KL and Variational Inference

This lecture introduces KL, ELBO, General EM, Factorized Inference with Mean Field Approximation, VI application for GMM

KL Divergence

Entropy is given by:

$$H(p) = -\sum_{x \sim p(x)} p(x) log(p(x))$$

KL divergence between two distributions p and q measures the additional information to describe/decode one distribution given another's coding scheme:

$$D_{KL}(P||Q) = E_{x \sim P}[log \frac{P(x)}{Q(x)}]$$

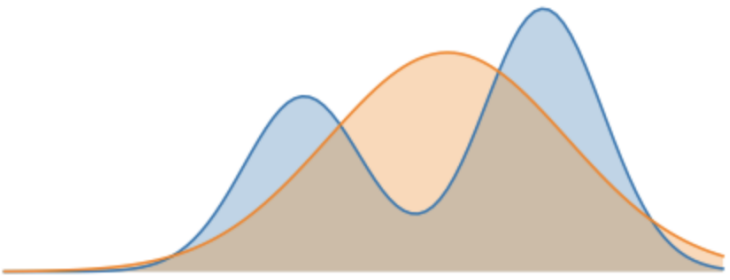
We can rewrite this in terms of entropy and cross entropy:

$$D_{KL}(P||Q) = E_{x \sim P}[-\log Q(x)] - H(P(x))$$

Depending on the ordering of P and Q in calculation, we would get **forward KL** or **reverse KL**:

1. Forward KL: $argmin_\theta D_{KL}(P|Q_\theta)$
   This brings us to the MLE estimate

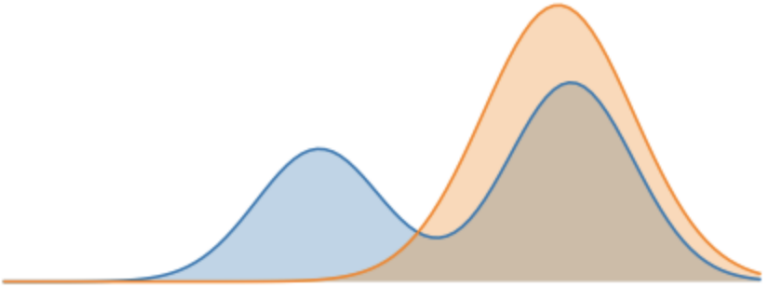
   We consider this *mean-seeking* behaviour, because the approximate distribution Q must cover all the modes and regions of high probability in P. The optimal "approximate" distribution for our example is shown below. Notice that the approximate distribution centers itself between the two modes, so that it can have high coverage of both. The forward KL divergence does not penalize Q for having high probability mass where P does not.



2. . Reverse KL: $argmin_\theta D_{KL}(Q_\theta|P)$

lie within a mode of P (since it's required that samples from Q have high probability under P). Notice that unlike the forward KL objective, there's nothing requiring the approximate distribution to try to cover all the modes. The entropy term prevents the approximate distribution from collapsing to a very narrow mode; typically, behaviour when optimizing this objective is to find a mode of P with high probability and wide support, and mimic it exactly.

The optimal "approximate" distribution for our example is shown below. Notice that the approximate distribution essentially encompasses the right mode of P. The reverse KL divergence does not penalize Q for not placing probability mass on the other mode of P.



Generalized EM

Rewrite the marginal likelihood in terms of observed and latent variables:

$$p(x|\theta) = \sum_z p(x, z|\theta)$$

Remember, the joint conditional distribution is given by the following:

$$p(x, z|\theta) = p(x|\theta)p(z|x, \theta)$$

Then marginal likelihood can be written as (after multiplying both nominator and denominator with q(z)):

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

Taking log on both sides and integrating over z, we obtain:

$$\log p(x|\theta) = \sum_z q(z) \log \frac{p(x, z|\theta)}{p(z|x, \theta)} + \sum_z q(z) \log \frac{q(z)}{p(z|x, \theta)}$$

which is equal to

$$\lambda(q, \theta) + D_{KL}(q||p)$$

EM :
E-step: fix θ, maximize $\lambda(q, \theta)$ which leads to minimization of D_{KL}. When D_{KL} is minimal (when $q(z) = p(z|x, \theta)$), lower bound is maximized. Log-likelihood value is equal to lower-bound in this case because we set KL to be zero.

M-step: $q(z)$ is fixed, maximize $\lambda(q, \theta)$ over θ to obtain θ^{new} that is better than $\lambda(q, \theta^{old})$

This causes the lower bound to increase which in turn means the log-likelihood must also increase. So, the increase in log-likelihood must be greater than the increase in the lower bound.

In each step, both lower bound and likelihood gets better