$f(m, \lambda \mid r, p) = \frac{\left(\frac{p}{1-p}\right)^r}{\Gamma(r)} \lambda^{r-1} e^{-p/1-p} \; \frac{e^{-\lambda}\, m!}{\lambda^m}$$

Integrate out λ:

$$\int_0^\infty f(m, \lambda \mid r, p)\, d\lambda =$$

$$f(m \mid r, p) = \frac{\Gamma(m+r)}{\Gamma(r)\, m!} p^r (1-p)^m$$

\# of r successes before m failures

Exponential family:

Given a family of distributions with real parameters, $\theta = [\theta_1, \dots, \theta_s]^T$

The family can be written

$$f(x|\theta) = h(x)\, g(\theta)\, \exp(\eta(\theta) \cdot T(x))$$

T = sufficient statistic

η = natural parameterization.

$$f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$$

$$\eta = \left[\frac{\mu}{\sigma^2} \quad -\frac{1}{2\sigma^2}\right]$$

$$h(y) = \frac{1}{\sqrt{2\pi}}$$

$$T(y) = (y, y^2)^T$$

$$g(\theta) = \frac{\mu^2}{2\sigma^2} + \log|\sigma|$$

Diaconis & Ylvisaker:

Given an exponential family likelihood conjugate priors satisfy

$$\mathbb{E}(\mathbb{E}(X \mid \theta) \mid X = x) = ax + b$$

Gaussian process - Infinite dimensional model.

Defn. A Gaussian process $\{X_t\}_{t \in T}$ indexed by a set T is a family of r.v.'s that for any finite subset $F \subset T$ $X_F := \{X_t\}_{t \in F}$ is MVN. If X_F is non-degenerate MVT for all F then $\{X_t\}$ is a non-degenerate GP.

For GP $\{X_t\}_{t \in T}$ with mean fcn.

$$\mu_t = \mathbb{E}(X_t)$$

and covariance kernel

$$R(X_s, X_t) = \mathrm{Cov}(X_s, X_t)$$

is positive (semi) definite

and $\{X_t\}$ is non-degenerate then for any finite $F \subset T$

$$X_F \sim MVN(\mu_F, \Sigma_F)$$

$$\mu_F = \begin{pmatrix} \mu(X_1) \\ \vdots \\ \mu(X_{F_n}) \end{pmatrix}$$

$$\Sigma_{F,i,j} = Cov(X_i, X_j)$$

GP examples:

Brownian motion - W_t

$$\mathbb{E}(W_s W_t) = \min(s,t)$$

Ornstein-Uhlenbeck - Y_t

$$\mathbb{E}(Y_s Y_t) = \exp(-|t-s|)$$

Brownian bridge - W_t°

$$\mathbb{E}(W_s^\circ W_t^\circ) = \min(s,t) - st$$

Some theory and practice of GP

GP prior regression

Prior - $f \sim GP(\mu(\cdot), K(\cdot,\cdot))$

$$\mu(x) = \mathbb{E}(f(x))$$

$$K(x_i, x_j) = Cov(f(x_i), f(x_j))$$

Likelihood model:

$$Y_i = f(x_i) + \varepsilon_i,$$

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$X = \begin{bmatrix} - x_1 - \\ - x_2 - \\ \vdots \\ - x_n - \end{bmatrix}, \quad X^* = \begin{bmatrix} - x_1^* - \\ \vdots \\ - x_m^* - \end{bmatrix},$$

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad Y^* = \begin{bmatrix} Y_1^* \\ \vdots \\ Y_m^* \end{bmatrix}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \quad \varepsilon^* = \begin{bmatrix} \varepsilon_1^* \\ \vdots \\ \varepsilon_m^* \end{bmatrix}$$

$$f = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}, \quad f^* = \begin{bmatrix} f^*(x_1) \\ \vdots \\ f^*(x_n) \end{bmatrix}$$

want $Y^* \mid Y, X, X^* \sim N(\mu^*, \Sigma^*)$

$$\begin{bmatrix} Y \\ Y^* \end{bmatrix} \Big| X^*, X = \begin{bmatrix} f \\ f^* \end{bmatrix} + \begin{bmatrix} \varepsilon \\ \varepsilon^* \end{bmatrix}$$

$$\sim N\left(0, \begin{bmatrix} K(X,X)+\sigma^2 I & K(X^*,X) \\ K(X^*,X) & K(X^*,X^*)+\sigma^2 I \end{bmatrix}\right)$$

$$\mu^* = K(X^*,X)\left(K(X,X)+\sigma^2 I\right)^{-1} Y$$

$$\Sigma^* = K(X^*,X^*) + \sigma^2 I - K(X^*,X)\left(K(X,X)+\sigma^2 I\right)^{-1} K(X,X^*)$$

Some theory - Empirical

Process theory

1) GPs on manifolds

Given compact Riemannian manifold (M, g) can we define a GP supported on M?

Näive idea.

$$K(x, x') = \sigma^2 \exp\left(- \frac{d_g(x, x')^2}{2\kappa^2}\right) \quad (5)$$

Theorem: If (5) is PSD for all $\kappa^2 > 0$ then M is isometric to a Euclidean space.

Solution:

$$\left(\frac{2\nu}{\kappa^2} - \Delta_M\right)^{\frac{\nu}{2} + \frac{d}{4}} f = \mathcal{W}_M$$

$$e^{-\frac{\kappa^2}{4}\Delta_M} f = \mathcal{W}$$

$\Uparrow$
solve SPDEs
to get
$K(x, x')$

2) Extrema of Gaussian processes

Def. 2.1: Let (T, d) be a compact metric space. For each $\varepsilon > 0$ the (Lebesgue) covering # $N(\varepsilon)$ is the minimum number of ε-balls to cover T.

Thm: Let d be the canonical metric

$$d(s,t) = \sqrt{\mathbb{E}\,|X_s - X_t|^2}$$

of a non-degenerate, centered, GP an $N(\varepsilon)$ is the covering #.

If for some $\rho > 0$

$$\int_0^\rho \sqrt{\log N(\varepsilon)}\, d\varepsilon < \infty$$

the GP has uniformly continuous sample paths and

$$\sup_{t \in T} X_t = \max_{t \in T} X_t < \infty.$$

Concentration of sup:

Thm: Let $\{X_t\}_{t \in T}$ be a centered GP on a countable T that is a.s. bounded. $X^* = \sup_t X_t$.

If $\sigma_T^2 := \sup_{t \in T} \mathbb{E} X_t^2 < \infty$,

then $\mathbb{E} X^* < \infty$ and for any $u > 0$

$$P(X^* \geq \mathbb{E} X^* + u) \leq e^{-u^2/2\sigma_T^2}$$

3) Reproducing Kernel Hilbert space

Given a positive definite Kernel $k(\cdot,\cdot)$ that is

$\forall\ t_1, \dots, t_n \in \mathcal{X}$ & $\forall\ a_1, \dots, a_n \in \mathbb{R}$

and all $n \in \mathbb{N}$

$$\sum_{ij}^{n} a_i a_j K(x_i, x_j) > 0.$$

Define $L_K : L_2(x) \to C(x)$

$$L_K f := \int_X K(s,t) f(t) dt = g(t)$$

Eigenvalues & eigenvectors

$$\int_X K(s,t) \phi_k(t) dt = \lambda_k \phi_k(t)$$

$\forall\ k$

$$K(s,t) = \sum_j \lambda_j \phi_j(s) \phi_j(t)$$

$$\| f(s) \|_{H_K}^2 = \sum_j \langle c_j \phi_j(s), c_j \phi_j(s) \rangle$$

$$:= \sum_j \frac{c_j^2}{\lambda_j}$$

$$\langle f, g \rangle = \left\langle \sum c_j \phi_j(s), \sum_j d_j \phi_j(s) \right\rangle$$

$$= \sum \frac{c_j d_j}{\lambda_j}$$

$$H_K = \overline{\left\{ f \mid f(s) = \sum_k c_k \phi_k(s) \text{ and } \|f\|_{\mathcal{H}_K} < \infty \right\}}$$

$$\langle f(\cdot), K(\cdot, x) \rangle_{\mathcal{H}_K} = f(x)$$