Entropy

Let X be a discrete RV defined on an alphabet \( \mathcal{X} \) drawn according to the pmf \( p(x) \). Notation: \( X \sim p(x) \).

The entropy of \( X \) is defined as

\[
H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} = E\left[ \log \frac{1}{p(X)} \right] \quad \text{Sometimes } H(X) = H(p(x))
\]

\[E[ ]\]

\( X \sim \text{Bern}(p) \)

\( X = \begin{cases} 1 & p \\ 0 & 1-p \end{cases} \Rightarrow H(X) = p \log \frac{1}{p} + (1-p) \log \frac{1}{1-p} := H(p) \quad \text{binary entropy function} \]

Properties of \( H(X) \):

1. \( H(X) \geq 0 \) - information is a non-negative quantity
2. \( H(p(x)) \) is concave in \( p(x) \)
3. \( H(x) \leq \log |\mathcal{X}| \) - when it's binary: \( \log_2 \)

Jensen's inequality:

1. If \( f(x) \) is convex then \( E[f(x)] \geq f[E(x)] \)
2. If \( f(x) \) is concave then \( E[f(x)] \leq f[E(x)] \)

Proof of property (3):

\[
H(X) = E\left[ \log \frac{1}{p(X)} \right] \leq \log E\left( \frac{1}{p(X)} \right) = \log \left( \sum_{x \in \mathcal{X}} p(x) \cdot \frac{1}{p(x)} \right) = \log |\mathcal{X}|
\]
Let \((X, Y)\) be a pair of discrete random variables.

Then the conditional entropy

\[
H(Y|X) = \sum_x p(x) \frac{H(p(y|x))}{H(Y|x)}
\]

\[
= \sum_x p(x) \left( \sum_y p(y|x) \log \frac{1}{p(y|x)} \right)
\]

\[
= E \left[ \log \frac{1}{p(Y|X)} \right]
\]

\[
H(Y|X) = \sum_x p(x) H(1/p(y|x)) \leq H(\sum_x p(x) p(y|x)) = H(Y)
\]

- \(H(Y|X) \leq H(Y)\) with \("=\) if \(X\) and \(Y\) are independent
- \(H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)\)

Let \((X, Y) \sim p(x, y)\)

Then their joint entropy is

\[
H(X, Y) = E[\log p(X, Y)] = E[\log \frac{1}{p(X|Y)}] + E[\log \frac{1}{p(Y|X)}]
\]

- Chain rule

Let \(X^n = (X_1, \ldots, X_n)\)

\[
H(X^n) = H(X_1) + H(X_2|X_1) + \ldots + H(X_n|X^{n-1}) = \sum_{i=1}^n H(X_i|X^{i-1})
\]

\[
\leq \sum_{i=1}^n H(X_i)
\]

with \("=\) if \(X_1, \ldots, X_n\) are mutually independent
Relative Entropy

Let \( p(x) \) and \( q(x) \) be probability measures on \( X \).

Then \( D(p\|q) = D(p \circ \log q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \).

KL divergence:
We say \( p(x) \) is absolutely continuous with respect to \( q(x) \) if \( q(x) = 0 \Rightarrow p(x) = 0 \).

\( D(p\|q) = \infty \) if \( p(x) \) is not absolutely continuous with respect to \( q(x) \).

Properties:
1. \( D(p\|q) \geq 0 \) with "=" iff \( p = q \).
2. \( D(p\|q) \neq D(q\|p) \) in general.
3. \( D(p\|q) \) is convex in \((p,q)\) i.e.
   \[ \forall \alpha \in [0,1], \quad \alpha D(p\|q_1) + (1-\alpha) D(p\|q_2) \leq D(\alpha p + (1-\alpha) q_1 \| \alpha q_1 + (1-\alpha) q_2) \]
4. Chain rule
   For any probability measures \( P \) and \( Q \)
   \[ D(p\|q) = \sum_{x \in X} p(x) D(p(x)\|q(x)) = D(p\|Q) + \sum_{x \in X} p(x) D(p(x)\|q(x)) \]

For any probability measures \( P \) and \( Q \) (Radon–Nikodym derivative)

\[ D(p\|Q) = \int \log \frac{dP}{dQ} \, dp, \] \( P \) is a.e. with respect to \( Q \).

As a special case, if \( P \) and \( Q \) are a.e. with respect to a finite measure \( \mu \) (Lebesgue measure or counting measure) then \( D(p\|Q) = D(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} \, d\mu \).
Let $X$ be finite and $U$ be the uniform pmf on $X$

Then $D(p \| q) = \sum p(x) \log \frac{p(x)}{q(x)} = \log |X| - H(U) = H(x) - H(x)$

- $f$-divergence

Let $f : [0, \infty) \to \mathbb{R}$ and $f(1) = 0$

Then for $p(x)$ and $q(x)$

$D_f(p \| q) = E_q \left[ f \left( \frac{p(x)}{q(x)} \right) \right] = \int f \left( \frac{p(x)}{q(x)} \right) q(x) \, dx$

If $f(t) = \log t$, then $D_{\log}(p \| q) = \int q(x) \log \frac{p(x)}{q(x)} \, dx$ = $\int p(x) \log \frac{p(x)}{q(x)} \, dx$ = $D(p \| q)$

- Mutual information

Let $(X, Y) \sim p(x, y)$ be a pair of discrete RVs, mutual information between $X$ and $Y$ is

$I(X; Y) = D(p(x, y) \| p(x) p(y)) = \sum p(x) p(y) \log \frac{p(x, y)}{p(x) p(y)} = E \left[ \log \frac{p(x, y)}{p(x) p(y)} \right]$

- Properties:

1. $I(X; Y) \geq 0$
2. $I(X; Y) = 0$ if and only if $X$ and $Y$ are independent
3. As a function of $p(x, y)$, $I(X; Y)$ is concave in $p(x)$ for a fixed $p(y|x)$