Introduction to Entropy

Erik Stern

December 10, 2025

Abstract

This seminar paper provides a foundational overview of Information Theory, focusing on Entropy, Mutual Information, and Relative Entropy. We establish key definitions and prove fundamental inequalities, using Jensen's Inequality and the Log-Sum Inequality. To build intuition, theoretical results are complemented by numerical simulations of discrete random variables. Finally, we derive the relationship between Relative Entropy and Cross-Entropy, demonstrating their practical application in optimizing neural networks for the MNIST digit classification task.

Contents

1 Entropy and Mutual information

1.1 Definitions and Conventions

Definition 1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, let $\mathcal{X}$ be a countable set and let $X : \Omega \to \mathcal{X}$ be a discrete random variable on $(\Omega, \mathcal{A}, \mathbb{P})$. We can then define

$$\textbf{Entropy wrt. base:} \quad H_b(X) = \mathbb{E}\left(-\log_b p_X(X)\right) = -\sum_{x \in \operatorname{supp}(X)} p_X(x) \log_b p_X(x)$$

$$\textbf{Entropy conventionally:} \quad H(X) = H_2(X)$$

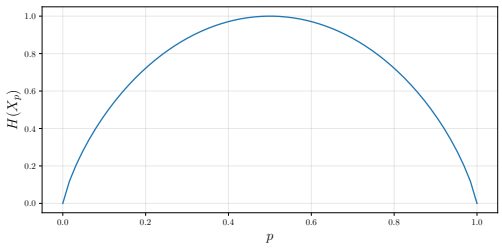
Remark 1.

1. Let $\mathcal{X}, \mathcal{Y}$ be countable sets and $X : \Omega \to \mathcal{X}$, $Y : \Omega \to \mathcal{Y}$ be discrete random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. From now on, the random variables X, Y are always available for use.
2. We do **not** use the shorthand notations $p(x) = \mathbb{P}[X = x]$ and $p(y) = \mathbb{P}[Y = y]$ from [CT05], to keep the notation easily understandable.
3. We use the convention $\log = \log_2$, as the entropy H is defined wrt. base 2.
4. We also use the following convention and justify it through a continuity argument:

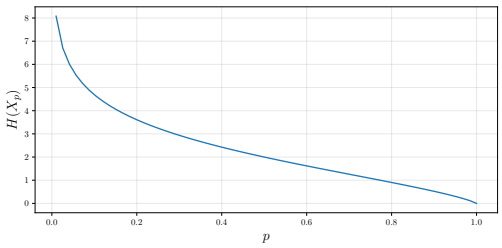
$$0 \log 0 = \lim_{x \to 0^+} x \log x = \lim_{x \to 0^+} \frac{\log x}{\frac{1}{x}} = \lim_{x \to 0^+} \frac{\frac{1}{\ln(2)x}}{-\frac{1}{x^2}} = \lim_{x \to 0^+} \frac{-x}{\ln(2)} = 0$$

   This choice is sensible, as $\log x$ is not defined for negative x.
5. The conventions, definitions and theorems are from Definitions, Theorems, Remarks and Exercises in *Elements of Information Theory, second edition* (see [CT05]).

Remark 2 (Existence of Entropy)**.** Note that if $|\mathcal{X}|$ is finite, $(\forall p \in \mathbb{R}_+ : H_b(X)$ finite) and $H(X) \leq |\mathcal{X}|$ (see Theorem 5). For $|\mathcal{X}|$ countably infinite, there are counterexamples where $H_b(X) = \infty$ (see [Mat]). From now on, we will assume that entropy is finite.



(a) Bernoulli Rand. Variable Entropy $H(X_p)$



(b) Geometric Rand. Variable Entropy $H(X_p)$

Example 1 (Entropy of bernoulli variable)**.** Let $p \in (0,1)$ and $X_p \sim B(1,p)$ be a weighted coin flip.
We can calculate the Entropy of X_p: $H(X_p) = -p\log p - (1-p)\log(1-p)$.
A visual inspection (see Figure 1a) reveals that $H(X_p)$ seems to be maximised for $p = 0.5$ and minimised for $p \in \{0,1\}$. An increase in uncertainty about the result of the coin flip seems to correspond with an increase in entropy.

Example 2 (Entropy of geometric variable)**.** Let $p \in (0,1)$ and $X_p \sim G(p)$ be the number of times a weighted coin is flipped, until the first head occurs. We will calculate the Entropy of X_p. It will require the two well-known series:

$$\forall r \in (0,1) : \sum_{n\in\mathbb{N}_0} r^n = \frac{1}{1-r} \tag{1}$$

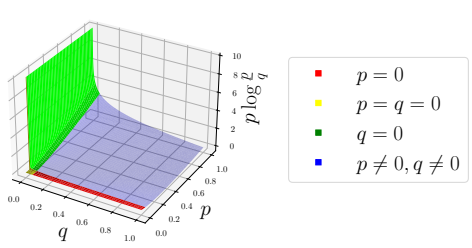
$$\forall r \in (0,1) : \sum_{n\in\mathbb{N}_0} nr^n = \frac{r}{(1-r)^2} \tag{2}$$

We can now directly calculate the Entropy of X_p:

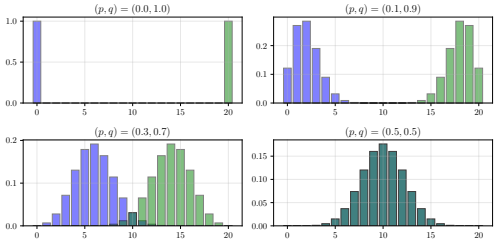
$$\begin{aligned}
H(X_p) &= \sum_{x\in\mathbb{N}} -p(x)\log p(x) && \text{(Def. of Entropy)}\\
&= -\sum_{x\in\mathbb{N}} (1-p)^{x-1}p\log\left((1-p)^{x-1}p\right) && \text{(Subst. in geometric mass function)}\\
&= -\sum_{x\in\mathbb{N}} (1-p)^{x-1}p\left((x-1)\log(1-p)+\log p\right) && \text{(Log rules)}\\
&= -p\log(1-p)\sum_{x\in\mathbb{N}_0}\left((1-p)^x x\right) - p\log p\sum_{x\in\mathbb{N}_0}\left((1-p)^x\right) && \text{(Factor out constants)}\\
&= -p\log(1-p)\frac{1-p}{(1-(1-p))^2} - p\log p\frac{1}{1-(1-p)} && \text{(Use series 1 and 2)}\\
&= -p\log(1-p)\frac{1-p}{p^2} - p\log p\frac{1}{p} && \text{(Simplify expr.)}\\
&= \frac{-(1-p)\log(1-p) - p\log p}{p} && \text{(Simplify expr.)}
\end{aligned}$$

For $p = 0.5$ we get $H(X_{0.5}) = \frac{-0.5\log 0.5 - 0.5\log 0.5}{0.5} = -2\log 0.5 = 2$.
We can visually inspect $(0,1) \to \mathbb{R}, p \mapsto H(X_p)$ (see Figure 1b) to get a feeling for the entropy of X_p. An increase in p is linked to lower variance and more concentration of the distribution towards zero. Based on the plot, that increase looks to be linked to a lower entropy and vice-versa.

We will now find a strategy, that calculates the number of flips until the first head occurs using simple yes/no questions. A simple strategy is to ask "Is $X = 1$?", "Is $X = 2$?" and so on. Under this strategy, the number of questions required is exactly the value of X. Thus, the average number of questions is simply $\mathbb{E}(X_{0.5}) = \frac{1}{0.5} = 2$. This matches the entropy $H(X_{0.5}) = 2$.



(a) Pointwise Relative Entropy



(b) Relative Entropies of Binomial Distributions

Definition 2.

We define the following:

$$\textbf{Conditional Entropy:} \quad H(X \mid Y) = -\mathbb{E}\left(\log p_{(X|Y)}(X \mid Y)\right)$$
$$\textbf{Joint Entropy:} \quad H(X, Y) = -\mathbb{E}\left(\log p_{(X,Y)}(X, Y)\right)$$

Let p and q be two probability mass functions on the same set $\mathcal{Z}$. We define the following:

$$\textbf{Relative Entropy:} \quad D(p\|q) = \mathbb{E}_{X\sim p}\left(\log \frac{p(X)}{q(X)}\right) \text{ with conventions from Remark 3}$$

(Sometimes also called KL-Divergence)

Using Relative Entropy, we can define the Mutual Information between the variables X, Y:

$$\textbf{Mutual Information:} \quad I(X;Y) = D(p_{(X,Y)}\|p_X\, p_Y)$$

Remark 3. We have

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

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

1. Case $p \in (0, 1], q = 0$: $\lim_{q\to 0^+} p \log \frac{p}{q} = \lim_{q\to 0^+} (p \log p - p \log q) = \infty$.
2. Case $p = 0, q \in (0, 1]$: $0 \log \frac{0}{q} = 0$.
3. Case $p = q = 0$: Case 1 logic yields $\lim_{q\to 0^+} p \log \frac{p}{q} = \infty$ and Case 2 logic yields $0 \log \frac{0}{0} = 0$. So what do we choose? 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 $0 \log \frac{0}{0} = 0$.

Figure 2a visualizes the pointwise relative entropy function $(p, q) \mapsto p \log \frac{p}{q}$.

Example 3. To understand the concept, we can now calculate the relative entropies for an example. Let $X \sim B(20, \alpha)$ and $Y \sim B(20, \beta)$ with $(\alpha, \beta) \in [0, 1]^2$.

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

Figure 2b visualizes the two discrete distribution functions in the cases above. Intuitively, the more overlap the distributions have, the closer to zero the relative entropy is.

1.2 Mutual Information and Chain Rules

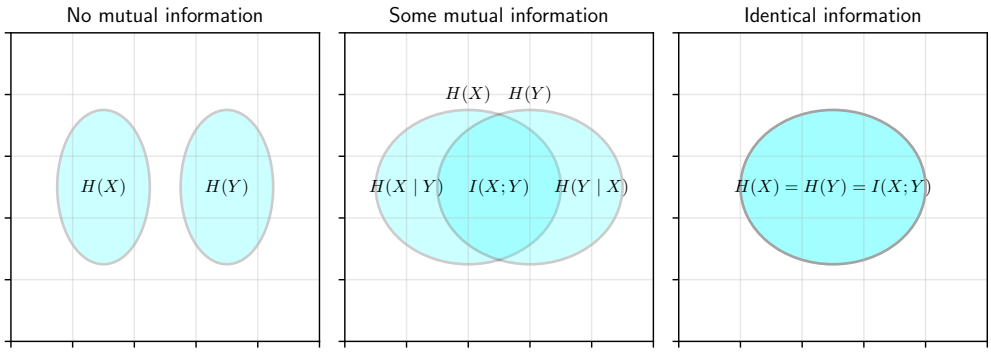


Figure 3: Relationship between Entropy, Conditional Entropy and Mutual Information

Theorem 1 (Chain Rule for Entropy)**.** *We have* $H(X,Y) = H(X) + H(Y \mid X)$

Proof.

$$\begin{aligned} H(X,Y) &= -\mathbb{E}\left(\log p(X,Y)\right) = -\mathbb{E}\left(\log p(Y \mid X)p(X)\right) \\ &= -\mathbb{E}\left(\log p(Y \mid X)\right) - \mathbb{E}\left(\log p(X)\right) = H(X) + H(Y \mid X) \end{aligned}$$

□

Theorem 2. *There are multiple equivalent ways to express Mutual Information (see Figure 3):*

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. $I(X;Y) = H(X) + H(Y) - H(X,Y)$
5. $I(X;X) = H(X)$

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

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

2. The definition of mutual information yields:

$$\begin{aligned} I(X;Y) &= D(P[X=x, Y=y] \| P[X=x]P[Y=y]) && \text{(by def. of mutual info.)} \\ &= D(P[Y=y, X=x] \| P[Y=y]P[X=x]) = I(Y;X) \end{aligned}$$

3. Follows directly from 1 and 2.

4.

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

5. Using the Defintion we get $H(X \mid X) = 0$. Using 1 we get $I(X;X) = H(X) - H(X \mid X) = H(X)$.

□

Definition 3. Let p and q be probability mass functions on $\mathcal{X} \times \mathcal{Y}$. We define the following:

$$\begin{aligned}
\textbf{Conditional Relative Entropy:} \quad & D(p_{Y|X} \| q_{Y|X}) \\
&= \mathbb{E}_{X \sim p_X} D(p(Y \mid X) \| q(Y \mid X)) \\
&= \sum_{x \in \mathcal{X}} p(x) D(p(y \mid x) \| q(y \mid x)) \\
&= \mathbb{E}_{(X,Y) \sim p(x,y)} \left(\log \frac{p(Y \mid X)}{q(Y \mid X)} \right) \\
&= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \log \frac{p(y \mid x)}{q(y \mid x)}
\end{aligned}$$

$$\textbf{Conditional Mutual Information:} \quad I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$$

Example 4. The Conditional Relative Entropy can be illustrated using a Christmas example:
Let $\Omega = \{1, \cdots, 12\}$, $\mathbb{P}$ uniformly distributed,
$\mathcal{X} = \{\textbf{JANUARY}, \cdots, \textbf{DECEMBER}\}$ and $\mathcal{Y} = \{0, 1\}$.
Our variable X gives us the month of the year:
We define it as $X : \Omega \to \mathcal{X}, \omega \mapsto$ **the corresponding month**.
Our variable Y answers the question "Is Christmas Advent?":
We define it as $Y : \Omega \to \mathcal{Y}, \omega \mapsto \mathbb{1}_{\{\textbf{NOVEMBER}, \textbf{DECEMBER}\}}(\omega)$.
We have lived on earth our entire lives. Our model of reality p is correct!
We define it as

$$p(Y = 1 \mid x) = \begin{cases} 1, & \text{if } x \in \{\textbf{NOVEMBER}, \textbf{DECEMBER}\} \\ 0, & \text{otherwise} \end{cases}$$

The alien ET visits earth. He sees the Christmas trees in January and assumes Advent continues into January. In November and December, he is uncertain too. His model of reality q is not correct!
We define it as

$$q(Y = 1 \mid x) = \begin{cases} 0.5, & \text{if } x \in \{\textbf{NOVEMBER}, \textbf{DECEMBER}, \textbf{JANUARY}\} \\ 0, & \text{otherwise} \end{cases}$$

Let us calculate the individual relative entropies:

1. $D(p(y \mid \textbf{NOV.}) \| q(y \mid \textbf{NOV.})) = 0 + 1 \log \frac{1}{0.5} = 1$
2. $D(p(y \mid \textbf{DEC.}) \| q(y \mid \textbf{DEC.})) = 0 + 1 \log \frac{1}{0.5} = 1$
3. $D(p(y \mid \textbf{JAN.}) \| q(y \mid \textbf{JAN.})) = 1 \log \frac{1}{0.5} + 0 = 1$
4. Otherwise: $D(p(y \mid x) \| q(y \mid x)) = 0$

We can now calculate the Conditional Relative Entropy:

$$\begin{aligned}
D(p_{Y|X} \| q_{Y|X}) &= \sum_{x \in \mathcal{X}} p_X(x) D(p(y \mid x) \| q(y \mid x)) \\
&= \frac{1}{12} (1 + 1 + 1) = \frac{3}{12} = 0.25
\end{aligned}$$

So the expression calculates how much of an error ET makes per month, on average.

Theorem 3 (Chain Rules)**.** *Let $n \in \mathbb{N}, n \geq 2$, $(X_1, \cdots, X_n) \sim p(x_1, \cdots, x_n)$ and Y a random variable. The following statements about Entropy and Mutual Information are called Chain Rules:*

1. $H(X_1, \cdots, X_n) = \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \cdots, X_1)$
2. $D(p(x)\|q(x)) = D(p(x \mid y)\|q(x \mid y)) + D(p(y)\|q(y))$
3. $I(X_1, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, \cdots, X_1)$

Proof. 1. Prove this result using induction by n.
Base case $n = 2$:

$$\begin{aligned} H(X_1, X_2) &= -\mathbb{E}\left(\log p(X_1, X_2)\right) \\ &= -\mathbb{E}\left(\log p(X_2 \mid X_1) p(X_1)\right) \\ &= -\mathbb{E}\left(\log p(X_2 \mid X_1)\right) + -\mathbb{E}\left(\log p(X_1)\right) \\ &= H(X_1) + H(X_2 \mid X_1) \end{aligned}$$

Assume the theorem holds for $n-1$. Induction case $n-1$ to n:

$$\begin{aligned} H(X_1, \cdots, H_n) &= H(X_n | X_1, \cdots, H_{n-1}) + H(X_1, \cdots, H_{n-1}) && \text{(apply base case)} \\ &= H(X_n | X_1, \cdots, H_{n-1}) + \sum_{i=1}^{n-1} H(X_i \mid X_{i-1}, \cdots, X_1) && \text{(induction hypothesis)} \end{aligned}$$

2.

$$\begin{aligned} D(p(x)\|q(x)) &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \frac{p(x)}{q(x)} = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \frac{p(x)}{q(x)} = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \frac{p(x \mid y) p(y)}{q(x \mid y) p(y)} \\ &= \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \frac{p(x \mid y)}{q(x \mid y)} + \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} p(x,y) \frac{p(y)}{p(y)} \\ &= D(p(x \mid y)\|q(x \mid y)) + D(p(y)\|q(y)) \end{aligned}$$

3.

$$\begin{aligned} I(X_1, \cdots, X_n; Y) &= H(X_1, \cdots, X_n) - H(X_1, \cdots, X_n \mid Y) \\ &= \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \cdots, X_1) + \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \cdots, X_1, Y) \\ &= \sum_{i=1}^{n} I(X_i; Y \mid X_{i-1}, \cdots, X_1) \end{aligned}$$

□

2 Inequalities for Entropy and Mutual Information

2.1 Convexity and Jensen's Inequality

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

Theorem 4 (Jensen's Inequality)**.** *Let $\mathcal{X} \subset \mathbb{R}$ and $f : \mathcal{X} \to \mathbb{R}$ 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 $\mathbb{E}f(X) \leq f(\mathbb{E}X)$.*
3. *If the inequality is strict we have $\mathbb{E}(X) = X$ almost surely.*

Proof. 1. Let $n \in \mathbb{N} \setminus \{1\}$ and $p_1, \cdots, p_n \in (0,1)$ such that $\sum_{i=1}^n p_i = 1$ and $x_1, \cdots, x_n \in \mathbb{R}$. Distinguish two cases. If $\mathcal{X}$ is finite: We show $f\left(\sum_{i=1}^n p_i x_i\right) \leq \sum_{i=1}^n p_i f(x_i)$ 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: If $\mathcal{X}$ is countably infinite, the claim follows using the limit of the previous case.

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

□

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

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

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

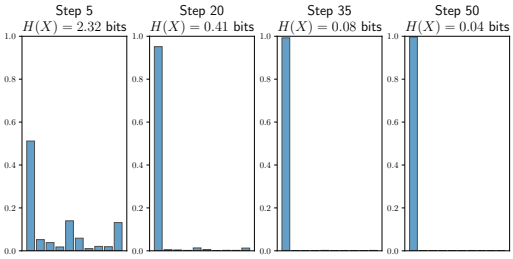
$$\begin{aligned}
-D(p(x)\|q(x)) &= -\mathbb{E}_p\left(\log\frac{p(X)}{q(X)}\right) = \mathbb{E}_p\left(\log\frac{q(X)}{p(X)}\right) \quad \text{(def. of relative entropy)} \\
&\leq \log\left(\mathbb{E}_p\frac{q(X)}{p(X)}\right) \\
&= \log\left(\sum_{x\in\mathcal{X}} p(x)\frac{q(x)}{p(x)}\right) = \log\left(\sum_{x\in\mathcal{X}}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

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

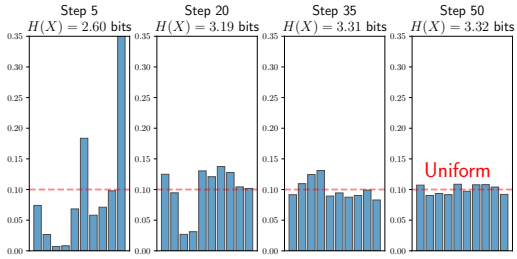
3. Follows from part 2: $I(X;Y) = D(p(x,y)\|p(x)p(y)) \geq 0$.

□

2.2 Advanced Properties of Entropy



(a) Peak Distribution minimises Entropy



(b) Uniform Distribution maximises Entropy

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. Can conditioning on more information increase the entropy?
   Figure 3 illustrates the idea, that conditioning on more information can never increase the entropy. We will prove this exact statement: $H(X \mid Y) \leq H(X)$.

2. What distribution minimises and what distribution maximises the value of Entropy?
   To get a computational intuition for this question, we need to limit our random variables to finite supports. Let $X_p \sim B(p_1, \cdots, p_n)$ be a categorical distribution.
   We can now write first the question as an optimisation problem: $\min_{p \in \mathbb{R}^n, \|p\|_1 = 1, p \geq 0} H(X_p)$. The constraints can be automatically achieved by passing the logits through a *Softmax* function. We start from a random initialisation. Figure 4b shows the solver working with Gradient Descent in Tinygrad [Tin]. A similar approach can answer the second question. Figure 4a shows the process. From the data, it seems plausible that the uniform distribution maximises entropy and a peaked distribution minimises entropy. We will prove these statements.

3. Can joint entropy increase if we add redundant information?
   Let $\Omega = \{1, 2\}$, $\mathbb{P}$ uniform, $X, Y : \Omega \to \mathbb{R}$, $X(\omega) = \omega$ and $Y(\omega) = 2\omega$. Y is redundant to X, as $Y = 2X$. We can now calculate the entropy and joint entropy:

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

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

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

4. 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, $N \sim U(\{0, 1\})$ the noise and let X, N be independent. Then $S := X + N$ is a noisy signal. We expect the entropy to never decrease. We will prove this later.

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

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

2. *A peaked distribution minimises entropy:*
   $H(X) \geq 0$ *and* $H(X) = 0 \iff \exists x \in \mathcal{X} : p_X(x) = 1$.

3. *The uniform distribution maximizes entropy:*
   $H(X) \leq \log |\mathcal{X}|$ *and* $H(X) = \log |\mathcal{X}| \iff X \sim U(\mathcal{X})$.

4. *If information from Y does not add anything to X, then Y must be derived from X:*
   $H(Y \mid X) = 0 \implies \exists f : Y = f(X)$ *almost surely*

5. *If independent noise is added to a random variable, entropy can only increase:*
   Set $Z = X + Y$. Then we have X, Y independent $\implies H(X) \leq H(Z) \wedge H(Y) \leq H(Z)$

6. *The joint entropy of many variables never exceeds the sum of the individual entropies:*
   $H(X_1, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$

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

2. Let X be a discrete random variable with an entropy. Then we have

$$\begin{aligned} H(X) = 0 &\iff \forall x \in \mathcal{X} : p_X(x) \log p_X(x) = 0 \\ &\iff \forall x \in \mathcal{X} : p_X(x) = 0 \oplus p_X(x) = 1 \end{aligned}$$

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

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

This is equivalent to $H(X) \leq \log |\mathcal{X}|$. Lastly, Jensen's Inequality (3) yields the equivalence.

4. We have $H(Y \mid X) = -\mathbb{E} \left(\log p(Y \mid X) \right) = \sum_{(x,y) \in \mathcal{X} \times \mathcal{Y}} -p(x,y) \log p(y \mid x)$.
Additionally, we have $\forall (x,y) \in \mathcal{X} \times \mathcal{Y} : -p(x,y) \log p(y \mid x) \geq 0$.
Combining those facts, we get

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

This tells us that either $p(x,y) = 0$ or $p(x,y) = p(y \mid x)p(x) = p(x) = 1$.
Define $(y_x)_{x \in \mathcal{X}}$ such that $\forall x \in \mathcal{X} : p(x, y_x) > 0$.
Set $f : \operatorname{supp}(X) \to \mathcal{Y}, x \mapsto y_x$. This gets us $Im(f) = \{y_x : x \in \mathcal{X}\} = \{y \in \mathcal{Y} : p(x,y) > 0\}$.
So $Y = f(X)$ almost surely.

5. We have

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

Similarly, $H(Z \mid Y) = H(X \mid Y)$.
Using indep. of X, Y and information never hurts, we get:
$H(X) = H(X \mid Y) = H(Z \mid Y) \leq H(Z)$.
Similarly, we get $H(Y) = H(Y \mid X) = H(Z \mid X) \leq H(Z)$.

6. Using the Chain rule, we have

$$\begin{aligned} H(X_1, \cdots, X_n) &= \sum_{i=1}^{n} H(X_i \mid X_{i-1}, \cdots, X_1) && \text{(chain rule)} \\ &\leq \sum_{i=1}^{n} H(X_i) && \text{(information never hurts)} \end{aligned}$$

□

2.3 The Log-Sum Inequality

Theorem 6 (Log-Sum Inequality)**.** *Let $n \in \mathbb{N}$, $a_1, \cdots, a_n, b_1, \cdots, b_n \geq 0$. Then we have*

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

Proof. TODO □

Theorem 7 (Convexity of Entropy)**.** *1. $D(.\|.)$ is a convex function. This means that*

$$\begin{aligned} \forall p_1, q_1, p_2, q_2, \forall \lambda \in [0,1]: \quad & D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \\ & \leq \lambda D(p_1 \| q_1) + (1-\lambda) D(p_2 \| q_2) \end{aligned}$$

2. Let $X_p \sim B(1,p)$ for all $p \in (0,1)$ and $h(p) := H(X_p)$.
Then $h \leq 1$ and h is a concave function.

Proof. 1. Let p_1, q_1, p_2, q_2 be probability mass functions on $\mathcal{X}$. Let $\lambda \in (0,1)$.
First of all, the set of probability densities is convex. So the above statement is well-defined.
Secondly, the inequality needs to be verified:

$$\begin{aligned} & D(\lambda p_1 + (1-\lambda)p_2 \| \lambda q_1 + (1-\lambda)q_2) \\ &= \sum_{x \in \mathcal{X}} (\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} && \text{(by definition)} \\ &\leq \sum_{x \in \mathcal{X}} \lambda p_1(x) \log \frac{p_1(x)}{q_1(x)} + (1-\lambda) p_2(x) \log \frac{p_2(x)}{q_2(x)} && \text{(log-sum inequality on the 2 two-sums)} \\ &= \lambda D(p_1 \| q_1) + (1-\lambda) D(p_2 \| q_2) && \text{(by definition)} \end{aligned}$$

2. First, we have

$$\begin{aligned} h(p) = H(X_p) &= -p \log p - (1-p) \log(1-p) && \text{(by definition)} \\ &= -(p \log p + (1-p) \log(1-p)) && \text{(factor out)} \\ &\leq -(p + 1 - p) \log \frac{p+1-p}{1+1} = -\log \frac{1}{2} = 1 && \text{(simplify)} \end{aligned}$$

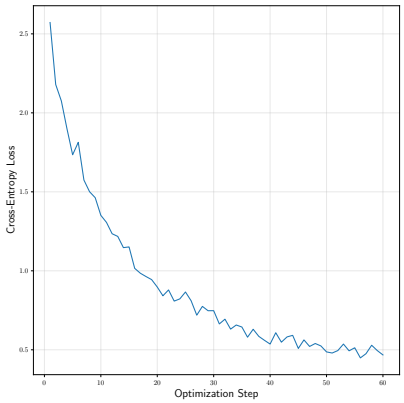
Second, for q with $\forall x \in \mathcal{X} : q(x) = \frac{1}{|\mathcal{X}|}$ we have Let $p_1, p_2 \in [0,1]$ and $\lambda \in (0,1)$.

$$\begin{aligned} h(\lambda p_1 + (1-\lambda)p_2) &= H(X_{\lambda p_1 + (1-\lambda)p_2}) && \text{(by definition)} \\ &= |\mathcal{X}| - D(\lambda p_1 + (1-\lambda)p_2 \| q) && \text{(using theorem 5)} \\ &\geq |\mathcal{X}| - \lambda D(p_1 \| q) - (1-\lambda) D(p_2 \| q) && \text{(using convexity)} \\ &= \lambda \left(|\mathcal{X}| - D(p_1 \| q) \right) - (1-\lambda) \left(|\mathcal{X}| - D(p_2 \| q) \right) && \text{(factor out twice)} \\ &= \lambda H(X_{p_1}) - (1-\lambda) H(X_{p_2}) = \lambda h(p_1) + (1-\lambda) h(p_2) && \text{(by definition, simplify)} \end{aligned}$$

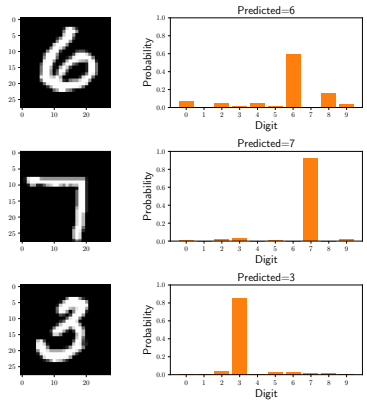
□

Remark 6 (Divergence vs Metric)**.** TODO

2.4 Application to Optimisation



(a) MNIST objective function



(b) MNIST images and pred. digit distr.

Definition 4 (Cross-Entropy)**.** Let p, q be probability mass functions over the set $\mathcal{X}$ and $X \sim p$. Then Cross-Entropy is defined as $\text{CrossEntropy}(p, q) = H(X) + D(p\|q)$.

Theorem 8. *1.* $\min_q \text{CrossEntropy}(p, q) = \min_q D(p\|q)$.

2. $\text{CrossEntropy}(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$.

Proof. 1. Cross-Entropy is simply Relative Entropy with a constant offset. So minimizing with respect to q yields the same value.

2. Let $X \sim p$. Then we have

$$\begin{aligned}\text{CrossEntropy}(p, q) = H(X) + D(p\|q) &= \sum_{x \in \mathcal{X}} -p(x) \log p(x) + \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \\ &= \sum_{x \in \mathcal{X}} -p(x) \log p(x) + p(x) \log p(x) - p(x) \log q(x) \\ &= -\sum_{x \in \mathcal{X}} p(x) \log q(x)\end{aligned}$$

□

Example 5 (MNIST Digit Classification)**.** We can now apply the concept of Relative Entropy to solve a common classification problem from machine learning. Let Ω be the set of 28x28 pixel images that contain exactly one handwritten digit from 0 to 9. The task is to predict the digit 0 to 9, based on the input image $\omega \in \Omega$.
In order to accomplish this task, we define a model $f : \mathbb{R}^n \to \mathbb{R}_+^{10}$ with $n \in \mathbb{N}$.
This function f takes parameters as inputs that allow it to output a probability mass function that indicates the digit in the image. So $\forall \omega \in \Omega : f(\omega) \geq 0 \wedge \sum_{i=1}^{10} f(\omega)_i = 1$.
The optimisation objective is

$$\Phi = \text{argmin}_{\phi \in \mathbb{R}^n} D(d\|f_\phi)$$

which according to Theorem 8 is equivalent to

$$\Phi = \text{argmin}_{\phi \in \mathbb{R}^n} \text{CrossEntropy}(d, f_\phi)$$

As we have shown in Theorem 7, the Relative Entropy and Cross-Entropy are convex functions. For f, we can use a two-layer convolutional neural network with dropout, as detailed in the official PyTorch Python example [Pyt]. The solver is Tinygrad [Tin]. In this case, the model f is not a convex function itself. So the optimisation objective is not convex either. But at least the loss function $\text{CrossEntropy}(d\|.)$ is convex. We can now use Stochastic Gradient Descent to optimise for 60 steps with step size 5e-3 and batch size 512. Figure 5a illustrates the Value of the Objective Function over time and Figure 5b illustrates the outputs of the finished model.

References

[CT05] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 1st ed. Wiley, Sept. 16, 2005. ISBN: 9780471241959 9780471748823. DOI: 10.1002/047174882X. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/047174882X (visited on 12/08/2025).

[Mat] *Can the entropy of a random variable with countably many outcomes be infinite?* URL: https://math.stackexchange.com/questions/279304/can-the-entropy-of-a-random-variable-with-countably-many-outcomes-be-infinite (visited on 08/12/2025).

[Pyt] *MNIST example*. URL: https://github.com/pytorch/examples/blob/main/mnist/main.py (visited on 09/12/2025).

[Tin] *Tinygrad autograd library*. URL: https://github.com/tinygrad/tinygrad (visited on 09/12/2025).