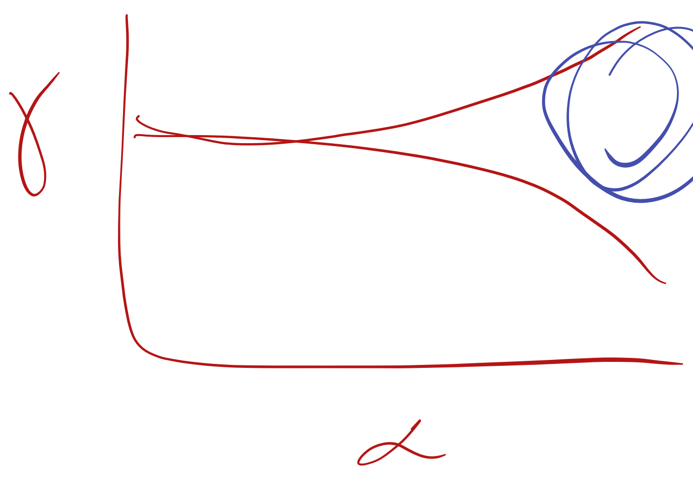
Approximate Inference

- HW 5 (last one), due 2 weeks
- Reading 13.7 (Variational Inference)
- Reminder: submit Project "proposals"



Approximate Inference

- MCMC can be very slow in high dimensional problems
- Idea: find a distribution that is easy to sample from which closely approximate $p(\theta \mid y)$
- A couple of examples
  - Laplace Approximation – Normal Approximation
  - INLA – Iterated Nested Laplace Approx.
  - Variational Bayes

(13.7)

Mean-Field Variational Bayes

Want: 1. g close to $p(\theta|y)$

- Solve an optimization problem:

$$\underset{\lambda}{\operatorname{argmin}} \operatorname{dist}(p(\theta \mid y), g_\lambda(\theta))$$

2. g easy to sample from

- In mean-field inference we restrict $g(\theta_1, \ldots, \theta_d) = \prod_i g(\theta_i)$
- For VB, we usually use KL-divergence to measure the "distance" between probability distributions
  - KL is not a true distance since $KL(p|q) \neq KL(q|p)$.

Kullback-Leibler Divergence

Mean-Field Variational Bayes

Reverse-KL

- Optimize $KL(q(\theta)||p(\theta \mid y)) = -E_q \log(\frac{p(\theta|y)}{q(\theta)})$
- Problem: expectation under q means that the variational distribution usually underestimates posterior uncertainty

Approximating Distn

$KL(q||p) \geq 0$

$KL(q||p) = 0$ iff $q = p$

$q(\theta) \ll p(\theta|y)$: $supp(q) \subseteq supp(p(\theta|y))$

$p(\theta|y) = 0 \Rightarrow q(\theta) = 0$

$$\text{Fwd: } KL(p \,\|\, q) = -E_p \log \frac{q}{p}$$

Typically can't use in practice

$$KL(q(\theta) \,\|\, p(\theta|y)) =$$

$$-E_q\left(\log \frac{p(\theta|y)}{q(\theta)}\right) = E_q \log q(\theta) - E_q\left[\log \overbrace{p(\theta|y)}^{\frac{p(\theta,y)}{p(y)}}\right]$$

$$= E_q \log q(\theta) - E_q \log p(\theta,y) + \underbrace{\log p(y)}_{\text{model evidence}}$$

$$\geq 0$$

$$\log p(y) \geq \underbrace{E_q \log p(\theta,y) - E_q \log q}_{\text{Evidence Lower Bound (ELBO)}}$$

ELBO: $E_q \log P(\theta, y) - E_q \log q =$

$E_q \log P(y|\theta)P(\theta) - E_q \log q =$

$E_q \log P(y|\theta) + \underbrace{E_q \log P(\theta) - E_q \log q}_{KL(q(\theta) \| P(\theta))}$

$= \underbrace{E_q \log P(y|\theta)}_{\text{Favors } q(\theta) \text{ which explain the data.}} + \underbrace{KL(q(\theta) \| P(\theta))}_{\text{Favor } q \text{ close to the prior}}$

Mean-Field VI.

$q(\theta_1, \dots \theta_d) = q(\theta_1) \dots q(\theta_d)$

Coordinate Updates:

Update one $q(\theta_i)$ at a time

conditional on the rest.

$$E_q \log \frac{p(\theta,y)}{q(\theta)} = \int_{\theta_d} \cdots \int_{\theta_j} \log\left(\frac{p(\theta,y)}{q(\theta_1)\cdots q(\theta_d)}\right) q(\theta_1)\cdots q(\theta_d)\, d\theta_1 \ldots d\theta_d$$

$$= \int_{\theta_{-j}} \left[\int_{\theta_j} \log\left(\frac{p(\theta,y)}{q(\theta_j)}\right) q(\theta_j)\, d\theta_j\right] d\theta_{-j} -$$

$$\int_{\theta_{-j}} q(\theta_{-j}) \log \theta_{-j}\, d\theta_{-j}$$

Ex: $j = 3$:

$$q(\theta_{-j}) = q_1(\theta_1)\, q_2(\theta_2)\, q_4(\theta_4) \cdots q_d(\theta_d)$$

To update $q_j(\theta_j)$ (fixing the rest)

$$\max_{q_j} \int_{\theta_{-j}} \int_{\theta_j} \log\left(\frac{p(\theta,y)}{q(\theta_j)}\right) q(\theta_j)\, d\theta_j -$$

$$\max_{q_j} \int_{\theta_j} E_{q_{-j}(\theta_{-j})} \log\left(\frac{P(\theta, y)}{q_j(\theta_j)}\right) q_j(\theta_j)\, d\theta_j$$

$$\max_{q_j} E_{q_j} E_{q_{-j}} \log P(\theta, y) - E_{q_j} \log q_j$$

call: $\tilde{P}_j(\theta_j, y) : E_{q_{-j}} \log P(\theta, y)$

$$\max_{q_j} E_{q_j} \frac{\tilde{P}_j(\theta_j, y)}{q_j} = -KL(q_j, \tilde{P}_j)$$

$$\log q_j(\theta) = E_{q_{-j}} \log P(\theta, y) + \text{const}$$

Mean Field Algo:

Until ELBO converges:

for j in 1:D {

$q_j(\theta) \leftarrow$ distribution $\propto \underline{E_{q_{-j}} \log P(\theta, y)}$

}

$$y_i \sim N(\theta_i, \sigma^2)$$

$$\theta_i \sim N(\mu, \tau^2)$$

$$p(\mu) \propto \text{const}$$

$$\log p(\theta_1, \ldots \theta_8, \mu, y_1, \ldots y_8) =$$

$$\log \prod_{i=1}^{8} \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \theta_i)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi\tau^2}} e^{-\frac{(\theta_i - \mu)^2}{2\tau^2}} \right]$$

until convergence:

update $q_1(\theta_1)$ given q_{-1}

$\vdots$

$q_8(\theta_8)$

$q(\mu)$ given others

$$q_j(\theta_j) \propto E_{q_{-j}}\left[-\frac{(y_j-\theta_j)^2}{2\sigma^2} - \frac{(\theta_j-\mu)^2}{2\tau^2}\right]$$

$$\propto e^{-\frac{(y_j-\theta_j)^2}{2\sigma^2} - E_{q_{-j}}\left[\frac{\theta_j^2-2\theta_j\mu+\mu^2}{2\tau^2}\right]}$$

$$\propto e^{-\frac{(y_j-\theta_j)^2}{2\sigma^2} - \frac{\theta_j^2}{2\tau^2} - \frac{2\theta_j E_{q(\mu)}\mu + E_{q(\mu)}[\mu^2]}{2\tau^2}}$$

$$q_\mu(\mu) \propto e^{E_{q_{-\mu}}\sum_{i=1}^{8}\frac{(\theta_i-\mu)^2}{2\tau^2}}$$

$$\propto e^{E_{q_{-\mu}}\sum\frac{\theta_i^2-2\mu\theta_i+\mu^2}{2\tau^2}}$$

$$\propto e^{\sum_{i=1}^{8}\frac{E_{q(\theta_i)}[\theta_i^2]-2E_{q(\theta_i)}[\mu]+\mu^2}{2\tau^2}}$$

Automatic Differentiation Variational Inference

See: https://arxiv.org/pdf/1603.00788.pdf (Blei et al)

ADVI implemented in STAN.

Remarks: $q \ll p(\theta|y)$

$\mathrm{supp}(q(\theta)) \subseteq \mathrm{supp}(p(\theta|y))$

Define 1-1 transformation T,

d

from $\mathrm{supp}(\theta) \longrightarrow \mathbb{R}^{d}$

$$\phi = T(\theta), \quad \phi \in \mathbb{R}^{d}$$

$$\underline{KL(q(\phi) \,\|\, P(\phi | y))}$$

$$P(\phi | y) = P_{\theta}(T^{-1}(\phi) | y) \, |J_{T^{-1}}(\phi)|$$

$$q(\phi) \sim N(\mu_{d}, \Sigma_{d \times d})$$

$$\min_{\mu, \Sigma} KL(q(\phi) \,\|\, P(\phi | y))$$

$$\underline{ELBO} = E_{q(\mu, \Sigma)} \log P_{\theta}(T^{-1}(\phi), y) + \log |J_{T^{-1}}(\phi)|$$

$$- E_{q} \log q(\phi)$$

$$\mu^*, \Sigma^* = \underset{\mu, \Sigma}{\arg\max}\ \mathrm{ELBO}(\mu, \Sigma)$$

$$\eta = S(\phi) = \Sigma^{-1/2}(\phi - \mu)$$

$$\Rightarrow q(\eta) \sim N(0, I)$$

ELBO:

$$\mathbb{E}_{N(0,I)} \overbrace{\log P(T^{-1}(S^{-1}_{\mu,\Sigma}(\eta)), y)}^{\log P(\theta, y)} + \log |J_{T^{-1}}(S^{-1}(\eta))| - \mathbb{E}_{q(\phi)} \log q(\phi)$$