

# Introduction to Entropy

Erik Stern

December 6, 2025

## Abstract

Your abstract.

## Contents

<b>1</b>	<b>Entropy and Mutual information</b>	<b>1</b>
1.1	Definitions and Conventions	1
1.2	Chain Rules and Mutual Information	3
<b>2</b>	<b>Inequalities for Entropy and Mutual Information</b>	<b>5</b>
2.1	Convexity and Jensen Inequality	5
2.2	The Information Inequality	8
2.3	Application to Optimisation	9

## 1 Entropy and Mutual information

### 1.1 Definitions and Conventions

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.** TODO: This is all incorrectly defined. 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  $H(X) = \mathbb{E}(-\log_2 p(X))$ . TODO: Add  $\log = \log_2$  convention.

**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:  $H_p(X) = \mathbb{E}(-\log_q(p(X))) = \mathbb{E}\left(\frac{1}{\log_q(p(X))}\right)$ .

**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:  $I(X; Y) := D(p(x, y) || p(x)p(y))$

Conditional Mutual information:  $I(X; Y | Z) := H(X | Z) - H(X | Y, Z)$

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

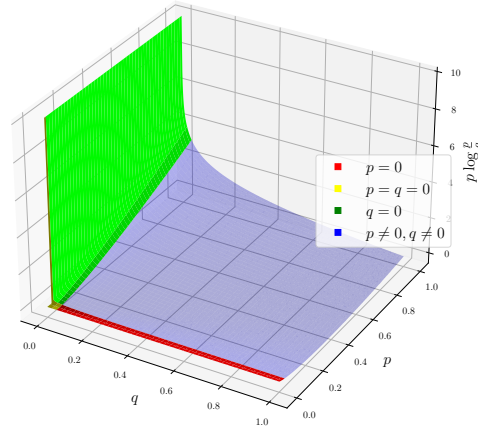
**Remark 2.** Under the assumptions of Definition 2, we have

$$\begin{aligned} D(p(x)||q(x)) &= \mathbb{E}_p \left( \log \frac{p(X)}{q(X)} \right) && \text{(def. of relative entropy)} \\ &= \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} && \text{(def. of expected value)} \end{aligned}$$

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

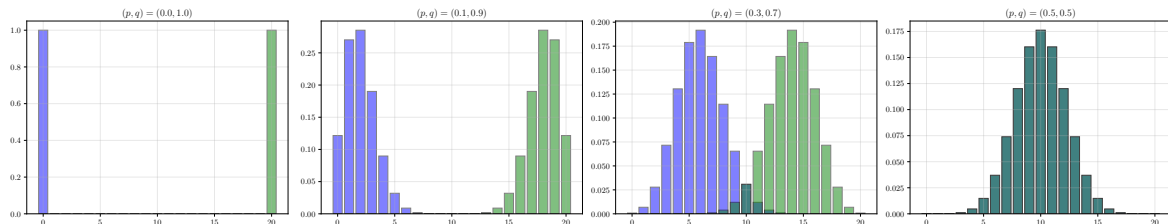
1. **Case**  $p \in (0, 1], q = 0$ :  $\lim_{q \rightarrow 0+} p \log \frac{p}{q} = \lim_{q \rightarrow 0+} (p \log p - p \log q) = \infty$ .
2. **Case**  $p = 0, q \in (0, 1]$ :  $0 \log \frac{0}{q} = 0$ .
3. **Case**  $p = q = 0$ : Case 1 logic yields  $\lim_{q \rightarrow 0+} p \log \frac{p}{q} =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



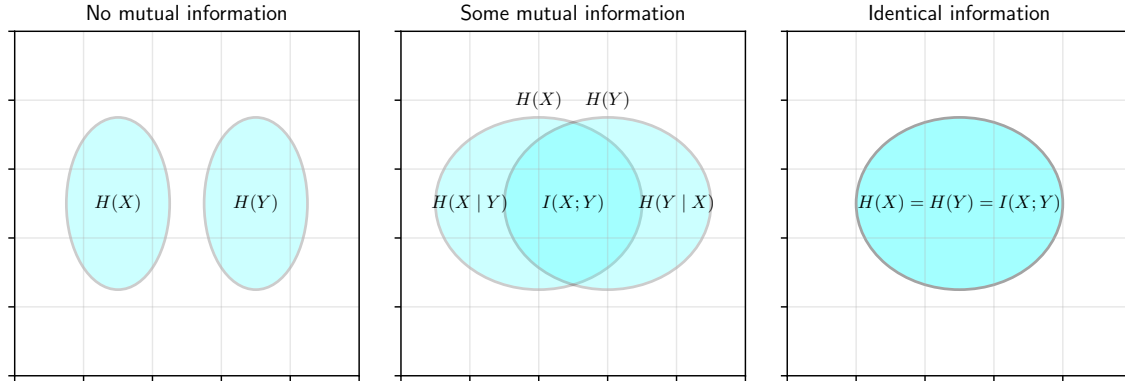
We can calculate the relative entropies for an example. The more alike the distributions are, the closer to zero the relative entropy is. Let  $X \sim B(20, \alpha)$  and  $Y \sim B(20, \beta)$  with  $(\alpha, \beta) \in [0, 1]^2$ .

$$\begin{aligned} D(p(x)||q(x)) &= \sum_{x=0}^{20} p(x) \log \frac{p(x)}{q(x)} \\ \alpha = 0, \beta = 1 : \quad D(p(x)||q(x)) &= 0 \log \frac{0}{1} + 1 \log \frac{1}{0} = 0 + \infty = \infty \\ \alpha = 0.1, \beta = 0.9 : \quad D(p(x)||q(x)) &\approx 50.7 \\ \alpha = 0.3, \beta = 0.7 : \quad D(p(x)||q(x)) &\approx 9.8 \\ \alpha = 0.5, \beta = 0.5 : \quad D(p(x)||q(x)) &= \sum_{x=0}^{20} p(x) \log 1 = 0 \end{aligned}$$



## 1.2 Chain Rules and Mutual Information

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



**Theorem 2.** The following statements about Entropy and Mutual Information are called Chain Rules:

1.

2.

*Proof.*

□

**Corollary 1.**

*Proof.*

□

**Corollary 2.** The following statements are specialized for two random variables  $X, Y$ :

1.  $H(X, Y) = H(X) + H(Y | X)$

2.

*Proof.*

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

□

**Theorem 3.** There are multiple equivalent ways to express Mutual Information:

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:

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

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 Inequalities for Entropy and Mutual Information

### 2.1 Convexity and Jensen Inequality

**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. Distinguish two cases. If  $\mathcal{X}$  is finite:

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:

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

Else:

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

□

**Corollary 3.** 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:  $I(X; Y) = D(p(x, y) \| p(x)p(y)) \geq 0$ .

□

**Remark 5.** There are multiple natural questions we can ask about Entropy. We will look at an example for each of them and then prove the results in the follow-up theorem.

1. What distribution maximises the value of Entropy? I simulated

From the data, it seems plausible that the uniform distribution maximises entropy. We

2. Can joint entropy increase if we add redundant information?

Let  $\Omega = \{1, 2\}$ ,  $p$  uniform TODO,  $X, Y : \Omega \rightarrow \mathbb{R}$ ,  $X(\omega) = \omega$  and  $Y(\omega) = 2\omega$ .  $Y$  is redundant to  $X$ , as  $Y = 2X$ .

$$H(X) = \sum_{i=1}^2 -0.5 \log 0.5 = -\log 0.5 = \log 2 = 1$$

$$H(X, Y) = \sum_{(i,j) \in \{(1,2), (2,4)\}} -0.25 \log 0.25 = 2 * 0.25 * 2 = 1$$

Pairs like  $(1, 4)$  have probability zero and do not contribute to the sum in  $H(X, Y)$ . This example suggest entropy never increases. We will prove that in the follow-up theorem.

3. What happens to the Entropy if we add independent noise to our measurements?

Let  $X \sim U(\{1, 2, 3, 4\})$  be the original signal,  $N \sim U(\{\})$  the noise and let  $X, N$  be independent. Then  $S := X + N$  is a noisy signal.

We can compute the crirical variance  $a$ .

**Theorem 5.** We can formalise the previous observations including some more:

1. More information can only decrease entropy:  $H(X | Y) \leq H(X)$ .
2. The uniform distribution maximizes entropy:  
 $H(X) \leq \log |\mathcal{X}|$  and  $H(X) = \log |\mathcal{X}| \iff X \sim U(\mathcal{X})$ .
3. If information from  $Y$  does not add anything to  $X$ , then  $Y$  must be derived from  $X$ :  
 $H(Y | X) = 0 \implies \exists f : Y = f(X)$  almost surely
4. If independent noise is added to a random variable, entropy can only increase:  
Set  $Z = X + Y$ . Then we have  $X, Y$  independent  $\implies H(X) \leq H(Z) \wedge H(Y) \leq H(Z)$

*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  $H(X) \leq \log |\mathcal{X}|$ . TODO: Iff part.

3. We have  $H(Y | X) = -\mathbb{E}(\log p(Y | X)) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} -p(x, y) \log p(y | x)$ .

Additionally, we have  $\forall (x, y) \in \mathcal{X} \times \mathcal{Y} : -p(x, y) \log p(y | x) \geq 0$ .

Combining those facts, we get

$$\begin{aligned} H(Y | X) &= 0 \\ \iff \forall (x, y) \in \mathcal{X} \times \mathcal{Y} : p(x, y) \log p(y | x) &= 0 \\ \iff \forall (x, y) \in \mathcal{X} \times \mathcal{Y} : p(x, y) &= 0 \oplus \log p(y | x) = 0 \\ \iff \forall (x, y) \in \mathcal{X} \times \mathcal{Y} : p(x, y) &= 0 \oplus p(y | x) = 1 \end{aligned}$$

This tells us, that either  $p(x, y) = 0$  or  $p(x, y) = p(y | x)p(x) = p(x) = 1$ .  
Now we can finish up the argument. Set  $A = \{x \in \mathcal{X} : p(x) > 0\}$ .  
Define  $(y_x)_{x \in \mathcal{X}}$  such that  $\forall x \in \mathcal{X} : p(x, y_x) > 0$ .  
Set  $f : A \rightarrow \mathcal{Y}, x \mapsto y_x$ . This gets us  $\text{Im} f = y_x : x \in \mathcal{X} = y \in \mathcal{Y} : p(x, y) > 0$ .  
So  $Y = f(X)$  almost surely.

4. We have

$$\begin{aligned}
H(Z | X) &= -\mathbb{E}(\log p(Z | X)) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} -p(x, y) \log P[Z = x + y | X = x] \\
&= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} -p(x, y) \log P[X + Y = x + y | X = x] \\
&= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} -p(x, y) \log P[Y = y | X = x] \\
&= -\mathbb{E}(\log p(Y | X)) = H(Y | X)
\end{aligned}$$

$$H(Z, X) = H(X) + H(Z | X) \quad H(Y, X) = H(X) + H(Y | X)$$

$$H(Y, X) = H(X) + H(Y | X) \quad \text{Using the Chain rule, we get } H(X, Y) = H(Y | X) + H(X)$$

□

## 2.2 The Information Inequality

Hello

## 2.3 Application to Optimisation

Hello