Exercise. \( H(X \mid Y) = 0 \iff X = g(Y) \) for some function \( g(\cdot) \)

**Chain rule for MI**

\[
I(X_1, \ldots, X_n ; Y) = \sum_{i=1}^{n} I(X_i ; Y \mid X_1, \ldots, \hat{X}_i, \ldots, X_n)
\]

**Proof:**

\[
\text{LHS} = H(X_1, \ldots, X_n) - H(X_1, \ldots, X_n \mid Y)
\]
\[
= \sum_{i=1}^{n} H(X_i \mid X_1, \ldots, \hat{X}_i, \ldots, X_n)
\]
\[
- \sum_{i=1}^{n} H(X_i \mid X_1, \ldots, \hat{X}_i, \ldots, X_n, Y) = \checkmark.
\]

M.I. under conditioning.

\[
I(X ; Y) \overset{?}{=} I(X ; Y \mid Z)
\]

1. if \( Z = Y \) \( \Rightarrow I(X ; Y \mid Z) = 0 \leq I(X ; Y) \).
2. \( Z = X \oplus Y \) \( \lor X \perp Y \) \( \Rightarrow I(X ; Y) = 0 < I(X ; Y \mid Z) = 1 \).
\[
H(x_1, \ldots, x_n) \leq H(x_1) + \cdots + H(x_n)
\]

\& \text{ if } x_i \text{ are independent.}

Lemma: If \( x_1, \ldots, x_n \) are independent,

\[
I(x_i; \ldots; x_n \mid y) \geq \sum_{i=1}^{n} I(x_i; y)
\]

LHS: \( H(x_1, \ldots, x_n) - H(x_1, \ldots, x_n \mid y) \)

\[
= \sum_{i=1}^{n} H(x_i) - H(x_i \mid x_1, \ldots, x_{i-1}, y)
\]

\[
= \sum_{i=1}^{n} H(x_i) - H(x_i \mid x_1, \ldots, x_{i-1}, y)
\]

\[
= \sum_{i=1}^{n} H(x_i) - H(x_i \mid x_1, \ldots, x_n \mid y)
\]

\[
\geq \sum_{i=1}^{n} I(x_i; y)
\]

\[
\geq \sum_{i=1}^{n} I(x_i; y)
\]
Relative Entropy / Information Divergence

Kullback–Liebler Divergence: \( (KL) \)

Let \( p, q \) be distributions on \( U \), then,

\[
D(p \parallel q) := \sum_{x \in U} p(x) \log \frac{p(x)}{q(x)}.
\]

\( \Delta \) Not symmetric!

* Clearly, for \( p = q \Rightarrow D(p \parallel q) = 0 \).

* \( D(p \parallel q) \) finite \( \Rightarrow \) \( \text{Supp}(p) \subseteq \text{Supp}(q) \).

* **Gibbs inequality**: \( D(p \parallel q) \geq 0 \)

\( \Delta = 0 \) iff \( p = q \).

**Proof**: (Concavity) \( \Delta \)

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

\( \text{Jensen} \)

\[
\leq \log E(Z) = \log 1 = 0. \quad (\text{equality iff } Z = E(Z) \text{ everywhere})
\]
Theorem. \[ I(X; Y) = D\left( p(x, y) \parallel p(x)p(y) \right) \]

in particular, if \( X \perp Y \Rightarrow D(\parallel) = 0 \).

\( \checkmark \)

on the other hand: if \( X = Y \Rightarrow D \) is maximal.

\[ \text{Proof.} \]

RHS. \[ = \sum_{x, y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \]

\[ = \sum_{x, y} p(x, y) \log \frac{p(x|Y)}{p(x)} \]

\[ = \sum_{x, y} p(x, y) \log \frac{1}{p(x)} - \sum_{x, y} p(x, y) \log \frac{1}{p(x|Y)} \]

\[ = H(X) - H(X|Y) \]

\[ = I(X; Y). \]

\( \square \)

Some viewpoints on KL-divergence.

1. Source Coding interpretations:

"Increase in expected encoding length due to incorrect knowledge of the distribution."
2. Rejection Sampling

$p, q \leftarrow \text{Universe } U.$

* **Setting:** Sequence $X_1, X_2, \ldots$ iid. $\leftarrow q$.

* **Goal:** Output $i^* \text{ s.t. } X_{i^*} \text{ is distributed according to } p.$ (with no modification)

* Say $p$ is uniform on \( \{1, \ldots, 50\} \)

\[ q \leftarrow \{1, \ldots, 100\}. \]

Then, sample from $q$ until you hit \( \{1, \ldots, 50\} \).

* but the inverse is not possible.

\[ \downarrow \]

* **Solution:** Output the first $i^*$ s.t. $X_{i^*} \leq 50$

\[ \Rightarrow \mathbb{E}(i^*) = 2, \quad \mathbb{E}(\text{length } i) = 1. \]

* **Thm:** Strategy that achieves

\[ \mathbb{E}(\text{length } i^*) \leq \mathcal{D}(p \parallel q). \]

* over $p$ & internal randomness of strategy.

* (No proof now!)
(Compressing protocols) \[ A \] knows \( p \) \[ B \] knows \( q \) (exists common shared randomness)

Goal: Communicate \( x \leftarrow p \) to \( B \).

Thm. \exists \text{ interactive protocol between} \ A \ & \ B \ \text{with Expected } D(p \parallel q) + \text{small} \ \text{bits of communication st. in the end,}

1. \ A \ \text{outputs} \ a \leftarrow p
2. \ B \ \text{outputs} \ b \ \text{st.}

\[ \forall x, \ Pr(b = x \mid a = x) \geq 1 - \varepsilon. \]

(i.e., \( D(11) \) is the complexity of "agreement").

(2 & 3) are useful for compression of protocols.
Data processing inequality:

Suppose:
\[ X, Y, g(Y) \text{ deterministic func.} \]

Then:
\[ I(X; Y) \geq I(X; g(Y)). \]

Proof next time!