ross Entropy:

KL divergence:

Coin 1

$$\begin{cases} 1/2 & \text{heads } (p_1) \\ 1/2 & \text{tails } (p_2) \end{cases}$$

Coin 2

$$\begin{cases} p & \text{heads} \\ q & \text{tails} \end{cases}$$

$$\frac{P(\text{observations} \mid \text{real coin})}{P(\;〃\; \mid \text{coin 2})} = \left(\frac{p_1^{N_H}\, p_2^{N_T}}{q_1^{N_H}\, q_2^{N_T}}\right)$$

$$\xrightarrow[\log]{\frac{1}{N}} \log\left(\frac{p_1^{N_H}\, p_2^{N_T}}{q_1^{N_H}\, q_2^{N_T}}\right)^{\frac{1}{N}} = \frac{1}{N}\log p_1^{N_H} + \frac{1}{N}\log p_2^{N_T}$$

$$-\frac{1}{N}\log q_1^{N_H} - \frac{1}{N}\log q_2^{N_T} = p_1 \log p_1 + p_2 \log p_2$$

$$-p_1 \log q_1 - p_2 \log q_2 = p_1 \log\frac{p_1}{q_1} + p_2 \log\frac{p_2}{q_2}$$

$$D_{KL}(P \| Q) = \sum_i P(i) \log\frac{P(i)}{Q(i)}$$

Now cross Entropy Loss:

$$D_{KL}\Big(\overbrace{p^*(y \mid x_i)}^{\text{True class distribution}} \;\Big\|\; \overbrace{p(y \mid x_i;\theta)}^{\text{predicted class distribution}}\Big) = \sum p^*(y \mid x_i) \log\frac{p^*(y \mid x_i)}{p(y \mid x_i;\theta)}$$

$$= \underbrace{\sum_y p^*(y|x_i) \log p^*(y|x_i)}_{\text{Doesn't depend on } \theta} - \sum_y p^*(y|x_i) \log (y|x_i;\theta)$$

$$\Rightarrow \operatorname{argmin} D_{KL}(p^* || p) \equiv \underset{\theta}{\operatorname{argmin}} - \sum_y p^*(y|x_i) \log p(y|x_i;\theta)$$

$$\Downarrow$$

$$\underset{\theta}{\operatorname{argmin}} H(p^*, p)$$