On EM, IEEE, numerics, etc

Min Seok Song

Definition 1. *Kullback-Leibler(KL) divergence between p.d.f.s g and f is given by*

$$d_{KL}(g\|f) = E_g[\log(\frac{g}{f})]$$

This is always nonnegative and it can be shown by Jensen's inequality. Intuitively, we have more 'confidence' on g whenever g is greater than f, whence the logarithm is positive.

Definition 2. *The ELBO (Evidence Lower-Bound) of a p.d.f. g with respect to an unnormalized p.d.f. $\tilde{f}$ is given by*

$$ELBO(g) := E_g[\log \frac{\tilde{f}}{g}]$$

Remark 1.
- Simple algebra yields $ELBO(g) \leqslant \log c$, where c is a normalizing constant.
- Let's think in Bayesian framework; our unnormalized function in this case is $\tilde{f}(\theta) := f(y|\theta)f(\theta)$, so $ELBO(g) \leqslant \log f(y)$.
- This justifies the name "evidence lower-bound" and this helps with the choice of modeling (essentially maximizing ELBO).
- We write ELBO(g), but really, what's omitted is that this is with respect to $\tilde{f}(\theta)$.

Remark 2.
- It follows that $ELBO(g) = E_g[\log f(y|\theta)] - d_{KL}(g(\theta)\|f(\theta))$
- In another Bayesian framework, similar equality (will be used in EM algorithm) is

$$\begin{aligned}
&ELBO(g, \theta)\\
&= \int \log(\frac{f(y, z|\theta)}{g(z)})g(z)dz\\
&= -d_{KL}(g(z)\|f(z|y, \theta)) + \log(f(y|\theta))
\end{aligned}$$

The first term promotes matching the data, and the second term promotes matching prior beliefs. The reason we work with log domain is that, except for obvious reasons, it helps with numerical stability.
Indeed, remark 2 motivates what's called EM algorithm.

Algorithm 1 EM Algorithm

Input: Initialization of θ_0

repeat

E-step: compute

$$E_{Z\sim f(z|y,\theta_l)}[\log f(y,Z \mid \theta)] = \int \log f(y,z \mid \theta) f(z \mid y,\theta_l) dz$$

M-step: compute

$$\theta_{l+1} = \arg\max_\theta E_{Z\sim f(z|y,\theta_l)}[\log f(y,Z \mid \theta)]$$

until some stopping criterion

Theorem 1. *We have* $\log f(y|\theta_l) \leqslant \log f(y|\theta l+1)$

Proof. Let $g_l(z) = f(z|y,\theta_l)$. We have

$$\begin{aligned}\log f(y|\theta_{l+1}) &= ELBO(g_l,\theta_{l+1}) + d_{KL}(g_l|f(z|y,\theta_{l+1})) \\ &\geqslant ELBO(g_l,\theta_l) + d_{KL}(g_l|f(z|y,\theta_l)) \\ &= \log f(y|\theta_l)\end{aligned}$$

□

- This shows that likelihood function is non-decreasing for each iteration, and since likelihood is always bounded by 1, we have established the convergence of the algorithm.
- GMM algorithm is a specific instance of this algorithm. Here, w_{ik} can be thought of as latent variable, corresponding to E-step, and computing θ and π corresponds to M-step.
- E-step can be computationally expensive and so we usually use Monte-Carlo method to approximate.

Remark 3.

- Let us now venture into variational inference, which we use when approximating the intractable distribution. For the choice of possible sets that f admits in $d_{KL}(g\|f)$, we can use **mean field family**, which assumes the independence for coordinate distributions.

Theorem 2. *Let* $g(x) = \prod_{i=1}^d g_i(x_i)$ *with* $g_{-i}(x_{-i})$ *fixed. Then*

$$g_i^*(x_i) \propto \exp(E_{g_{-i}}[\log f(x_i, x_{-i}]))$$

maximizes $ELBO(g)$.

Proof. By routine algebra (we need to use independence at some point), one can show that

$$ELBO(g) = E_{g_i}[\log(\exp(E_{g_{-i}}[\log f(x_i, x_{-i})])) - \log g_i(x_i)] + C$$

and notice that the first term can be phrased as -$d_{KL}(g^* \| g)$, whence $g^* = g$ gives the optimization solution. □

Remark 4.
- The intuition is to average out the effect of x_{-i} on log expected value in order to incorporate the independence between g_i's.
- This leads to the CAVI algorithm, which approximates the unnormalized target density $\tilde{f}$. After initialization, updating g_i will increase ELBO for each i, so may iterate until ELBO converges.
- The disadvantage is that it may be computationally expensive and accuracy might be not so good.