Variational formulations of Bayesian Inference

Overview of materials

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

2) Infinite dimensional Bayesian inference - Gaussian process model
   Lec 2

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

4) Posterior consistency
   Lec 3
   a) Classic test fcn/sieve approach
   b) Variational approach
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 \( f(x_1, \ldots, x_n | \theta) \), \( \theta \in \Theta \)
\[
f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i | \theta) \quad \text{iid}
\]
Prior: \( \pi(\theta) \) - belief you see data

Posterior:
\[
\pi(\theta | x_1, \ldots, x_n) = \frac{f(x_1, \ldots, x_n | \theta) \pi(\theta)}{\int_\Theta f(x_1, \ldots, x_n | \theta) \pi(\theta) d\theta}
\]

marginal likelihood or \( p(x | \Theta) \)

There is an asymmetry.
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
\[ f(x \mid \lambda, \mu) = \binom{n}{x} \mu^x (1 - \mu)^{n-x} \]
\[ \Pi(\mu) \sim \text{Beta}(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} \mu^{\alpha-1} (1 - \mu)^{\beta-1} \]
\[ \alpha, \beta > 0 \]

\[ \Pi(\mu \mid x, \alpha, \beta) = \frac{\mu^{x+\alpha-1} (1 - \mu)^{n-x+\beta-1}}{\int_0^1 \mu^{x+\alpha-1} (1 - \mu)^{n-x+\beta-1} \, d\mu} \]
\[ = \text{Beta}(x+\alpha, \, n-x+\beta) \]

The Beta is conjugate prior for the binomial likelihood.

Posterior consistency: The posterior \( \pi_n(\theta | x_1, \ldots, x_n) \) is consistent at \( \theta_0 \) if for every neighborhood \( V \) of \( \theta_0 \)

\[ \pi_n(V) \rightarrow 1 \quad \text{as} \quad n \rightarrow \infty \]

For observations with finite number of values at any point \( \theta_0 \) which belongs in the support of \( \pi_n \).
Counterexample:
Infinite multinomial - Infer a pmf on the set of integers.
Let \( \Theta_0 : \Pr(X = k) = (1 - \rho)^{\infty} \rho \)
be geometric.

One can construct a prior \( \pi \) that gives positive mass
to every neighborhood of \( \Theta_0 \) but the posterior concentrates around
\( \Theta_n : \Pr(X = k) = (1 - \rho)^{\infty} \rho \)

Bayesian hierarchical model:
Often count data is modeled as \( \text{Poisson} \)
\( f(m|\lambda) = \frac{m! e^{-\lambda}}{\lambda^m} \)
\( \mathbb{E}(m) = \text{Var}(m) = \lambda \)
Often \( \text{Var}(m) > \mathbb{E}(m) \)
\[ \lambda \sim \text{Gamma}(r, \frac{p}{1-p}) \]
\[ f(\lambda | r, p) = \frac{\left( \frac{p}{1-p} \right)^r \lambda^{r-1} e^{-\lambda \frac{p}{1-p}}}{\Gamma(r)} \]

\[ m | \lambda \sim \text{Pois}(\lambda) \]
\[ f(m | \lambda) = \frac{e^{-\lambda} \lambda^m}{\lambda^m} \]

\[ f(m, \lambda | r, p) = \frac{\left( \frac{p}{1-p} \right)^r \lambda^{r-1} e^{-\lambda \frac{p}{1-p}}}{\Gamma(r)} \frac{e^{-\lambda} \lambda^m}{\lambda^m} \]

Integrate out \( \lambda \):
\[ \int_0^\infty f(m, \lambda | r, p) \, d\lambda = \]
\[ f(m | r, p) = \frac{\Gamma(m+r)}{\Gamma(r) m!} \frac{\rho^r (1-p)^m}{\Gamma(r) m!} \]

\# of successes before \( m \) failures
Exponential family:

Given a family of distributions with real parameters, \( \Theta = (\alpha, \ldots, \theta)^T \), the family can be written

\[
   f(x|\theta) = h(x) g(\theta) \exp(\eta(\theta) \cdot T(x))
\]

\( T \) is sufficient statistic

\( \eta \) is natural parameterization.

\[
   f(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left( -\frac{(y-\mu)^2}{2\sigma^2} \right)
\]

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

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

\[
   T(y) = (y, y^2)^T
\]

\[
   g(\theta) = \frac{\theta^2}{2\sigma^2} + \log(\theta)
\]

Diaconis & Ylvisaker:

Given an exponential family, likelihood conjugate priors satisfy
\[ E(E(x_1) | X=x) = ax + b \]

**Gaussian process - Infinite dimensional model.**

**Defn.** A Gaussian process \( \{X_t\}_{t \in T} \) on a set \( T \) is a family of r.v.'s that for any finite subset \( F \subset T \) \( \{X_F\} \) is MVN. If \( X_F \) is non-degenerate MVN for all \( F \) then \( \{X_t\} \) is a non-degenerate GP.

For GP \( \{X_t\}_{t \in T} \) with mean fn.

\[ \mu_t = E(X_t) \]

and covariance kernel positive definite

\[ R(X_s, X_t) = \text{Cov}(X_s, X_t) \]
and \( \mathbb{E}(X_t) \) is non-degenerate then for any finite FCT

\[
X_F \sim \text{MVN}(\mu_F, \Sigma_F)
\]

\[
\mu_F = \begin{pmatrix} \mu(x_i) \\ \vdots \\ \mu(x_m) \end{pmatrix}
\]

\[
\Sigma_{s,t} = \text{cov}(x_i, x_j)
\]

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 - \( W^*_t \)

\[
\mathbb{E}(W^*_s, W^*_t) = \min(s, t) - st
\]
Some theory and practice of GP

GP prior regression

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

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

$K(x_i, x_j) = \text{Cov}(f(x_i), f(x_j))$

Likelihood model:

$Y_i = f(x_i) + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$, $X^* = \begin{bmatrix} x_{n+1} \\ \vdots \\ x_m \end{bmatrix}$,

$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$, $Y^* = \begin{bmatrix} y_{n+1} \\ \vdots \\ y_m \end{bmatrix}$

$\epsilon = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$, $\epsilon^* = \begin{bmatrix} \epsilon_{n+1} \\ \vdots \\ \epsilon_m \end{bmatrix}$
\[
f = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}, \quad f^* = \begin{bmatrix} f^*(x_1) \\ \vdots \\ f^*(x_n) \end{bmatrix}
\]
want \( y^* | X, X, x^* \sim N(\mu^*, \Sigma^*) \)

\[
\begin{bmatrix} y \\ y^* \end{bmatrix} | X, x^* \sim \begin{bmatrix} f \\ f^* \end{bmatrix} + \begin{bmatrix} \varepsilon \\ \varepsilon^* \end{bmatrix}
\]

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

\[
\mu^* = K(x^*, x) (K(X, x) + \sigma^2 I)^{-1} y
\]

\[
\Sigma^* = K(x^*, x^*) + \sigma^2 I - K(x^*, x) (K(X, x) + \sigma^2 I)^{-1} K(x, x^*)
\]

---

Some theory - Empirical
1) GPs on manifolds

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

Naive idea.

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

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

Solution:

$$\left(\frac{2\nu}{x^2} - \Lambda_n\right)^{\frac{\kappa}{2}} \frac{d}{dx} f = \mathcal{V}_n$$

$$e^{-\frac{\kappa}{2} \Lambda_n} f = \mathcal{W}$$
Let $d$ be the canonical metric of a non-degenerate, centered, $N(0, E)$.

Theorem: Let $d$ be the canonical metric $d(s, t) = \sqrt{\int_E |x_s - x_t|^2 dt}$.

Def. 2.1: Let $(T, d)$ be a compact metric space. For each $\varepsilon > 0$, let $N(\varepsilon)$ be the minimal number of $\varepsilon$-balls to cover $T$. Extrema of Gaussian processes to solve \( k(x, x') \)
If for some $\sigma > 0$
\[\int_0^\infty \sqrt{\log N(t)} \, \mathrm{d}t < \infty\]
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 \(u > 0\)
\[\mathbb{P}(X^* \geq \mathbb{E} X^* + u) \leq e^{-\frac{u^2}{2 \sigma_T^2}}\]
3) Reproducing Kernel Hilbert space

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

\[
\forall t_1, \ldots, t_n \in \mathbb{R} \quad \forall a, \ldots, a_n \in \mathbb{R}
\]

and all \( n \in \mathbb{N} \)

\[
\sum_{i, j} a_i a_j k(x_i, x_j) > 0.
\]

Define \( L_k : L_2(x) \to \mathbb{C} \)

\[
L_k f := \int \! K(s, t) f(t) \, dt = g(s)
\]

Eigenvalues and eigenvectors

\[
\int \! K(s, t) \varphi_k(t) \, dt = \lambda_k \varphi_k(s)
\]

\( \forall k \)

\( K(s, t) = \sum_i \lambda_i \varphi_i(s) \varphi_i(t) \)

\[
\|f(s)\|_{H_k}^2 = \sum_i c_i \varphi_i(s), c_i \varphi_i(s)
\]
\[
\hat{c}_i = \sum_i \frac{c_i}{\lambda_i}
\]

\[
\langle f, g \rangle = \langle \sum_i c_i \delta_i(s), \sum_j d_j \delta_j(u) \rangle
\]

\[
= \sum_i \frac{c_i d_i}{\lambda_i}
\]

\[
H_k = \{ f \mid f(s) = \sum_k c_k \delta_k(s) \text{ and } \| f \|_{H_k} < \infty \}
\]

\[
\langle f(c), k(\cdot,x) \rangle_{H_k} = f(x)
\]