CS281 Section #8 Notes, v 1.1

1 The Evidence Lower Bound (ELBO)

Our goal in variational inference is to approximate the posterior distribution over some latent variables z, given a prior $p(z)$ and a generative model $p(x|z)$. As we mentioned in the previous section, the exact posterior $p(z|x)$ is intractable, because

$$p(z|x) = \frac{p(z,x)}{p(x)}$$

which requires the intractable value $p(x)$. Instead, we choose a new distribution q, which is called the *variational approximation*, to approximate $p(z|x)$. We choose the optimal q^* from a family Q of distributions such that

$$q^*(z) = \underset{q(z)\in Q}{\arg\min}\, \mathrm{KL}(q(z)||p(z|x))$$

but how do we know if something is close to $p(z|x)$ if we don't know $p(z|x)$?

$$\begin{aligned}\mathrm{KL}(q(z)||p(z|x)) &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z|x)] \\ &= \mathbb{E}_q[\log q(z)] - \mathbb{E}_q[\log p(z,x)] + \log p(x)\end{aligned}$$

where we note that $\mathbb{E}_q[\log p(x)] = \log p(x)$, since it's a constant with respect to our choice of q. We want to minimize this quantity with respect to q, but since $\log p(x)$ doesn't matter in the optimization, it suffices to maximize:

$$\mathrm{ELBO}(q) = \mathbb{E}_q[\log p(z,x)] - \mathbb{E}_q[\log q(z)]$$

This is called the **evidence lower bound** (ELBO). It's the negative of the KL divergence above, plus $\log p(x)$. Maximizing the ELBO is equivalent to minimizing the divergence. We can also rewrite it as:

$$\begin{aligned}\mathrm{ELBO}(q) &= \mathbb{E}_q[\log p(z)] + \mathbb{E}_q[\log p(x|z)] - \mathbb{E}_q[\log q(z)] \qquad (1)\\ &= \mathbb{E}_q[\log p(x|z)] - \mathrm{KL}(q(z)||p(z)) \qquad (2)\end{aligned}$$

where the first term is the expected log likelihood, and the second term is the KL divergence between the prior over z, and the variational posterior.

We call the ELBO the evidence lower bound because

$$\mathrm{ELBO}(q) \le \mathrm{ELBO}(q) + \mathrm{KL}(q(z)||p(z|x)) = \log p(x)$$

where the first inequality comes from the fact that KL divergences are positive, which we get from Jensen's inequality.

Exercise Which values of z will this objective encourage $q(z)$ to place its mass on?

Exercise Why don't we try to minimize $\mathrm{KL}(p(z|x)||q(z))$ ("the forward KL") instead? If we could minimize it, what kind of properties would the learned q distribution have instead?

2 Gibbs Sampling

3 Review

Gibbs sampling is the first MCMC algorithm we've seen in this class. The goal is to get samples from a distribution $p(\mathbf{x})$, for some purpose (for example, we can use the samples to calculate $\mathbb{E}_p[f(\mathbf{x})]$). In **Markov**

Chain Monte Carlo (MCMC), we construct a Markov chain $\mathbf{x}_t$ such that the *stationary distribution* of the Markov chain is $p(\mathbf{x})$. Note that the samples $\mathbf{x}_i$ are correlated with each other! As long as they are unbiased, we still get an estimate of $\mathbb{E}_p[f(\mathbf{x})]$ from it. After a long time, we expect the chain to converge to $p(\mathbf{x})$, and we can use our samples.

For Gibbs sampling in particular, we want to sample this chain by sampling each element x_i from the conditional distribution given every other element: $x_i \sim p(x_i|\mathbf{x}_{-i})$. The proof that the stationary distribution of a chain sampled this way is $p(\mathbf{x})$ is not covered in this course.

When to use MCMC? It's a slow method, so in smaller models where the dimensionality is low, it might be easier to get samples via rejection sampling or importance sampling. It is used primarily for high dimensional problems. Depending on what information you have, variational inference might be better.

4 Example/ Walkthrough

Suppose $y_i \sim \mathcal{N}(\mu, \sigma^2)$ are some data we observe, and we'd like to sample from the posterior distribution of (μ, σ^2).

To simplify our sampling, we should choose conjugate prior distributions for μ, σ^2.

$$\begin{aligned} \mu &\sim \mathcal{N}(\mu_0, \sigma_0^2) \\ \sigma^2 &\sim \Gamma^{-1}(\alpha, \beta) \end{aligned}$$

(If we didn't specify a conjugate prior, sampling will have to be done with a Metropolis step. More details coming in MCMC section.)

Full posterior:

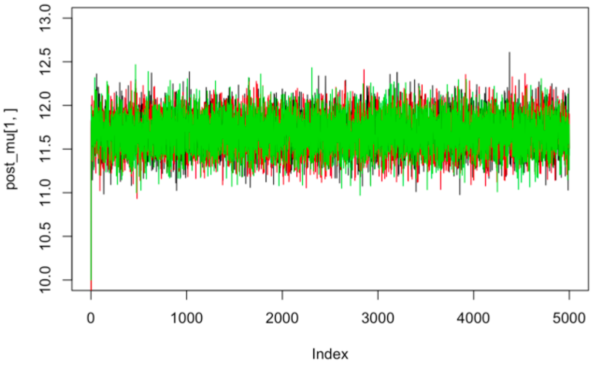
$$p(\mu, \sigma^2|y) \propto p(y|\mu, \sigma^2)p(\mu, \sigma^2)$$

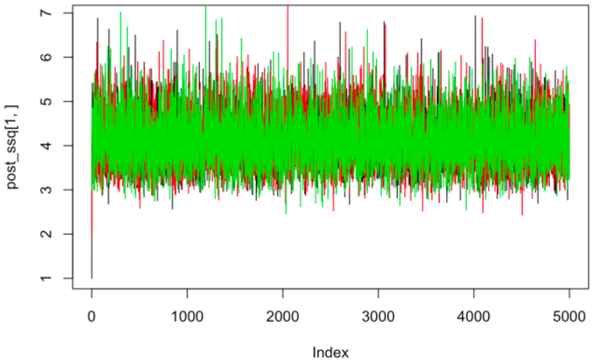
Calculating conditional posterior distributions gives:

$$\begin{aligned} \mu|y, \sigma^2 &\sim \mathcal{N}(\mu_p, \sigma_p^2) \\ \sigma_p^2 &= \Big(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}\Big)^{-1} \\ \mu_p &+ \sigma_p^2\Big(\frac{\sum y_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\Big) \\ \sigma^2|y, \mu &\sim \Gamma^{-1}(\alpha_p, \beta_p) \\ \alpha_p &= \alpha + \frac{n}{2} \\ \beta_p &= \beta + \frac{1}{2}\sum(y_i - \mu)^2 \end{aligned}$$

Now that we have the full conditionals, we can implement Gibbs sampling. We choose starting values $\mu^{(0)}, \sigma^{2(0)}$, and iterate sampling over $\mu|y, \sigma^2$ and $\sigma^2|y, \mu$ with the most current updates of μ, σ^2.

For demonstration, let $y_i \sim \mathcal{N}(12, 4)$ with priors $\mu \sim \mathcal{N}(0, 10000), \sigma^2 \sim \Gamma^{-1}(0.001, 0.001)$. Below are the results obtained from running 3 chains with 5000 iterations each. Initiate chains with different starting values: $\mu^{(0)} = (-10, 0, 10), \sigma^{2(0)} = (1, 2, 3)$





5 Loopy BP

While we know belief propagation on trees is an exact method that converges after just 2 iterations, we can run it on graphs with cycles and it will sometimes work:

Loopy BP for pairwise MRF:

- Initialize all the messages $m_{s \to t}(x_t) = 1$ for all edges, and the beliefs $\text{bel}_s(x_s) = 1$ for all nodes.
- Repeat:
  - Update all the messages
  - Update all the beliefs

  until beliefs don't change significantly

(see Murphy, Algorithm 22.1 for the exact steps).

6 References

1. CS281 Lectures on Info Theory and Mixture Models (2017), Sasha Rush
2. Bayesian Mixture Models and the Gibbs Sampler (2015), David M. Blei*
3. Variational Inference: A Review for Statisticians (2017), David M. Blei, Alp Kucukelbir, Jon D. McAuliffe*

* These notes borrow heavily from David Blei's tutorials. They are very good resources.