Jensen's Inequality

$$f \text{ convex} \quad \Rightarrow \quad f(\mathbb{E}X) \leq \mathbb{E}(f(X))$$

In other words, if $\lambda_1, \dots, \lambda_n$ are such that $\lambda_i \geq 0$ for $i = 1, \dots, n$ & $\sum_{i=1}^{n} \lambda_i = 1$,

then

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i \cdot f(x_i)$$

Application 1: $H(X) \leq \log |\Omega|$ for any finite sample space Ω, and $X : \Omega \to \mathbb{R}$.

Application 2

Theorem (Shannon Inequality)

For any finite sample spaces Ω_1, Ω_2, and $X: \Omega_1 \to \mathbb{R}$ and $Y: \Omega_2 \to \mathbb{R}$,

$$H(X) \geq H(X|Y)$$

(i.e. conditioning never increases entropy.)

Moreover, $H(X) = H(X|Y)$ iff X and Y are independent [HW2]

Proof

Let $f(a) = a \cdot \ln a$. Then $\frac{df}{da} = a \cdot \frac{1}{a} + \ln a$

$= 1 + \ln(a)$

$\frac{d^2 f}{da^2} = \frac{1}{a} > 0$, $a > 0$.

$\Rightarrow f$ is convex.

Now, $\sum_{i=1}^{n} \lambda_i f(\alpha_i) \geq f\left(\sum_{i=1}^{n} \lambda_i \alpha_i\right)$. for convex weights λ_i, $i=1,\dots,n$.

Let $\lambda_i = p(x_i)$. Let $\alpha_i = p(y_j \mid x_i)$ (j free variable).

Then $\sum_{i=1}^{n} p(x_i) \; f(p(y_j \mid x_i)) \geq f\left(\sum_{i=1}^{n} p(x_i) \cdot p(y_j \mid x_i)\right)$.

i.e. $\sum_{i=1}^{n} p(x_i) \; p(y_j \mid x_i) \cdot \log\left(p(y_j \mid x_i)\right) \geq \left(\sum_{i=1}^{n} p(x_i) \; p(y_j \mid x_i)\right) \log\left(\sum_{i=1}^{n} p(x_i) \cdot p(y_j \mid x_i)\right)$

i.e. $\sum_{i=1}^{n} p(x_i, y_j) \log p(y_j \mid x_i) \geq \left(\sum_{i=1}^{n} p(x_i, y_j)\right) \log p(y_j)$

$= p(y_j) \log p(y_j)$

Summing over j,

$$\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log p(y_j \mid x_i) \geq \sum_{j=1}^{m} p(y_j) \log p(y_j)$$

Hence $-H(Y|X) \geq -H(Y)$

$\Leftrightarrow H(Y) \geq H(Y|X)$. $\square$.

Kullback-Leibler Divergence.

Let $P = (p_1, \dots, p_n)$ & $Q = (q_1, \dots, q_n)$ be two probability distributions. Then the KL divergence of Q from P is defined by

$$D(P\|Q) = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} = \sum p_i \log p_i + \sum p_i \log \frac{1}{q_i}$$

$$= \sum p_i \log\left(\frac{1}{q_i}\right) - \left(-\sum p_i \log p_i\right)$$

$$= \sum p_i \log\left(\frac{1}{q_i}\right) - H(P).$$

By Huffman coding, an optimal compressor for P, the average code length for any message with symbol distribution P is $H(P)$.

Suppose you thought that the actual distribution is Q (maybe sample error)

The true distribution is P.

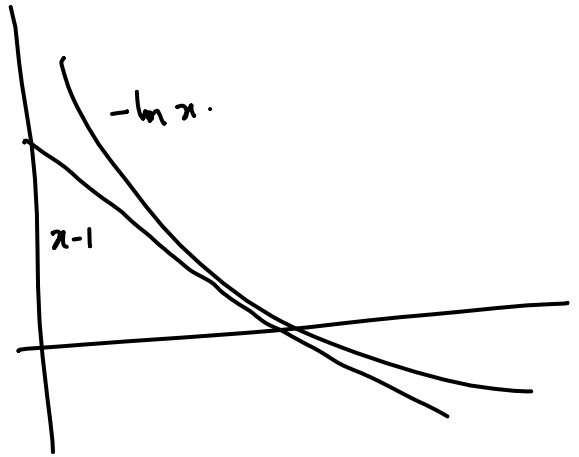
So the average code length $= \sum p_i \cdot \log \frac{1}{q_i}$

$$D(P||Q) = \left(\sum p_i \log \tfrac{1}{q_i}\right) - \left(\sum p_i \log \tfrac{1}{p_i}\right).$$

We show $D(P||Q) \geq 0$, and $D(P||Q) = 0$ iff $P = Q$..

Lemma.

$D(p\|q) \geq 0$.



$1-x \leq -\ln x$

Proof

$$D(p\|q) = \sum p_i \log \frac{p_i}{q_i}$$

$$= \sum p_i \left(-\log \frac{q_i}{p_i}\right)$$

$$\geq \sum p_i \left(1 - \frac{q_i}{p_i}\right) \quad \text{[fundamental inequality]}$$

$$= \sum_{i=1}^{n} (p_i - q_i) = \left(\sum_{i=1}^{n} p_i\right) - \left(\sum_{i=1}^{n} q_i\right) = 1 - 1 = 0.$$