

# Your Paper

You

December 5, 2025

## Abstract

Your abstract.

## 1 Entropy and Mutual information

Let  $(\mathcal{X}, \mathcal{A}, \mathbb{P}_{\mathcal{A}})$  be a probability space. Let  $X : \mathcal{X} \rightarrow \mathbb{R}$  be a discrete random variable on the space with probability density function  $f_X : \mathcal{X} \rightarrow \mathbb{R}_+$ . We use the shorthand notation  $p(x) = \mathbb{P}_{\mathcal{A}}[X = x]$ . Let  $(\mathcal{Y}, \mathcal{B}, \mathbb{P}_{\mathcal{B}})$  be a probability space. Let  $Y : \mathcal{Y} \rightarrow \mathbb{R}$  be a discrete random variable on the space with probability density function  $f_Y : \mathcal{Y} \rightarrow \mathbb{R}_+$ . We use the shorthand notation  $p(y) = \mathbb{P}_{\mathcal{B}}[Y = y]$ .

This is a test.

**Definition 1.** Let  $X$  be a discrete random variable with distribution  $p(x)$ .

We define entropy as  $H_q(X) = \mathbb{E}(-\log_q p(X))$ . Since entropy was originally defined in the context of compression by Shannon in TODO, we usually use  $q = 2$  and  $H(X) = \mathbb{E}(-\log_2 p(X))$ .

**Theorem 1.** *Existence of Entropy: If  $\mathcal{X}$  is finite,  $H_q(X)$  exists.*

*Example, when Entropy does not exist: Let  $X =$ ,*

*Proof.*

□

Convention:  $0 \log(0) = 0$ .

Using definition of expected value:  $H_p(X) = \mathbb{E}(-\log_q(p(X))) = \mathbb{E}\left(\frac{1}{\log_q(p(X))}\right)$ .

Properties of Entropy:

**Definition 2.** Let  $X, Y$  be discrete random variables with marginal distributions  $p(x), p(y)$  and with joint distribution  $p(x, y)$ . We define the following:

Conditional Entropy:  $H(X | Y) = -\mathbb{E}(\log p(X | Y))$

Joint Entropy:  $H(X, Y) = -\mathbb{E}(\log p(X, Y))$

Relative Entropy:  $D(p(x) \| q(x)) = \mathbb{E}_p\left(\log \frac{p(X)}{q(X)}\right)$  with the conventions...

Mutual information:  $I(X; Y) := D(p(x, y) \| p(x)p(y))$

**Remark 1.** From now on, we always define the two discrete random variables  $X, Y$  from Definition 2.

**Remark 2.** Under the assumptions of Definition 2, we have

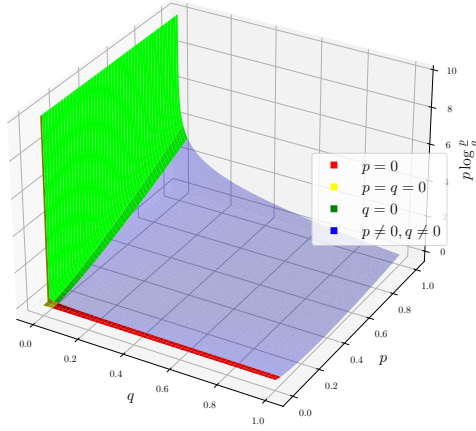
$$D(p(x)||q(x)) = \mathbb{E}_p \left( \log \frac{p(X)}{q(X)} \right) \quad (\text{def. of relative entropy})$$

$$= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \quad (\text{def. of expected value})$$

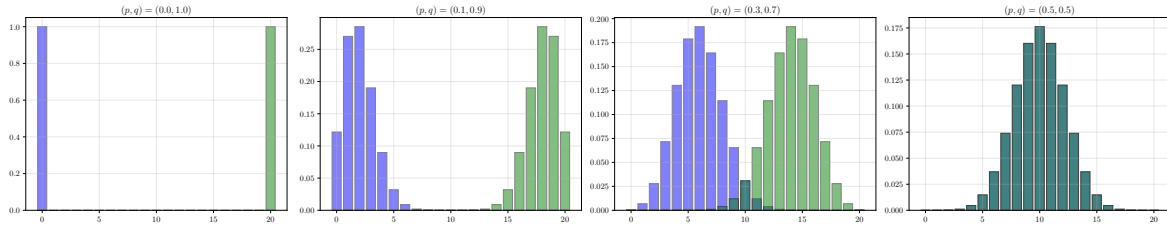
To understand the conventions, we can look at the limit cases:

1. **Case**  $p \in (0, 1], q = 0$ :  $\lim_{q \rightarrow 0^+} p \log \frac{p}{q} = \lim_{q \rightarrow 0^+} (p \log p - p \log q) = \infty$ .
2. **Case**  $p = 0, q \in (0, 1]$ :  $0 \log \frac{0}{q} = 0$ .
3. **Case**  $p = q = 0$ : Case 1 logic yields  $\lim_{q \rightarrow 0^+} p \log \frac{p}{q} = \infty$  and Case 2 logic yields  $0 \log \frac{0}{0} = 0$ .

As we want  $\sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$  to sum over  $x \in \mathcal{X}, p(x) > 0$ , we choose the convention  $0 \log \frac{0}{0} = 0$ . We can visualize the the pointwise relative entropy function  $(p, q) \mapsto \log \frac{p}{q}$ :



We can calculate the relative entropies for an example. The more alike the distributions are, the closer to zero the relative entropy is. Let  $X \sim B(20, \alpha)$  and  $Y \sim B(20, \beta)$  with  $(\alpha, \beta) \in [0, 1]^2$ .



$$D(p(x)||q(x)) = \sum_{x=0}^{20} p(x) \log \frac{p(x)}{q(x)}$$

$$D(p(x)||q(x)) = 0 \log \frac{0}{1} + 1 \log \frac{1}{0} = 0 + \infty = \infty \quad (\alpha = 0, \beta = 1)$$

$$D(p(x)||q(x)) \approx 35.2 \quad (\alpha = 0.1, \beta = 0.9)$$

$$D(p(x)||q(x)) \approx 6.8 \quad (\alpha = 0.3, \beta = 0.7)$$

$$D(p(x)||q(x)) = \sum_{x=0}^{20} p(x) \log 1 = 0 \quad (\alpha = 0.5, \beta = 0.5)$$

How do these concepts relate?

TODO: Add a venn diagram visualisation.

**Theorem 2.** *The chain rule holds:*

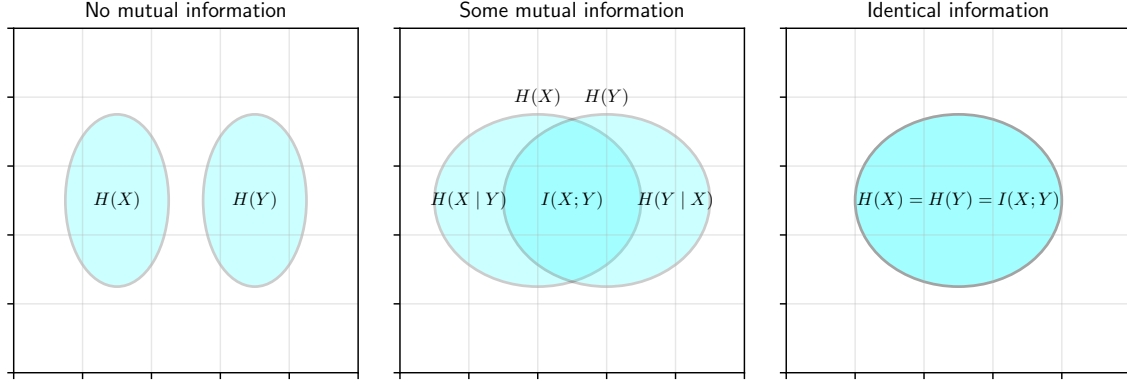
$$H(X, Y) = H(X) + H(Y \mid X)$$

*Proof.*

$$\begin{aligned} H(X, Y) &= -\mathbb{E}(\log p(X, Y)) && \text{(def. entropy)} \\ &= -\mathbb{E}(\log p(X)p(Y \mid X)) && \text{(conditional prob.)} \\ &= -\mathbb{E}(\log p(X)) + -\mathbb{E}(\log p(Y \mid X)) && \text{(log sum property)} \\ &= H(X) + H(Y \mid X) && \text{(def. entropy)} \end{aligned}$$

□

**Remark 3.** The relationship between entropy, conditional entropy and mutual information can be visualized:



**Theorem 3.** We can formalise the visual insight from before:

1.  $I(X;Y) = H(Y) - H(Y|X)$
2.  $I(X;Y) = I(Y;X)$
3.  $I(Y;X) = H(X) - H(X|Y)$
4.  $I(X;Y) = H(X,Y)$
5.  $I(X;X) = H(X)$

*Proof.* 1. We can use the definition of mutual information and relative entropy to obtain:

$$\begin{aligned}
 I(X;Y) &= D(p(x,y) \| p(x)p(y)) && \text{(by def. of mutual info.)} \\
 &= \mathbb{E}_{p(x,y)} \left( \log \frac{p(x,y)}{p(x)p(y)} \right) && \text{(by def. relative entropy)} \\
 &= \mathbb{E}_{p(x,y)} \left( \log \frac{p(x)p(y|x)}{p(x)p(y)} \right) && \text{(using cond. probability)} \\
 &= \mathbb{E}_{p(x,y)} \left( \log \frac{p(y|x)}{p(y)} \right) && \text{(simplify fraction)} \\
 &= \mathbb{E}_{p(x,y)} (\log p(y|x)) - \mathbb{E}_p(x,y) (\log p(y)) && \text{(simplify logarithm)} \\
 &= H(Y|X) - H(Y) && \text{(by def. of entropy)}
 \end{aligned}$$

2. The definition of mutual information yields:

$$\begin{aligned}
 I(X;Y) &= D(p(x,y) \| p(x)p(y)) && \text{(by def. of mutual info.)} \\
 &= D(p(y,x) \| p(y)p(x)) && \text{(TODO: Idk)} \\
 &= I(Y;X)
 \end{aligned}$$

3. Follows directly from 2 and 3.

4.

$$\begin{aligned}
 I(X;Y) &= H(Y) - H(X|Y) && \text{(by 1)} \\
 &= H(Y) - (H(X,Y) - H(X)) && \text{(chain rule)} \\
 &= H(X) + H(Y) - H(X,Y)
 \end{aligned}$$

5. Using 1 we get  $I(X;X) = H(X) - H(X|X) = H(X)$ .

□

## 2 Properties of Entropy and Mutual Information

**Remark 4.** We use the common definition of convex functions, concave functions from analysis.

**Theorem 4.** Let  $X$  be a random variable (not necessarily discrete) and  $f$  a function.

1. If  $f$  is convex, we have  $\mathbb{E}f(X) \geq f(\mathbb{E}X)$ .

2. If  $f$  is concave, we have  $\mathbb{E}f(X) \leq f(\mathbb{E}X)$ .

*Proof.* 1. We show  $f(\sum_{i=1}^n p_i x_i) \leq \sum_{i=1}^n p_i f(x_i)$  by induction.

The definition of convexity yields the base case  $i = 2$ :  $f(p_1 x_1 + p_2 x_2) \leq p_1 f(x_1) + p_2 f(x_2)$ .

We assume the claim holds for  $n - 1$  and the induction case goes as follows:

$$\begin{aligned} f\left(\sum_{i=1}^n p_i x_i\right) &= f\left(p_1 x_1 + (1 - p_1) \sum_{i=2}^n \frac{p_i}{1 - p_1} x_i\right) \\ &\leq p_1 f(x_1) + (1 - p_1) f\left(\sum_{i=2}^n \frac{p_i}{1 - p_1} x_i\right) \quad (\text{def. of convexity}) \\ &\leq p_1 f(x_1) + (1 - p_1) \sum_{i=2}^n \frac{p_i}{1 - p_1} f(x_i) \quad (\text{induct. hypo. applies bc. of } \sum_{i=2}^n \frac{p_i}{1 - p_1} = 1) \\ &= \sum_{i=1}^n p_i f(x_i) \end{aligned}$$

If  $\mathcal{X}$  is finite, the claim follows. If  $\mathcal{X}$  is countably infinite, the claim follows using ... .

2. Follows from part 1 applied to  $-f$ .

□

**Corollary 1.** Entropy and Mutual Information are non-negative:

1.  $0 \leq H(X)$

2.  $0 \leq D(p(x) \| q(x))$

3.  $0 \leq I(X; Y)$

*Proof.* 1. Note that  $\log(\frac{1}{[0,1]}) = \log([1, \infty]) = [0, \infty]$  and  $p(X)(\mathcal{X}) \in [0, 1]$ .

Using the monotonicity of the expected value, we obtain

$$0 \leq \mathbb{E}\left(\log\left(\frac{1}{p(X)}\right)\right) = -\mathbb{E}(\log(p(X))) = H(X)$$

2. We can prove this using Jensens inequality on a *concave* function:

$$\begin{aligned} -D(p(x) \| q(x)) &= -\mathbb{E}_p\left(\log\frac{p(X)}{q(X)}\right) && (\text{def. of relative entropy}) \\ &\leq -\log\left(\mathbb{E}_p\frac{p(X)}{q(X)}\right) = \log\left(\mathbb{E}_p\frac{q(X)}{p(X)}\right) && (-\log \text{ is convex, property of log}) \\ &= \log\left(\sum_{x \in \mathcal{X}} p(x) \frac{q(x)}{p(x)}\right) = \log\left(\sum_{x \in \mathcal{X}} q(x)\right) && (\text{def. of exp. value, simplify expr.}) \\ &= \log(1) = 0 && (q \text{ is a prob. function}) \end{aligned}$$

So equivalently, we have  $D(p(x) \| q(x)) \geq 0$ .

3. Follows from part 1:  $I(X; Y) = D(p(x, y) \| p(x)p(y)) \geq 0$ .

□

**Corollary 2.** 1. *More information can only decrease entropy:*  $H(X | Y) \leq H(X)$ .

2. *The uniform distribution maximizes entropy:*

$$H(X) \leq \log |\mathcal{X}| \text{ and } H(X) = \log |\mathcal{X}| \iff X \sim U(\mathcal{X}).$$

3.  $p \mapsto H(p)$  is concave.

*Proof.* 1.  $0 \leq I(X; Y) = H(X) - H(X | Y) \iff H(X | Y) \leq H(X)$

2. Let  $Y \sim U(\mathcal{X})$  st.  $\forall x \in \mathcal{X} : q(x) = \frac{1}{|\mathcal{X}|}$ .

$$\begin{aligned} 0 &\leq D(p(x) \| q(x)) \\ &= \mathbb{E}_p \left( \log \frac{p(X)}{q(X)} \right) \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log (p(x) |\mathcal{X}|) \\ &= \sum_{x \in \mathcal{X}} p(x) \log p(x) + |\mathcal{X}| \sum_{x \in \mathcal{X}} p(x) \\ &= -H(X) + |\mathcal{X}| = |\mathcal{X}| - H(X) \end{aligned}$$

This is equivalent to  $H(X) \leq \log |\mathcal{X}|$ .

3.

$$H(p) = -p \log p =$$

□

### 3 Axiomatic Definition