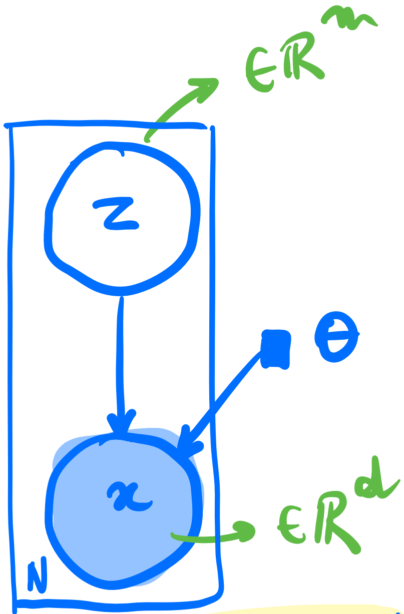
LEC06 Expectation - Maximization

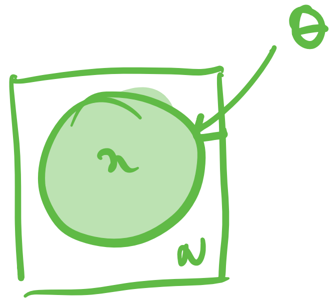


+ GALTON!

$p(z)$

$p_\theta(x|z)$

Start with



How to estimate θ?

How to estimate θ?

$\max p(x_{1:n}) = \prod p(x_i) \iff \sum \log p(x_i)$

$$\max_\theta p_\theta(x) = \int p_\theta(x,z)\,dz$$

$$= \int p(z)\, p_\theta(x|z)\,dz$$

$$= \mathbb{E}_{p(z)}\left[p_\theta(x|z)\right]$$

$$\approx \frac{1}{B}\sum_b p_\theta(x|z_b)$$

MC approximation

⚠

Poor approximation when n is large due to the curse of dimensionality.

I. Expectation - Maximization

Check derivation in DL

$$p_\theta(x) = \mathbb{E}_{p(z)}\left[p_\theta(x|z)\right]$$

IS

$$= \mathbb{E}_{q(z)}\left[\frac{p(z)}{q(z)}\, p_\theta(x|z)\right]$$

if we hold N x_n

$\Rightarrow \prod p(x_n)$

$\rightarrow \sum \log p(x_n)$

$$\log p_\theta(x) = \log \mathbb{E}_{q(z)}\left[\frac{p(z)}{q(z)}\, p_\theta(x|z)\right]$$

Jensen

$$\geq \mathbb{E}_{q(z)}\left[\log \frac{p(z)}{q(z)}\, p_\theta(x|z)\right]$$

$$= \mathbb{E}_{q(z)}\left[\log \frac{p_\theta(x,z)}{q(z)}\right]$$

① ②

ELBO

① $$= \mathbb{E}_{q(z)}\left[\log \frac{p_\theta(x,z)}{q(z)}\right]$$

$$= \mathbb{E}_{q(z)}\left[\log p_\theta(x|z)\right] - KL\left(q(z)\,\|\,p(z)\right)$$

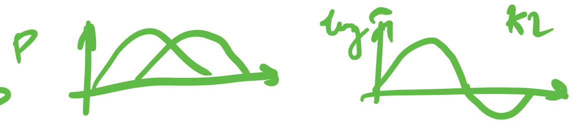
ELBO(q, θ)

or negative of the variational free energy

② $$= \mathbb{E}_{q(z)}\left[\log \frac{p(z)}{q(z)}\, p_\theta(x|z)\, \frac{p_\theta(x)}{p_\theta(x)}\right]$$

$$= \mathbb{E}_{q(z)}\left[\log \frac{p_\theta(z|x)}{q(z)}\, p_\theta(x)\right]$$

$$= \log p_\theta(x) - KL\left(q(z)\,\|\,p_\theta(z|x)\right)$$



$$\boxed{KL(p\,\|\,q) = \mathbb{E}_p\left[\log \frac{p}{q}\right]}$$

B... ... gained for

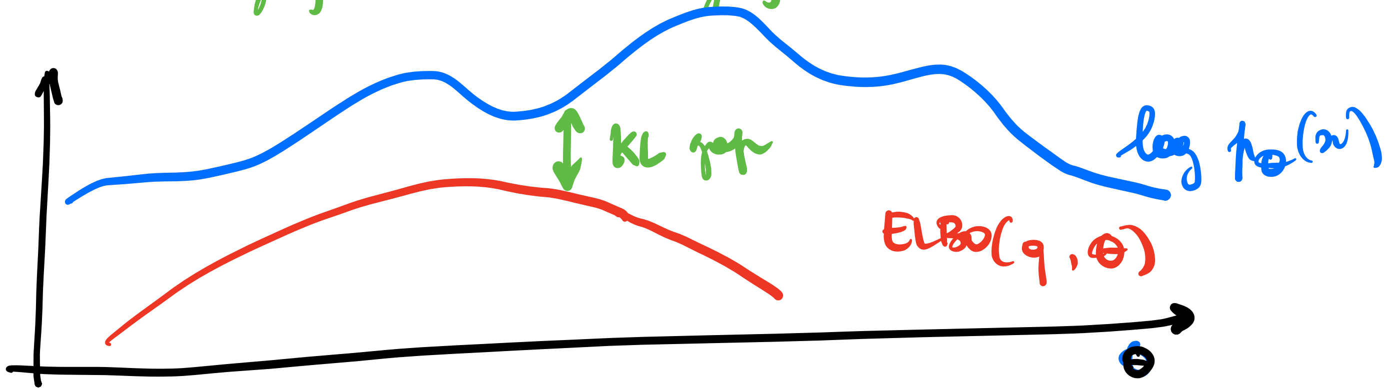
Bayesian: information gained by
revising one's prior beliefs from Q to P

①+② $\Rightarrow \log p_\theta(x) = \mathrm{ELBO}(q, \theta) + KL\left(q(z) \,\|\, p_\theta(z|x)\right)$

Since $KL \geq 0$,

$$\log p_\theta(x) \geq \mathrm{ELBO}(q, \theta)$$

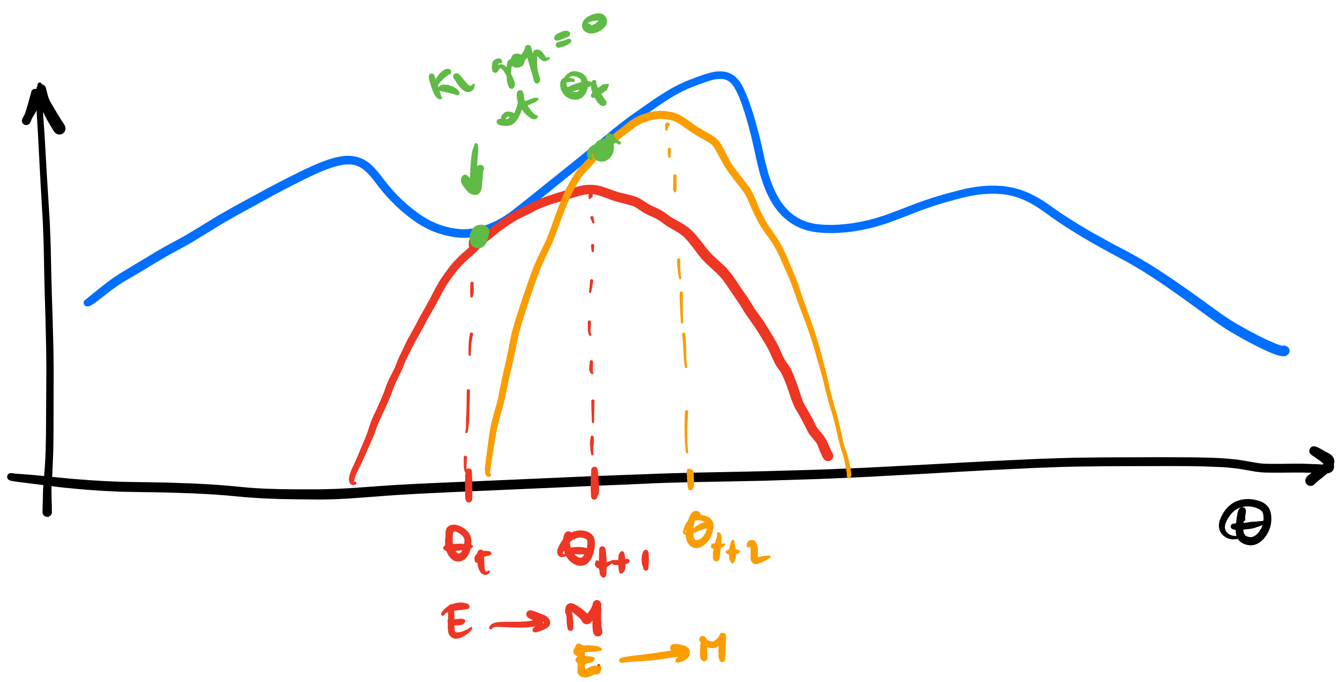
maximizing $\Leftrightarrow$ maximizing



How to pick q? Assume a candidate θ_t,
make the bound tight and set
$(KL = 0 \Rightarrow q(z) = p_{\theta_t}(z|x))$

$$\boxed{q(z) := p_{\theta_t}(z|x)}$$

Assuming the posterior can be derived analytically.



Hence,

$$\mathrm{ELBO}(q,\theta) = \mathbb{E}_{q(z)}\left[\log \frac{p_\theta(x,z)}{q(z)}\right]$$

$$= \mathbb{E}_{p_{\theta_t}(z|x)}\left[\log p_\theta(x,z)\right] - \mathbb{E}_{p_{\theta_t}(z|x)}\left[\log p_{\theta_t}(z|x)\right]$$

$$= Q(\theta,\theta_t) + \mathbb{H}\left[p_{\theta_t}(z|x)\right]$$

EM: $\theta_0 = \mathrm{rand}()$

① E-step : Define $Q(\theta,\theta_t) = \mathbb{E}_{p_{\theta_t}(z|x)}\left[\log p_\theta(x,z)\right]$

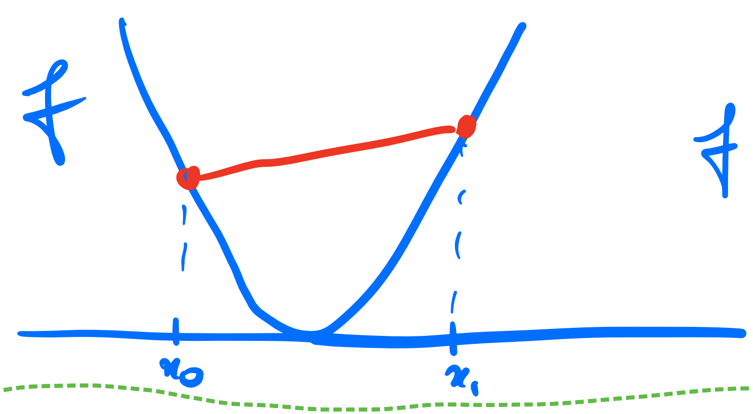
Because $\mathbb{H}$ does not depend on θ.

② M-step : $\theta_{t+1} := \arg\max_\theta Q(\theta,\theta_t)$ $(+\log p(\theta))$ (for MAP estimation)

Jensen's inequality

if f is convex then $f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$

concave $f(\mathbb{E}[x]) \geq \mathbb{E}[f(x)]$



$$f\left(\sum \lambda_i x_i\right) \leq \sum \lambda_i f(x_i)$$

$\mathbb{E}_{p(x)}[\cdot] \Leftrightarrow \sum_{x_i} p(x_i)$

II. Variational inference

What if $p_\theta(z|x)$ is not tractable ?

Replace q with a variational family $\underline{q_\phi(z)}$

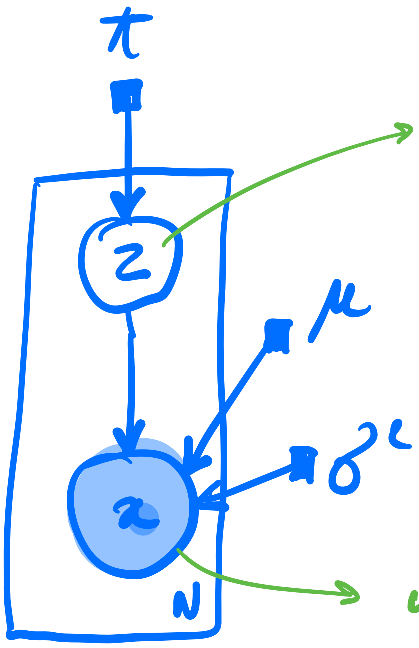
and solve

$$\boxed{\begin{aligned} &\text{VI :} \\ &\underbrace{\theta^*, \phi^*}_{\text{both together!}} = \arg\max_{\theta,\phi} \text{ELBO}(q_\phi, \theta) \\ &\qquad = \arg\max_{\theta,\phi} \mathbb{E}_{q_\phi(z)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z)}\right] \end{aligned}}$$

→ Coordinate ascent VI
Stochastic VI

⚠ EM and VI provide algorithms for both fitting θ (to $p_\theta(x)$) and estimating the posterior $p_{\theta^*}(z|x)$.

Notebook example



$\text{multinomial}(\pi) \hookrightarrow \{1..K\}$

$\rightarrow \mathbb{R}^K$

$\rightarrow \mathbb{R}^{+K}$

$\rightarrow \mathcal{N}(x_i \mid \mu_{z_i}, \sigma^2_{z_i})$

- incomplete data likelihood:

$$\mathcal{L}(\pi, \mu, \sigma^2, x) = \prod_{x_i} \sum_k \pi_k \, \mathcal{N}(x_i \mid \mu_k, \sigma_k^2)$$

- complete data likelihood

$$\mathcal{L}(\pi, \mu, \sigma^2, x, z) = \prod_{x_i} \prod_k \left[\pi_k \, \mathcal{N}(x_i \mid \mu_k, \sigma_k^2)\right]^{1(z_i = k)}$$

E - step

$$p(z = k \mid x_i) = \frac{\pi_k \mathcal{N}(x_i \mid \mu_k, \sigma_k^2)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_i \mid \mu_{k'}, \sigma_{k'}^2)} := r$$

$$= \frac{p(x_i \mid z)\, p(z)}{p(x_i)}$$

M-step

$$\mathbb{E}_z \left[p(x, z \mid \Theta) \right]$$

⇝ Principled alternative to K-Means

(K-means ≈ MLE for z ⟹ provides a partition)