2 KL divergence and Maximum Likelihood

(a) Nonnegativity

Prove the following:

$$\forall P, Q, D_{KL}(P \parallel Q) \geq 0$$

And

$$D_{KL}(P||Q) = 0 \iff P = Q$$

PROOF 1:

$$\begin{aligned}
D_{KL}(P \parallel Q) &= \sum_x P(x) log \frac{P(x)}{Q(x)} \\
&= -\sum_x P(x) log \frac{Q(x)}{P(x)} \\
&>= -log(\sum_x P(x) \frac{Q(x)}{P(x)}) \\
&= -log(\sum_x Q(x)) \\
&= -log(1) = 0
\end{aligned}$$

PROOF 2:

$$D_{KL}(P \parallel Q) = 0 \iff P = Q$$

a. If $P = Q$, $D_{KL}(P \parallel Q) = \sum_{x \in X} P \log \frac{P}{Q} = \sum_{x \in X} P \log(1) = 0$

b. If $D_{KL}(P \parallel Q) = 0$, given $-\log x$ is strictly convex, then

$$\begin{aligned} E[-log(\frac{P}{Q})] &\geq -log(E[\frac{P}{Q}]) = 0 \\ &= log(E[\frac{Q}{P}]) \\ &= log(\sum[P\frac{Q}{P}]) \\ &= log(1) \\ &= 0 \end{aligned}$$

The equality holds iff $\frac{Q}{P}$ is a constant with probability 1, given the fact that both P and Q are pdf, we can only have $P = Q$.

(b) Chain rule for KL divergence

Prove that:

$$D_{KL}(P(X,Y) \parallel Q(X,Y)) = D_{KL}(P(X) \parallel Q(X)) + D_{KL}(P(Y|X) \parallel Q(Y|X))$$

PROOF:

$$\begin{aligned}
D_{KL}(P(X) \parallel Q(X)) + D_{KL}(P(Y|X) \parallel Q(Y|X)) &= \sum_x P(x) \log \frac{P(x)}{Q(x)} + \sum_x P(x)(\sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)}) \\
&= \sum_x P(x) \left(\log \frac{P(x)}{Q(x)} + \sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)} \right) \\
&= \sum_x P(x) \left(\sum_y P(y|x) \log \frac{P(x)}{Q(x)} + \sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)} \right) \\
&= \sum_x P(x) \left(\sum_y P(y|x) \left(\log \frac{P(x)}{Q(x)} + \log \frac{P(y|x)}{Q(y|x)} \right) \right) \\
&= \sum_x \left(\sum_y P(y|x) P(x) (\log \frac{P(x)P(y|x)}{Q(x)Q(y|x} \right) \\
&= \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{Q(x,y)} \\
&= D_{KL}(P(X,Y) \parallel Q(X,Y))
\end{aligned}$$

(c) KL and maximum likelihood

Prove that

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

$$\begin{aligned}
D_{KL}(\hat{P} \parallel P_\theta) &= \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P_\theta(x)} \\
&= -\sum_x \hat{P}(x) \log \frac{P_\theta(x)}{\hat{P}(x)} \\
&= -\sum_x \frac{1}{m} \sum_{i=1}^{m} 1\{x^{(i)} = x\} \log \frac{P_\theta(x)}{\frac{1}{m}\sum_{i=1}^{m} 1\{x^{(i)} = x\}} \\
&= -\frac{1}{m} \sum_{i=1}^{m} \log \frac{P_\theta(x^{(i)})}{\frac{1}{m}\sum_{i=1}^{m}} \\
&= -\frac{1}{m} \sum_{i=1}^{m} \log P_\theta(x^{(i)}) \\
&= -\frac{1}{m} \text{log-likelihood}
\end{aligned}$$

Which implies that

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

.