Intro to Variational Inference & BBVI

Overview

Observed data
$\{x_n, y_n\}$

Assume Model to Generate Data
(1) Prior $p(z)$
(2) Likelihood $p(y|z,x)$

Goal: Estimate posterior $p(z|x,y)$ given observed data

When can I use posterior inference methods?

	General MCMC	HMC (specific)	General VI	BBVI
both prior + lik need to be Gaussian	✓	✓	✓	✓
prior only is a Gaussian	✓	✓	✓	✓
z Continous → prior + lik have differentiable log pdf (z is cont.)	✓	✓	✓	✓
z discrete → prior + lik have pdfs we cannot differentiate (z has discrete)	✓	✗	✓	✓

$$\text{PDFCat}(z) = \prod_{k=1}^{K} p_k^{\delta(z=i)}$$

What tunable settings does HMC require?
#leapfrog steps, step size

What tunable setting does BBVI require?
step size, # monte carlo samples

Variational Inference

Goal: Estimate a posterior distribution $q(z|x,y)$ (z or w here)

Any distr. can be represented two possible ways

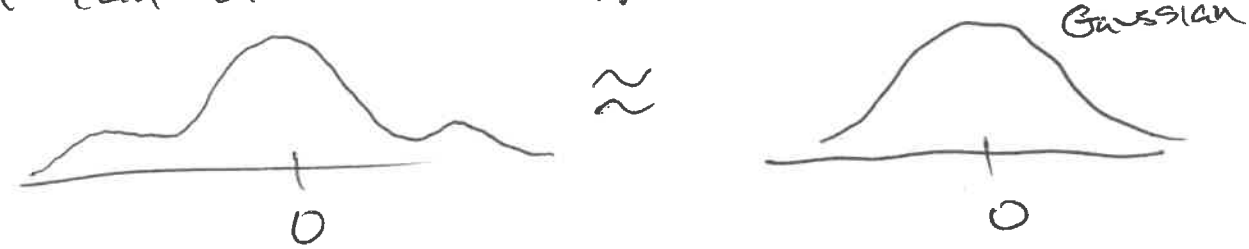
- set of samples $\{w^s\}_{s=1}^{S}$
- as a known density function w/ known parameters

Normal w/ mean + stddev

Poisson w/ mean

etc.

Not all distributions belong to known family, but can often make approximations



Idea of VI:

Choose an approximating family of density functions call this $q_\nu(w)$ with possible parameters $\nu \in V$

Goal: Make $q(w)$ as "close" as possible to $p(w|y,x)$ by finding specific value of ν within V

- $p(w|y,x)$

$\min KL(q||p)$

"Kullback Leibler"

$$KL(q\|p) \triangleq \mathbb{E}_{q(w)}\left[\log \frac{q(w)}{p(w|x,y)}\right]$$

(→ textbook says →)

what is KL if $q = p$ everywhere? exactly 0

what if $q \neq p$ everywhere? $KL(q\|p) > 0$

Properties:

not symmetric

useful to measure "distance" between q & p

Motivated by information thy

measures expected num. "bits" (if $\log_2$)

required to code samples from q using best possible code for p

Look at expanding KL

$$KL \triangleq \mathbb{E}_{q(w)}\left[\log \frac{q(w)}{p(w|x,y)}\right]$$

Bayes thm: $p(w|x,y) = \frac{p(w)p(y|w,x)}{p(y|x)}$

$$= \mathbb{E}_{q(w)}\left[\log q(w) - \log \frac{p(w)p(y|w,x)}{p(y|x)}\right]$$

$$= \mathbb{E}_{q(w)}\left[\log q(w) - \log p(w)p(y|w,x)\right] + \log p(y|x)$$

(marginal lik; constant)

want to do $\min_q KL(q\|p(w|y,x))$

can instead $\min_q \mathbb{E}_q\left[\log q(w) - \log p(w) - \log p(y|w,x)\right]$

$\max_{v \in V} \mathcal{L}(v)$

Evidence lower bound objective

$$\mathcal{L}(v) = \mathbb{E}_{q_{(w|v)}}[\log p(w) + \log p(y|w,x) - \log q(w)]$$

Two interps

$$\mathbb{E}_{q(w)}[\log p(y,w|x)] + \mathbb{E}_q[-\log q(w)]$$

joint prob. — entropy

$$\mathbb{E}_{q(w)}[\log p(y|x,w)] + \mathbb{E}_q\left[\log \frac{p(w)}{q(w)}\right]$$

reconstruction likelihood — KL to prior

Two choices:

(1) Density family for q w/ V space → if $\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} \in \mathbb{R}^2$ could be

(2) Specific parameters $v \in V$

simple ↕ flexible

$N\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, I\right)$

$N\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix}\right)$

$N\left(\begin{bmatrix} m_1 \\ m_2 \end{bmatrix}, \Sigma\right)$

want to balance simplicity & flexibility in making approximations

How can we compute $\mathcal{L}(v)$?

Monte Carlo approximation!

$$\mathbb{E}_{q(w)}\left[\log p(y|w,x) + \log p(w) - \log q(w)\right]$$

$$\approx \frac{1}{S}\sum_{s=1}^{S} \log p(y|x,w^s) + \log p(w^s) - \log q(w^s)$$

when $w^s \overset{iid}{\sim} q(w^s|v)$

works for any model (lik & prior pdf choice) as long as we can

(1) sample from q

(2) evaluate logpdf of q at w

How can we find the variational parameter v that maximizes $\mathcal{L}$?

If we could estimate gradients, just go uphill

how to estimate gradients?

can't use samples w^s from MC approx.

Once we've drawn, we've lost dependence on v

$$\frac{1}{S}\sum_{s} \nabla_v\left[\log p(y|x,w^s) + \ldots\right]$$

doesn't work!

Idea: use small identity that's true for any distrib.

Given $q(w|v)$ w/ continuous param. v

$$\nabla_v \log q(w|v) = \frac{1}{q(w|v)} \nabla_v [q(w|v)]$$

by chain rule of log function

How does this help?

$$\nabla_v \mathbb{E}_{q(w|v)}[\log q(w|v)] = \mathbb{E}_{q(w|v)}[\nabla_v \log q(w|v)]$$

$$= \mathbb{E}_{q(w|v)}\left[\frac{1}{q(w|v)} \nabla_v q(w|v)\right]$$

$$= \int \frac{q(w|v)}{q(w|v)} \nabla_v q(w|v)\, dw$$

$$= \nabla_v \int q(w|v)\, dw$$

$$= \nabla_v [1] = 0$$

this "trick" lets us simplify expectations of log probs.

Now for ELBO objective $\mathcal{L}(v)$

$$\nabla_v \mathbb{E}_q \left[\log p(y,w|x) - \log q(w|v) \right]$$

$$\nabla_v \int q(w) \left[\log p(y,w|x) - \log q(w|v) \right] dw$$

$$\int \nabla_v \left[\underbrace{q(w)}_{A} \left[\log p(y,w|x) - \log q(w|v) \right]_{B} \right] dw$$

product rule says

$$\int \boxed{\text{grad } A * B} \left[\nabla_v q(w) \right] \left(\log p(y,w|x) - \log q(w|v) \right) dw$$

$$+ \int \nabla_v \left[\log p(y,w|x) - \log q(w|v) \right] q(w)\, dw$$

$\log p(y,w|x)$: const wrt v, so grad $= 0$

$\log q(w|v)$: we know gradient is 0

$$\boxed{\text{grad } B * A}$$

→ whole term equals zero!

so keep first term & apply trick again.

$$\nabla_v \mathcal{L}(v) = \int \left[\nabla_v q(w) \right] \left(\log p(y,w|x) - \log q(w|v) \right) dw$$

$$= \int \left[\nabla_v \log q(w|v) \right] q(w) \left(\quad // \quad \right) dw$$

(LD trick)

(expectation!)

$$= \text{Something we can estimate via Monte Carlo!}$$

$$\frac{1}{S} \sum_{s=1}^{S} \left[\nabla_v \log q(w^s|v) \right] \left(\log p(y,w^s|x) - \log q(w^s|v) \right)$$