2. Alirezaei

\[ H(X) = - \sum_{i=1}^{m} p_i \cdot \log p_i \]
\[ \text{Prob}(X = x_i) = p_i, \quad i = 1, \ldots, m \]

**Theorem 2.1.8.**

a) \( 0 \leq H(X) \leq \log m \)

"=" in (i) \( \Leftrightarrow \) \( X \) has a single tone dist., i.e.,
\[ \exists x_i : \text{Prob}(X = x_i) = 1 \]

"=" in (ii) \( \Rightarrow \) \( X \) is uniformly dist., i.e.,
\[ \text{Prob}(X = x_i) = \frac{1}{m}, \quad \forall i = 1, \ldots, m \]

b) \( 0 \leq H(X | Y) \leq H(X) \)

"=" in (i') \( \Leftrightarrow \) \( \text{Prob}(X=x_i | Y=y_j) = 1 \forall i,j \)

with \( \text{Prob}(X=x_i, Y=y_j) > 0 \), i.e.,
\( X \) is totally dependent on \( Y \).

"=" in (ii') \( \Rightarrow \) \( X, Y \) are stock. independent.

c) \( H(X) \leq H(X, Y) \leq H(X) + H(Y) \)

"=" in (i) \( \Leftrightarrow \) \( Y \) is totally dependent on \( X \)

"=" in (ii) \( \Rightarrow \) \( X, Y \) are stock. independent.

d) \( H(X | Y, Z) \leq \text{min} \{ H(X | Y), H(X | Z) \} \)
Proof:

a) (ii) \[ H(K) = - \sum_{i=1}^{m} \Pi_i \cdot \log \Pi_i = \sum_{i=1}^{m} \Pi_i \cdot \log \frac{\Pi_i}{1} \]

Lemma 2.1.5

\[ \leq \log \left( \sum_{i=1}^{m} \Pi_i \cdot \frac{1}{\Pi_i} \right) \]

\[ = \log (\sum_{i=1}^{m} 1) = \log m \]

(c) obvious

b) Exercise

c) (i) By the chain rule, Th. 2.1.5.:

\[ H(X, Y) = H(X) + H(Y | X) \geq H(X) \]

equality follows from \( \leq \) b) (ii).

(ii) \( 0 \leq H(X) - H(X | Y) \leq H(Y) - [H(X, Y) - H(Y)] \)

\[ \Rightarrow H(X) + H(Y) \geq H(X, Y) \]

with equality from b) (i).

d) Analogous to b) (ii).

Ref. Cover & Thomas, Math.

Thinking the Future
Zukunft denken
Definition 2.1.9.

Let $X$, $Y$, $Z$ discrete r.v.

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

is called mutual information or synergy of $X$ and $Y$.

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

is called conditional mutual information of $X$ and $Y$ given $Z$.

Interpretation: $I(X; Y)$ is the reduction in uncertainty about $X$ when $Y$ is given or the amount of information about $X$ provided by $Y$.

see. Figures on the next page.

$$I(X; Y) = H(X) - H(X | Y)$$
$$H(Y | X) = H(X, Y) - H(X)$$
$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$
Thinking the Future
Zukunft denken
Note: by The 2.1.8 b) $\Sigma(X;Y) \geq 0$.

By definition it holds that:

$$\Sigma(X;Y) = - \sum_{i,j} p(x_i, y_j) \log p(x_i) + \sum_{i,j} p(x_i, y_j) \log p(x_i | y_j)$$

$$= - \sum_{i,j} p(x_i, y_j) \log p(x_i) + \sum_{i,j} p(x_i, y_j) \log p(x_i | y_j)$$

$$= \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i | y_j)}{p(x_i)}$$

$$= \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i | y_j)}{p(x_i) \cdot p(y_j)}$$

which shows symmetry in $X$ and $Y$.

**Example 2.1.10.** Binary symmetric channel (BSC)

Symbol error with probability $\epsilon$, $0 \leq \epsilon \leq 1$. 

Thinking the Future
Zukunft denken
Hence:

\[ P(Y=0 \mid X=0) = P(Y=1 \mid X=1) = 1-\varepsilon \]

\[ P(Y=0 \mid X=1) = P(Y=1 \mid X=0) = \varepsilon \]

Assume that input symbols are uniformly distributed: \( P(X=0) = P(X=1) = \frac{1}{2} \).

Then for the joint distributions:

\[ P(X=0, Y=0) = P(Y=0 \mid X=0) \cdot P(X=0) = (1-\varepsilon) \cdot \frac{1}{2} \]

\[ P(X=1, Y=0) = P(Y=0 \mid X=1) \cdot P(X=1) = \varepsilon \cdot \frac{1}{2} \]

\[ P(X=0, Y=1) = P(Y=1 \mid X=0) \cdot P(X=0) = (1-\varepsilon) \cdot \frac{1}{2} \]

\[ P(X=1, Y=1) = P(Y=1 \mid X=1) \cdot P(X=1) = \varepsilon \cdot \frac{1}{2} \]

Further:

\[ P(X=0 \mid Y=0) = \frac{P(X=0, Y=0)}{P(Y=0)} = 1-\varepsilon \]

\[ P(X=1 \mid Y=0) = 1-\varepsilon \]

\[ P(X=0 \mid Y=1) = P(X=1 \mid Y=0) = \varepsilon \]
For \( \log = \log_2 \)

\[ \Rightarrow H(K) = H(Y) = -\frac{1}{2} \cdot \log \frac{1}{2} - \frac{1}{2} \cdot \log \frac{1}{2} = 1 \text{ bit} \]

\[ H(K,Y) = 1 - (1-\varepsilon) \cdot \log (1-\varepsilon) - \varepsilon \cdot \log \varepsilon \]

\[ H(K|Y) = -(1-\varepsilon) \cdot \log (1-\varepsilon) - \varepsilon \cdot \log \varepsilon \]

\[ 0 \leq I(K;Y) = 1 + (1-\varepsilon) \cdot \log (1-\varepsilon) + \varepsilon \cdot \log \varepsilon \leq 1 \]

**Def. 2.1.11** (Kullback-Leibler divergence)

Let \( p = (p_1, \ldots, p_m) \), \( q = (q_1, \ldots, q_m) \) be stochastic vectors.

\[ D(p \parallel q) = \sum_{i=1}^{m} p_i \cdot \log \frac{p_i}{q_i} \]

is called KL divergence between \( p \) and \( q \) (or relative entropy).

\( D(p \parallel q) \) measures the divergence (distance, dissimilarity) between distributions \( p \) and \( q \). However, it is not a metric, neither symmetric nor satisfies the triangle inequality.

It measures how difficult it is for \( p \) to pretend it were \( q \).
Theorem 2.1.12.

a) \( D(p \| q) \geq 0 \) with \( = \) if \( p = q \).

b) \( D(p \| q) \) is convex in the pair \( (p, q) \).

c) \[ I(K; Y) = D((p(K, Y)_{ii} || (p(K, Y)_{ij}))) \]

proof.

a) By definition and Cor. 2.1.8.

b) Use the log-sum inequality Lemma 2.1.7.

Let \( p, r, \) and \( q, s \) be stochastic vectors.

For all \( i = 1, \ldots, m \), it holds that

\[
\lambda \frac{r_i}{q_i} + (1-\lambda) \frac{s_i}{q_i} \leq \frac{\lambda r_i}{q_i} + (1-\lambda) \frac{s_i}{q_i}
\]

Summing over all \( i = 1, \ldots, m \), it follows for all \( \lambda \in [0,1] \)

\[
D(\lambda p + (1-\lambda) q) \leq \lambda D(p \| q) + (1-\lambda) D(q \| p)
\]

c) By definition.

Note: \( D(p \| q) \neq D(q \| p) \)