Example.

$$
\begin{aligned}
p &= \left(\frac{1}{2}, \frac{1}{2}\right) \\
q &= \left(\frac{1}{4}, \frac{3}{4}\right) \\
D(p||q) &= 0{,}207\ldots \\
D(q||p) &= 0{,}1887\ldots
\end{aligned}
$$

Definition. *Let X, Y be two RVs, values in $\mathcal{X}, \mathcal{Y}$. The* mutual information *of X and Y is*

$$I(X;Y) = D(p_{X,Y}||p_X \otimes p_Y)$$

Recall $p_{X,Y}(x,y) = \mathbb{P}[X = x, Y = y]$.

$$
\begin{aligned}
p_X(x) = \mathbb{P}[X = x] &= \sum_y p(x,y) \\
p_X \otimes p_Y(x,y) &= p_X(x)p_Y(y)
\end{aligned}
$$

Note: If X, Y are independent, then $I(X;Y) = 0$. We will see later that the converse is also true ($I(X;Y) = 0 \implies X, Y$ are independent).

$$
\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y) \log_2 \frac{p(x,y)}{p_X(x)p_Y(y)} \\
&= \sum_{x,y} p(x,y) \log_2 p(x,y) - \sum_x \underbrace{\sum_y p(x,y)}_{=p_X(x)} \log_2 p_X(x) - \sum_y \underbrace{\sum_x p(x,y)}_{=p_Y(y)} \log_2 p_Y(y) \\
&= H(X) + H(Y) - H(X,Y)
\end{aligned}
$$

Lemma 0.1.

$$
\begin{aligned}
I(X;Y) &= H(X) + H(Y) - H(X,Y) \\
&= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X)
\end{aligned}
$$

$$
\begin{aligned}
I(X;X) &= H(X) + H(X) - H(X,X) \\
&= H(X)
\end{aligned}
$$

since there is no new information from the second X:

$$p_{X,X}(x,x') = \mathbb{P}[X = x, X = x'] = \begin{cases} p_X(x) & , x = x' \\ 0 & , x \neq x' \end{cases}$$

Definition (Conditional mutual information)**.**

$$
\begin{aligned}
I(X;Y|Z) &= \sum_{z \in \mathcal{Z}} I(X;Y|Z = z)p_Z(z) \\
\ldots &= H(X|Z) - H(X|Y,Z) \\
&= H(Y|Z) - H(Y|X,Z)
\end{aligned}
$$

Remark.

$$p_{X,Y,Z}(x,y,z) = \mathbb{P}[X = x, Y = y, Z = z]$$

$$p_{X;Y|Z=z}(x,y) = \mathbb{P}[X = x, Y = y | Z = z] = \frac{p_{X,Y,Z}(x,y,z)}{p_Z(z)}$$

Recall chain rule:

$$H(X_1, \ldots, H_n) = \sum_{k=1}^{n} H(X_k | X_{k-1}, \ldots, X_1)$$

Lemma 0.2 (Version of Chain rule)**.**

$$I(X_1, \ldots, X_n; Y) = \sum_{k=1}^{n} I(X_k; Y | X_{k-1}, \ldots, X_1)$$

Remark (Recall from Analysis 1)**.** *Let I be an interval. A function $f : I \to \mathbb{R}$ is convex if*

$$f(px + (1-p)y) \leq pf(x) + (1-p)f(y)$$

for all $x, y \in I$ and all $p \in (0,1)$. add pic

A function f is called strictly convex if

$$f(px + (1-p)y) < pf(x) + (1-p)f(y)$$

for all $x, y \in I$ and all $p \in (0,1)$.

Remark (Theorem from Analysis 1)**.** *Let f be convex on an open interval I. Then*

1. *f continuous on I*
2. *for all $a \in I$ there exists $f'(a-), f'(a+)$, where $a-, a+$ are differential ..., and $f'(a-) \leq f'(a+)$* add pic
3. *Let $\alpha \in [f'(a-), f'(a+)]$. Then*

   $$f(x) \geq f(a) + \alpha(x-a)$$

   for all $x \in I$ and in the strictly convex case

   $$f(x) > f(a) + \alpha(x-a)$$

   for all $x \in I$ and $x \neq a$.

Theorem 0.3 (Jensens's inequality)**.** *Let f be convex on an open interval $I \subset \mathbb{R}$. And let X be an I-valued random variable such that $\underbrace{\mathbb{E}(X)}_{\in I}$ and $\mathbb{E}(f(X))$ exist and are finite. Then*

$$\mathbb{E}(f(X)) \geq f(\mathbb{E}(X)).$$

If f is strictly convex and X is not a.s. constant, then $\mathbb{E}(f(X)) > f(\mathbb{E}(X))$.

Proof. Set $a = \mathbb{E}(X)$, $\alpha \in [f'(a-), f'(a+)]$. Then

$$\begin{aligned} f(X) &\geq f(a) + \alpha(X - a) \\ \implies \mathbb{E}(f(X)) \geq \mathbb{E}(f(a) + \alpha(X - a)) = f(a) + \alpha \underbrace{(\mathbb{E}(X) - a)}_{=0} &= f(\mathbb{E}(X)) \end{aligned}$$

Let f be strictly convex and X not a.s. constant. Then

$$\mathbb{P}[X \neq a] = \mathbb{P}[\{\omega \in \Omega : X(\omega) = a\} \in \mathcal{A}] > 0$$

on $A = \{\omega \in \Omega : X(\omega) = a\}$,

$$f(X) > f(a) + \alpha(X - a)$$

Thus,

$$\mathbb{E}(f(X)) > f(a).$$

□

Remark.

$$\begin{aligned} Z = f(X) - (f(a) + \alpha(X - a)) \\ Z \geq 0 \end{aligned}$$

If $\mathbb{E}(Z) = 0$*, then* $Z = 0$ *a.s.*

Example. $A_n = [Z \geq \frac{1}{n}]$*:*

$$\begin{aligned} \mathbb{E}(Z) &\geq \mathbb{E}(Z \cdot \not\Vdash_{A_n}) \\ &\geq \mathbb{E}(\frac{1}{n} \not\Vdash_{A_n}) = \frac{1}{n}\,\mathbb{P}(A_n) \end{aligned}$$

$\mathbb{P}(A_n) = 0$ *and* A_n *is monotone increasing*

$$\bigcup_{n=1}^{\infty} A_n = [Z > 0]$$

Thus, $p[Z > 0] = 0$.

In our case: $\mathbb{E}(f(X)) = \sum_{k=1}^{n} f(x_k)p_k$ and $\mathbb{E}(X) = \sum_{k=1}^{n} x_k p_k$. X: values $x_k \in I$, $\mathbb{P}[X = x_k] = p_k$ for $k = 1, \ldots, n$

$$\sum_{k=1}^{n} p_k f(x_k) \geq f(\sum_{k=1}^{n} p_k x_k)$$

Example. *Prove the statement above*

$$\sum_{k=1}^{n} p_k f(x_k) \geq f(\sum_{k=1}^{n} p_k x_k)$$

by induction.

Theorem 0.4 (Information inequality)**.** *Let* $p(\cdot)$ *and* $q(\cdot)$ *be two probability distributions on* $\mathcal{X}$ *(finite). Then* $D(p||q) \geq 0$ *and* $D(p||q) = 0 \iff p(\cdot) = q(\cdot)$.

Proof. $-\log_2 : (0,\infty) \to \mathbb{R}$ is strictly convex.

$$D(p||q) = \mathbb{E}(-\log_2 \frac{q(X)}{p(X)})$$

The denominator $p(X)$ is always positive a.s.

$$\mathbb{E}(-\log_2 \frac{q(X)}{p(X)}) \geq -\log_2 \mathbb{E}(\frac{q(X)}{p(X)}) \tag{1}$$

Consider

$$\begin{aligned}\mathbb{E}(\frac{q(X)}{p(X)}) &= \sum_{x:p(x)>0} p(x) \cdot \frac{q(x)}{p(x)} \\ &= \sum_{x:p(x)>0} q(x) \leq 1\end{aligned}$$

Since $-\log$ is monotone decreasing,

$$-\log_2 \mathbb{E}(\frac{q(X)}{p(X)}) \geq -\log_2(1) = 0$$

If $D(p||q) = 0$, then we have equality in 1 and **??**. By **??**,

$$\sum_{x:p(x)>0} q(x) = 1$$

If $p(x) = 0$, then $q(x) = 0$.
$\frac{q(X)}{p(X)} = C$ is constant, $-\log_2 C = 0$ Then $C = 1$ and $p(x) = q(x)$ for all x. $\square$

Corollary 0.5. *$I(X;Y) \geq 0$*

$$I(X;Y) = 0 \iff X, Y \text{ are independent}$$

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

$$H(X) = H(X|Y) \iff X, Y \text{ are independent}$$