PGM - Variational inference

[Pres → Detail #1]

Detail #2: The objective is to minimize

$$KL(q \,\|\, p(\cdot|y)) = \mathbb{E}_q\left[\log \frac{q(x)}{\underbrace{p(x|y)}_{= \frac{p(x,y)}{p(y)}}}\right]$$

$$= \underbrace{\mathbb{E}_q[\log q(x)] - \mathbb{E}_q[\log p(x,y)]}_{\text{minimizing KL is the same as minimizing this expression, or equivalently maximizing}} + \underbrace{\log p(y)}_{\text{constant (wrt } q)}$$

$$ELBO(q) \overset{\text{def}}{=} \underbrace{\mathbb{E}_q[\log p(x,y)]}_{\text{encourages } q \text{ to concentrate at the mode of } p(x|y)} \underbrace{- \mathbb{E}_q[\log q(x)]}_{= \text{ entropy of } q, \text{ encourages } q \text{ to "spread out"}}$$

ELBO = evidence lower bound

$$\log p(y) = \underbrace{KL(q \,\|\, p(\cdot|y))}_{\geq 0} + ELBO(q) \Rightarrow \log p(y) \geq ELBO(q)$$

$p(y)$ is often referred to as model evidence (or marginal likelihood).

[Pres → Detail #4, CAVI]

Consider updating the kth component $q_k(x_k)$ with the other components $q_{-k}(\mathbb{X}_{-k}) \stackrel{\text{def}}{=} \prod_{j \neq k} q_j(x_j)$ being fixed.

$$\mathrm{ELBO}(q_k; q_{-k}) = \mathbb{E}_{q_k}\left[\mathbb{E}_{q_{-k}}\left[\log p(x_k, \mathbb{X}_{-k}, \mathbb{Y})\right]\right]$$

$$- \mathbb{E}_{q_k}\left[\log q_k\right] + \text{const}$$

Define $$q_k^*(x_k) \propto \exp\left(\mathbb{E}_{q_{-k}}\left[\log p(x_k, \mathbb{X}_{-k}, \mathbb{Y})\right]\right)$$
$$\propto \exp\left(\mathbb{E}_{q_{-k}}\left[\log p(x_k | \mathbb{X}_{-k}, \mathbb{Y})\right]\right)$$

Then $\mathrm{ELBO}(q_k; q_{-k}) = -\mathrm{KL}(q_k, q_k^*) + \text{const}$

which is maximized by setting $q_k^{new} = q_k^*$

[Pres → "What does", reveal $\prod_{k=1}^{K} \theta_{d,k}^{x_k + c_{d,k}}$]

We get

$$\log p(\theta_d \mid \beta, z, w)$$

$$= \sum_{k=1}^{K} (\alpha_k + c_{d,k} - 1) \log \theta_{d,k} + \text{const}$$

$$\Rightarrow \mathbb{E}_{q(-\theta_d)}\left[\log p(\theta_d \mid \beta, z, w)\right]$$

$$= \sum_{k=1}^{K} \left\{ \alpha_k + \mathbb{E}_q[c_{d,k}] - 1 \right\} \log \theta_{d,k} + \text{const}$$

Consequently, $q^*(\theta_d) \propto \exp(\cdot) = \prod_{k=1}^{K} \theta_{d,k}^{\alpha_k + \mathbb{E}_q[c_{d,k}] - 1}$

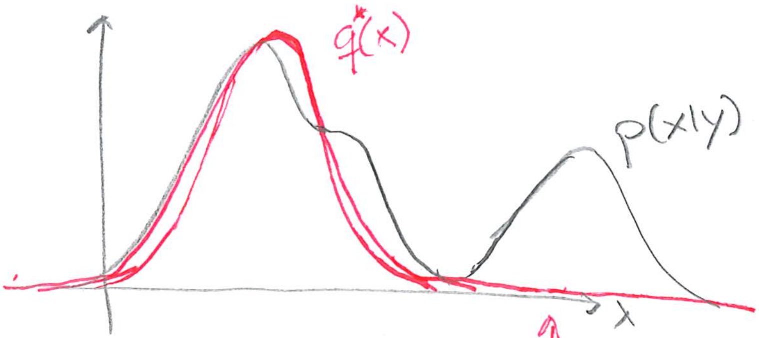
$$\propto \mathrm{Dir}(\theta_d \mid \alpha + \mathbb{E}_q[c_d])$$

[Pros → Limitations]

"Mode seeking" behavior of

$$KL(q \| p(\cdot|y)) = \mathbb{E}_q \left[\log \frac{q(x)}{p(x|y)} \right]$$

we "average the discrepancy" wrt the approximation itself.



We don't care as much about the discrepancy here, where q is low

This typically leads to underestimation of the posterior variance, which VI is infamous for.