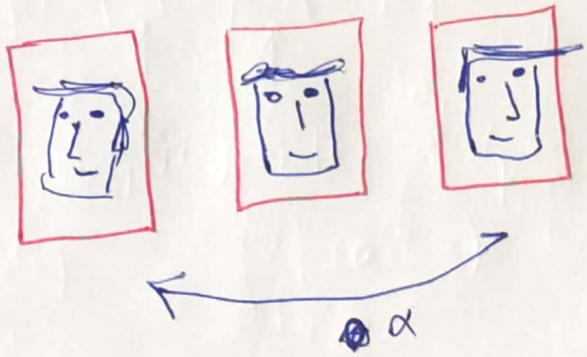
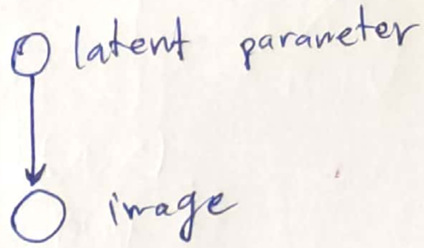
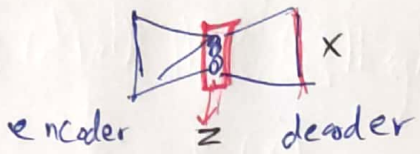
Variational Learning for latent variable models







$\boxed{p_\theta(x)}$

---

model $\quad p_\theta(x) = \sum_z p_\theta(x,z)$ latent variable model

Data $= \{x^1, x^2, \dots, x^m\} = D$

$$\max_\theta \; ll(\theta, D) = \sum_{i=1}^{m} \log p_\theta(x^i) = \sum_{i=1}^{m} \log \sum_z p_\theta(x^i, z)$$

Usually $p_\theta(x,z) = \boxed{p_\theta(x|z)}\, p(z)$

$\boxed{\text{or } \; p_{\theta_1}(x|z)\, p_{\theta_2}(z)}$

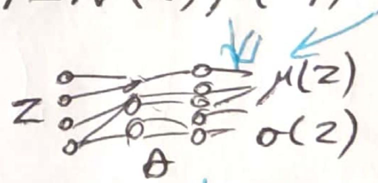
Example

$p_\theta(x,z) = p_\theta(x|z)\, p(z) \qquad p_\theta(x|z) \triangleq N(x; \mu(z), I\sigma(z))$

1 $z \sim p(z)$

2 compute $\boxed{\mu(z), \sigma(z)}$

3 $x \sim p_\theta(x|z) = N(x; \mu(z), I\sigma(z))$



a deep Neural Net

How to train the neural network

Need to comput $\frac{\partial}{\partial \theta_i} \sum_{i=1}^{m} \log \sum_z p_\theta(x^i, z)$

Intractible

$= \frac{\partial}{\partial \theta_i} \sum_{i=1}^{m} \log \sum_z p_\theta(x^i | z)\, p(z)$

for EM Need $p_\theta(z|x) \to$ posterior $\to$ intractable!

$$P_\theta(y) = \frac{1}{Z(\theta)} \tilde{P}(y)$$ MRF

$$Z(\theta) = \sum_y \tilde{p}(y)$$

minimize over q

$$KL(q \,\|\, P_\theta(y))$$

$$P_\theta(z|x) = \frac{1}{P_\theta(x)} P_\theta(x,z)$$ latend variable model

$$P_\theta(x) = \sum_z P_\theta(x,z)$$

minimize over $q(z|x)$

$$KL(\underbrace{q(z|x)}_{\text{approximate posterior}} \,\|\, P_\theta(z|x))$$

Example: $q(z|x) = \prod q_i(z_i|x)$

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

$$= \sum_z q(z|x) \log \frac{q(z|x)}{\frac{P_\theta(x,z) \to \text{easy}}{P_\theta(x) \to \text{hard}}}$$

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

$$= \sum_z q(z|x) \log \frac{q(z|x)}{P_\theta(x,z)} + \sum_z q(z|x) \overbrace{\log P_\theta(x)}^{\text{independent of } z}$$

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

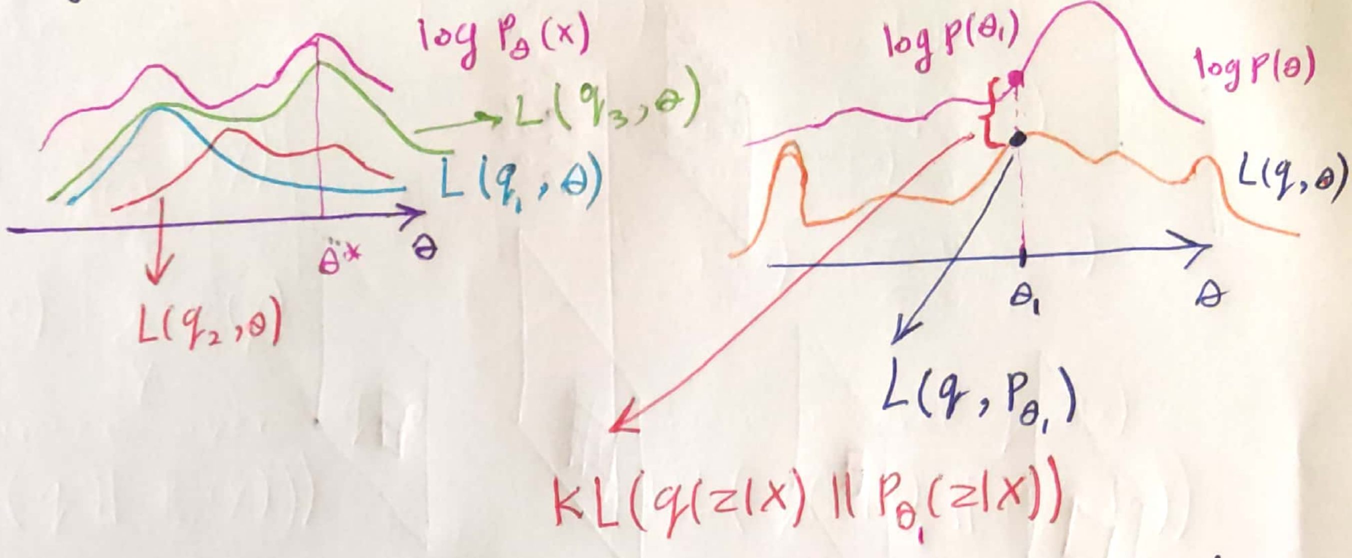
$$\underbrace{\sum_z q(z|x) \log \frac{P_\theta(x,z)}{q(z|x)}}_{\mathcal{L}(q,\, P_\theta)} = \log P_\theta(x) - KL(q(z|x) \,\|\, P_\theta(z|x))$$

$\mathcal{L}(q, P_\theta)$: variational lower bound

Evidence Lower BOund (ELBO)

$\mathcal{L}(q, P_\theta) \leq \log P_\theta(x)$ for any choice of q

single data x



$$L(q, P_\theta) = \log P_\theta(x) - KL(q(z|x) \| P_\theta(z|x))$$

1: Maximizing $L(q, P_\theta)$ with respect to θ pushes $\log P_\theta(x)$ up.

2: Maximizing $L(q, P_\theta)$ w.r.t θ works better when $q(z|x)$ is closer to $P_\theta(z|x)$ [$KL(q(z|x) \| P_\theta(z|x))$ is smaller]

3: Maximizing $L(q, P_\theta)$ w.r.t. q minimizes $KL(q(z|x) \| P_\theta(z|x))$

Variational Learning

$$\max_{q,\theta} \mathcal{L}(q, P_\theta) = \sum_z q(z|x) \log \frac{P_\theta(z,x)}{q(z|x)}$$

How to compute (approximate) $\mathcal{L}(q, P_\theta) = \sum_z \ldots$?

$$\mathcal{L}(q, P_\theta) = E_{q(z|x)} \left\{ \log \frac{P_\theta(z,x)}{q(z|x)} \right\}$$

$$\simeq \frac{1}{P} \sum_{i=1}^{P} \log \frac{P_\theta(z^i, x)}{q(z^i|x)}$$ where $z^1, z^2, \ldots, z^P \sim q(z|x)$

Variational Learning

Data $x^1, x^2, x^3, \dots, x^m$

$$\max_\theta \; ll(\theta) = \sum_{i=1}^m \log P_\theta(x^i) = \sum_{i=1}^m \log \sum_z \log P_\theta(x^i, z)$$

instead

$$\max_{\theta, q} \; \mathcal{L}(q, P_\theta) = \sum_{i=1}^m \sum_Z q_i(z|x^i) \log \frac{P_\theta(x^i, z)}{q_i(z|x^i)}$$

$$= \mathcal{L}(q, \theta) = \sum_{i=1}^m E_{q_i(z|x^i)} \{\log P_\theta(x^i, z)\} + H\{q(z|x^i)\}$$

For each x^i a different q might be optimum.

$\Rightarrow$ A different q_i for each x^i

Variational Learning Algorithm

Start from some $\theta \leftarrow \theta_0, q_1, q_2, \dots, q_m$

optimize w.r.t. θ

$$\frac{\partial}{\partial\theta} L(q, P_\theta) = \sum_{i=1}^m E_q \{ \frac{\partial}{\partial\theta} \log \underbrace{P_\theta(x^i, z)}_{P_\theta(x^i|z)\,p(z)} \}$$

$$= \sum_{i=1}^m E_q \{ \frac{\partial}{\partial\theta} \log P_\theta(x^i|z) \}$$

$$\sim \sum_{i=1}^m \sum_{j=1}^{m'} \frac{\partial}{\partial\theta} \log P_\theta(x^i|z_j^i)$$

$z_1^i, z_2^i, \dots, z_{m'}^i \sim q_i(z|x^i)$

usually $m' = 1$

$$\theta \leftarrow \theta + \lambda \frac{\partial}{\partial\theta} L(q, P_\theta)$$

maximize w.r.t. $q_1, q_2, \dots, q_m$

can do independently for each q_i