1. Mutual Information

- The mutual information is the average amount of information that you get about $X$ from observing the value of $Y$

$$I(X;Y) = H(X) - H(X|Y) = H(X) + H(Y) - H(X,Y)$$

- The mutual information is symmetrical

$$I(X;Y) = I(Y;X)$$

Proof:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) = H(Y) + H(X) - H(Y,X) = I(Y;X)$$

2. Conditional Mutual Information

- Definition
  
  Conditional Mutual Information:

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z)$$

(The above result follows directly from the definition of $I(X;Y)$)

- Chain Rule for Mutual Information

$$I(X_1,X_2,...,X_n : Y) = \sum_{i=1}^{n} I(X_i;Y|X_{i-1},X_{i-2},...,X_1)$$

Proof:

$$I(X_1,X_2,...,X_n;Y) = H(X_1,X_2,...,X_n) - H(X_1,X_2,...,X_n|Y)$$

$$= \sum_{i=1}^{n} H(X_i|X_{i-1},X_{i-2},...,X_n) - \sum_{i=1}^{n} H(X_i|X_{i-1},X_{i-2},...,X_n,Y)$$

$$= \sum_{i=1}^{n} I(X_i;Y|X_{i-1},X_{i-2},...,X_1)$$

- Example of using chain rule for mutual information

$$Y = Z - X$$
Find $I(X, Z; Y)$

Solution: We have

$$I(X, Z; Y) = I(X; Y) + I(Z; Y|X) = H(X) - H(X|Y) + H(Z|X) - H(Z|Y, X)$$

$H(X) = \log 2 = 1$ bit

From Table 1. and $Y = Z - X$, we can derive the p.m.f of $(X, Y)$ as shown in the Table 2. Also form Table 1., we can see that $P(X = x, Z = z) = \frac{1}{4} = \frac{1}{2} \times \frac{1}{2} = P(X = x)P(x = z) \Rightarrow X, Z$ are independent.

<table>
<thead>
<tr>
<th>$p(X, Z)$</th>
<th>$Z = 0$</th>
<th>$Z = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
</tbody>
</table>

Table 1: The p.m.f of $(X, Z)$

<table>
<thead>
<tr>
<th>$p(X, Y)$</th>
<th>$Y = 0$</th>
<th>$Y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X = 0$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
<tr>
<td>$X = 1$</td>
<td>$\frac{1}{4}$</td>
<td>$\frac{1}{4}$</td>
</tr>
</tbody>
</table>

Table 2: The p.m.f of $(X, Y)$

From Table 2., we have:

$$H(X|Y) = - \sum_{x,y} p(x, y) \log p(x|y) = 4 \times \frac{1}{4} \log \frac{1}{2} = 1 \text{ bit}$$

$H(Z|X) = H(Z) = \log 2 = 1$ bit since $X, Z$ are independent.

$H(Z|Y, X) = 0$ bit due to $Z = X + Y$, thus, when known $Y, X, Z$ provides no information.

Therefore, $I(X, Z; Y) = H(X) - H(X|Y) + H(Z|X) - H(Z|Y, X) = 1 - 1 + 1 - 0 = 1$ bit.

3. Concave and Convex Functions

- Definition
  
  $f(x)$ is strictly convex over $(a, b)$ if
  $$f(\lambda u + (1 - \lambda)v) < \lambda f(u) + (1 - \lambda)f(v) \quad \forall u \neq v \in (a, b), 0 < \lambda < 1$$

  $f(x)$ is strictly concave over $(a, b)$ if
  $$f(\lambda u + (1 - \lambda)v) > \lambda f(u) + (1 - \lambda)f(v) \quad \forall u \neq v \in (a, b), 0 < \lambda < 1$$

- Examples
  
  * Strictly convex functions: $f(x) = x^2, f(x) = e^x, f(x) = x \log x$ (x > 0)
  * Strictly concave functions: $f(x) = \log x$ (x > 0), $f(x) = \sqrt{x}$ (x > 0)

- Technique to determine the convexity of a function: $\frac{d^2f(x)}{dx^2} > 0 \Rightarrow f(x)$ is convex.

  Note: $f(x)$ is convex (or concave) $\Leftrightarrow$ replace $<$ (or $>$) with $\leq$ (or $\geq$) in the above definitions
• Jensen’s Inequality

\[ f(X) \text{ convex } \Rightarrow E[f(X)] \geq f(E[X]) \]  \hspace{1cm} (1)
\[ f(X) \text{ strictly convex } \Rightarrow E[f(X)] > f(E[X]) \]  \hspace{1cm} (2)

Proof:
Assume that \( X \) is discrete. We will use induction to prove (1).

– In the case \( |X| = 1 \) then \( P(X) = 1 \Rightarrow f(X) = E[f(X)] = f(E[X]) \)

– In the case \( |X| = 2 \), then suppose \( X \) has 2 elements \( x_1, x_2 \) with corresponding probabilities \( p \) and \( 1 - p \). We have:

\[ E[f(X)] = pf(x_1) + (1-p)f(x_2) \geq f(px_1 + (1-p)x_2) = f(E[X]) \]

– Now, suppose (1) is true for the case \( |X| = n \). We consider the case \( |X| = n + 1 \):

\[ E[f(X)] = \sum_{i=1}^{n+1} p_i f(x_i) = \sum_{i=1}^{n} p_i f(x_i) + p_{n+1} f(x_{n+1}) \]

\[ = (1 - p_{n+1}) \sum_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} f(x_i) + p_{n+1} f(x_{n+1}) \]

\[ \geq (1 - p_{n+1}) f(\sum_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} x_i) + p_{n+1} f(x_{n+1}) \]

by the definition of convexity

\[ \geq f((1 - p_{n+1}) \sum_{i=1}^{n} \frac{p_i}{1 - p_{n+1}} + p_{n+1} x_{n+1}) \]

\[ = f(\sum_{i=1}^{n} p_i x_i) = f(E[X]) \]

Similarly, if \( f(X) \) is strictly convex, we have \( E[f(X)] > f(E[X]) \)

4. Relative Entropy

• Definition

Relative Entropy of Kullback-Leibler Divergence between two probability mass vectors (functions) \( p \) and \( q \) is defined as:

\[ D(p||Q) = \sum_{x \in A} p(x) \log \frac{p(x)}{q(x)} = E_p \left[ \log \frac{p(x)}{q(x)} \right] = E_p [- \log q(x)] - H(X) \]

• Properties

(a) \( D(p||q) \geq 0 \)
(b) \( D(p||q) \neq D(q||p) \)

• Example
### 5. Information Inequalities

- The Relative Entropy of Kullback-Leibler Divergence is non-negative
  \[ D(p||q) \geq 0 \]

**Proof:**

\[
D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_x p(x) \log \frac{q(x)}{p(x)}
\]

Now, \(-\log z\) is a convex function:

\[
D(p||q) \geq -\log \left[ \sum_x \frac{p(x)}{p(x)} q(x) \right] \geq -\log 1 = 0\]

Equality \(\iff p(x_i) = q(x_i) \ \forall i\)

- Uniform distribution has the highest entropy
  \[
  H(X) \leq \log |A|
  \]

**Proof:**

We have \(D(p||q) \geq 0\). Let \(q(x) = \frac{1}{|A|} \ \forall x\).

\[
D(p||q) = \sum_x p(x) \log q(x) - H(X) = \sum_x p(x) \log |A| - H(X) = \log |A| - H(X) \geq 0
\]

\(\Rightarrow H(X) \leq \log |A|\)

- Mutual Information is non-negative
  \[
  I(X;Y) \geq 0
  \]

**Proof:**
\[ I(X; Y) = H(X) + H(Y) - H(X, Y) \]
\[ = - \sum_x p(x) \log p(x) - \sum_y p(y) \log p(y) + \sum_{x,y} p(x,y) \log p(x,y) \]
\[ = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \]
\[ = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} \quad \text{where } q(x,y) = p(x)p(y) \]
\[ = D(p\|q) \geq 0 \]

- **Conditioning reduces entropy**

\[ H(X|Y) \leq H(X) \]

**Proof:**
\[ I(X; Y) = H(X) - H(X|Y) \geq 0 \]
\[ \Rightarrow H(X|Y) \leq H(X) \]

- **Independence bound**

\[ H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^n H(X_i) \]

**Proof:**
\[ H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^n H(X_i|X(i-1), \ldots, X_1) \leq \sum_{i=1}^n H(X_i) \]

- **Conditional independence bound**

\[ H(X_1, X_2, \ldots, X_n|Y_1, Y_2, \ldots, Y_n) \leq \sum_{i=1}^n H(X_i|Y_i) \]

**Proof:**
\[ H(X_1, X_2, \ldots, X_n|Y_1, Y_2, \ldots, Y_n) = \sum_{i=1}^n H(X_i|X_{i-1}, \ldots, X_1, Y_1, Y_2, \ldots, Y_n) \leq \sum_{i=1}^n H(X_i|Y_i) \]

- **Mutual information independence bound**

If \( X_1, X_2, \ldots, X_n \) or \( Y_1, Y_2, \ldots, Y_n \) are independent then
\[ I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) \geq \sum_{i=1}^n I(X_i; Y_i) \]
Proof If \( X_1, X_2, \ldots, X_n \) are independent then

\[
I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) = H(X_1, X_2, \ldots, X_n) - H(X_1, X_2, \ldots, X_n | Y_1, Y_2, \ldots, Y_n)
\geq \sum_{i=1}^{n} H(X_i) - \sum_{i=1}^{n} H(X_i | Y_i)
= \sum_{i=1}^{n} H(X_i) - H(X_i | Y_i)
= \sum_{i=1}^{n} I(X_i; Y_i)
\]

Similarly, this inequality is also true when \( Y_1, Y_2, \ldots, Y_n \) are independent since \( I(X; Y) \) is symmetrical □.