a) $KL(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$

Prove $KL(P\|Q)$ is non negative.

a)

$$KL(P\|Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}.$$

$$= -\sum_x P(x) \log \frac{Q(x)}{P(x)}.$$

$$= -E\left[\log \frac{Q}{P}\right]$$

$$\geq -\log\left(E\left[\frac{Q}{P}\right]\right) \quad \{ \text{Jensen's Inequality.}$$

$$= -\log\left(\sum_{x \in X} P(x) \frac{Q(x)}{P(x)}\right)$$

$$= 0.$$

The inequality is introduced due to the application of Jensen's Inequality & the concavity of log.

b) ~~Ans~~ $KL(P(x,y)\|Q(x,y)) = \cancel{E} KL(P(x)\|Q(x)) + \cancel{K}KL(P(y|x)\|Q(y|x))$

Proof

$$KL(P(x,y)\|Q(x,y)) = \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{Q(x,y)}$$

$$= \sum_x \sum_y P(x,y) \log\left[\frac{P(x)\,P(y|x)}{Q(x)\,Q(y|x)}\right] \quad \{ \text{By conditional probability}$$

$$= \cancel{\sum_x \sum_y P(x,y) \log \frac{P(x)}{Q(x)} + \sum_x \sum_y P(x,y) \log \frac{P(y|x)}{Q(x|y)\,Q(y|x)}}$$

$$= \cancel{KL(P(x)\|Q(x))}$$

$$= \sum_x \sum_y P(x,y) \log \frac{P(x)}{Q(x)} + \sum_x \sum_y P(x,y) \log \frac{P(y|x)}{Q(y|x)}.$$

$$= \sum_x \sum_y P(x,y) \log \frac{P(x)}{Q(x)} + \sum_x P(x) \sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)}$$

$$= KL(P(x)\|Q(x)) + KL(P(y|x)\|Q(y|x))$$

Hence proved.

Q For $\hat{P}(x) = 1/m \sum_{i=1}^{m} 1\{x^{(i)} = x\}$, for family of distributions P_θ.

Prove that

$$\arg\min_{\theta} KL(\hat{P} \| P_\theta) = \arg\max_{\theta} \sum_{i=1}^{m} \log P_\theta(x^{(i)})$$

This indicates that finding the maximum likelihood estimate for the parameter θ is equivalent to finding P_θ with minimal KL divergence from $\hat{P}$.

Ans By KL divergence.

$$KL(P \| Q) = \int P(x) \log \frac{P(x)}{Q(x)} dx \quad \text{—①}$$

$\Rightarrow$ let $\hat{P}(x)$ be the emperical distribution.

$$\hat{P}(x) = 1/m \sum_{i=1}^{m} 1\{x^{(i)} = x\}.$$

$$KL[\hat{P}(x) \| P(x|\theta)] = \int \hat{P}(x) \log \frac{\hat{P}(x)}{P(x|\theta)} dx$$

$$= -H(\hat{P}) - \int \hat{P}(x) \log[P(x|\theta)] dx \quad \text{where } H(\hat{P}) = -\int \hat{P}(x) \log \hat{P}(x) dx$$

$\quad$ ②.

From ② it follows that

$$\arg\min_{\theta} KL[\hat{P}(x) \| P(x|\theta)] \| P(x|\theta)] = \arg\max_{\theta} \langle \log P(x|\theta) \rangle_{\hat{P}}$$

where $\langle \dots \rangle_{\hat{P}}$ represents expectation with respect to the distribution of $\hat{P}$.

using ① in ② RHS.

$$\langle \log p(x|\theta) \rangle_{\hat{P}} = \frac{1}{m} \int \sum_{t=1} \hat{P}(x) \log P(x|\theta)\, dx.$$

$$= \frac{1}{m} \sum_{t=1}^{m} \log P(x_t|\theta)$$

Apart from the scaling factor $1/m$, this is a log-likelihood funtion. H.

2) EM for MAP estimation

$$l(\theta) = \sum_{i=1}^{m} \log P(x^{(i)}|\theta) + \log P(\theta)$$

$$= \sum_{i=1}^{m} \log \sum_{z^i} P(x^i, z^i|\theta) + \log P(\theta)$$

$$= \sum_{i=1}^{m} \log \sum_{z} Q_i(z^i) \frac{P(x^i, z^i|\theta)}{Q_i(z^i)} + \log P(\theta).$$

$$>= \sum_{i=1}^{m} \sum_{z^i} Q_i(z^i) \log \frac{P(x^i, z^{(i)}|\theta)}{Q_i(z^i)} + \log P(\theta)$$

The above equality holds when

$$\frac{P(x^i, z^i|\theta)}{Q_i(z^i)} = c \quad \& \text{ since } \sum_z Q_i(z^i) = 1,$$

$$Q_i(z^i) = \frac{P(x^i, z^i|\theta)}{P(x^i|\theta)}$$

EM for MAP is following:

Repeat until convergence

{ (E-step) For each i, $Q_i(z^i) = P(z^i|x^i, \theta)$

(M-step) set

$$\theta = \arg\max_{\theta} \sum_{i=1}^{m} \sum_{z^i} Q_i(z^i) \log \frac{P(x^i, z^i|\theta)}{Q_i(z^i)} + \log P(\theta)$$

}

Prove that $l(\theta) = \sum_{i=1}^{m} \log P(x^i|\theta) + \log P(\theta)$ monotonically increases with each iteration. This is to just prove that $l(\theta^t) <= l(\theta^{t+1})$

Firstly, given $Q_i(z^i) := p(z^i|x^i, \theta)$ the following equality holds.

$$l(\theta^t) = \sum_{i=1}^{m} \sum_{z^i} Q_i(z^i) \log \frac{P(x^i, z^i|\theta^t)}{Q_i(z^i)} + \log P(\theta^t) = L(\theta^t)$$

with respect to the lower bound,

$$L(\theta) = \sum_{i=1}^{m} \sum_{z^i} Q_i(z^i) \log \frac{P(x^i, z^i|\theta)}{Q_i(z^i)} + \log P(\theta).$$

The next iteration is done by explicitly choosing θ^{t+1} w.r.t the posteriori q $q_i^{(t)}$ to maximize $L(\theta)$ which means

$$L(\theta^t) <= L(\theta^{t+1})$$

w.r.t the loss, the equality

$$l(\theta^{t+1}) = L(\theta^{t+1})$$
$$= \sum_{i=1}^{m} \log P(x^i | \theta^{t+1}) + \log P(\theta^{t+1})$$

holds only for

$$q_{t+1}(z^i) = P(z^i | x^i, \theta^{t+1})$$

which means

$$q_{t+1}(z^{(i)}) = P(z^i | x^i, \theta^{(t+1)})$$
$$l(\theta^{t+1}) = L(\theta^{t+1}) >= L(\theta^t) = l(\theta^t)$$