# 1 Information Theory

## 1.1 Entropy

**Definition 1.1** (Entropy)**.** Let $X$ be a discrete random variable taking values in a finite set $\mathcal{X}$ with probability mass function $p(x) = P(X = x)$. The *entropy* of $X$, denoted $H(X)$, is defined as:

$$H(X) \coloneqq -\sum_{x \in \mathcal{X}} p(x) \log p(x),$$

where the logarithm is typically taken base 2 (bits) or base $e$ (nats).

**Remark 1.1.** If $p(x) = 0$, we set $p(x) \log p(x) \coloneqq 0$. This ensures that $p(x) \log p(x)$ is continuous on $[0, 1]$.

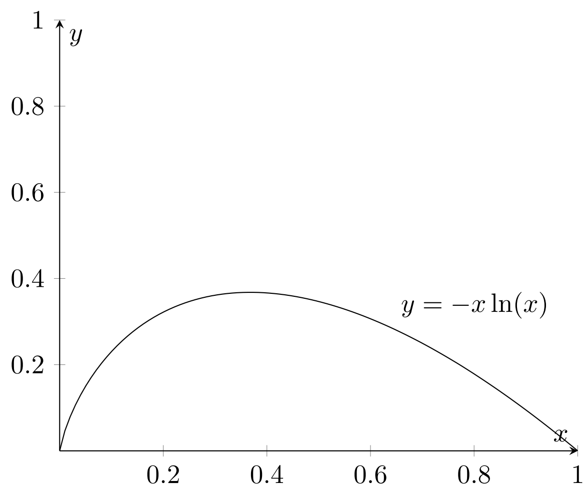


Figure 1: Plot of the function $y = -x \ln(x)$.

**Remark 1.2.** Entropy measures the uncertainty or information content of a random variable. Higher entropy indicates more unpredictability.

**Proposition 1.1** (Non-Negativity of Entropy)**.** *For any discrete random variable $X$, we have $H(X) \geq 0$.*

*Proof.* Since $0 \leq p(x) \leq 1$ and $-\log p(x) \geq 0$, each term in the sum is non-negative, so their total sum is non-negative. $\square$

**Lemma 1.1** (Jensen's Inequality)**.** *Let $X \in \mathcal{X}$ be a random variable over a finite set $\mathcal{X}$, and let $\phi$ be a convex function defined for all $X$. Then:*

$$\phi(E[X]) \leq E[\phi(X)] \quad .$$

*Proof.* We use induction over $n = |\mathcal{X}|$. The base case $n = 1$ is trivial. Hence, assume that the claim holds for some $n$. We now prove the claim for $n + 1$. Clearly, for $n > 1$, we must have $P(X = x_k) < 1$ for some $x_k \in \mathcal{X}$. Without loss of generality, we assume $k = n + 1$. Hence:

$$\begin{aligned}
\phi(E[X]) &= \phi\left(\sum_{i=1}^{n+1} p(x_i)x_i\right) \\
&= \phi\left(\left[(1 - p(x_{n+1}))\sum_{i=1}^{n} \frac{p(x_i)}{1 - p(x_{n+1})}x_i\right] + p(x_{n+1})x_{n+1}\right) \\
&\underset{\text{convexity}}{\leq} (1 - p(x_{n+1}))\phi\left(\sum_{i=1}^{n} \frac{p(x_i)}{1 - p(x_{n+1})}x_i\right) + p(x_{n+1})\phi(x_{n+1}) \\
&\underset{\text{inductive hypothesis}}{\leq} (1 - p(x_{n+1}))\sum_{i=1}^{n} \frac{p(x_i)}{1 - p(x_{n+1})}\phi(x_i) + p(x_{n+1})\phi(x_{n+1}) \\
&= \sum_{i=1}^{n+1} p(x_i)\phi(x_i) = E[\phi(X)] \quad .
\end{aligned}$$

$\square$

**Remark 1.3.** For strictly convex $\phi$, it can be shown that

$$\phi(E[X]) = E[\phi(X)] \text{ is maximized } \iff X \text{ is sampled from a uniform distribution} \quad .$$

**Proposition 1.2** (Maximum Entropy)**.** *For a discrete random variable $X$ over $n$ outcomes, entropy is maximized when $X$ is uniform:*

$$H(X) \leq \log n \quad .$$

*Proof.* We have:

$$\begin{aligned}
-H(X) &= -E[-\log(p(X))] \\
&= E\left[-\log\left(\frac{1}{p(X)}\right)\right] \\
&\underset{\text{Jensen's Inequality}}{\geq} -\log\left(E\left[\frac{1}{p(X)}\right]\right) \\
&= -\log n \quad ,
\end{aligned}$$

where we assumed $p(X) > 0$. Of course, the cases where $p(X) = 0$ follow directly, since $p(X)\log p(X) = 0$.

$H(X) \leq \log n$ follows directly. Note that we have equality iff $X$ has uniform distribution (since $-\log(x)$ is strictly convex). □

### 1.1.1 Joint, Conditional, and Cross Entropy

**Definition 1.2** (Joint Entropy)**.** For a pair of discrete random variables $X$ and $Y$, the joint entropy is:

$$H(X,Y) \coloneqq -\sum_{x,y} p(x,y)\log p(x,y) \quad .$$

**Definition 1.3** (Conditional Entropy)**.** The conditional entropy of $Y$ given $X$ is defined as:

$$H(Y \mid X) \coloneqq \sum_{x} p(x) H(Y \mid X = x) = -\sum_{x,y} p(x,y)\log p(y \mid x).$$

**Corollary 1.1.** *We immediately see from the first equation that $H(Y \mid X) \geq 0$.*

**Theorem 1.1** (Chain Rule for Entropy)**.**

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

*Proof.* We have:

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y) \log p(x,y) \\ &= -\sum_{x,y} p(x,y) \log \left(p(x)p(y \mid x)\right) \\ &= -\sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y \mid x) \\ &= H(X) + H(Y \mid X) \quad . \end{aligned}$$

□

**Corollary 1.2.** $H(X,Y) \geq 0$ *follows directly.*

**Definition 1.4** (Cross-Entropy)**.** Let $p$ and $q$ be two probability distributions over a finite set $\mathcal{X}$, with $p(x) > 0 \Rightarrow q(x) > 0$. The *cross-entropy* of $p$ relative to $q$ is defined as:

$$H_q(p) := -\sum_{x \in \mathcal{X}} p(x) \log q(x) \quad .$$

**Remark 1.4.** Cross-entropy measures the expected number of bits required to encode samples from $p$ using a code optimized for the distribution $q$.

**Remark 1.5.** Cross-entropy is non-negative (see section 1.2).

### 1.1.2 Properties of Entropy

**Proposition 1.3.** *Conditional entropy satisfies:*

$$H(Y \mid X) \leq H(Y) \quad ,$$

*with equality if and only if $X$ and $Y$ are independent.*

*Proof.* From the chain rule:

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

which implies:

$$H(Y \mid X) = H(Y) + H(X \mid Y) - H(X) = H(Y) - I(X;Y) \quad ,$$

with mutual information $I(X;Y) \geq 0$ (see section 1.3). Equality holds if and only if $I(X;Y) = 0$, i.e., $X$ and $Y$ are independent. □

**Corollary 1.3** (Subadditivity of Entropy)**.** *For any two random variables $X$ and $Y$,*

$$H(X,Y) \leq H(X) + H(Y) \quad ,$$

*with equality if and only if $X$ and $Y$ are independent.*

*Proof.* From the chain rule:

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

since $H(Y \mid X) \leq H(Y)$ based on proposition 1.3. Equality holds if and only if $H(Y \mid X) = H(Y)$, i.e., $X$ and $Y$ are independent. □

**Theorem 1.2** (Concavity of Entropy)**.** *The entropy function $H(p)$, where $p \in \Delta$ is a probability vector, is concave on the probability simplex $\Delta$.*

*Proof.* This follows from the fact that $f(x) = -x \log x$ is concave for $x \in [0, 1]$, and entropy is the sum of such terms. Therefore, for every convex combination $p = \lambda p_1 + (1 - \lambda)p_2$:

$$H(p) \geq \lambda H(p_1) + (1 - \lambda)H(p_2) \quad .$$

□

**Summary of Key Properties**

- Non-negativity: $H(X) \geq 0$
- Maximum entropy: $H(X) \leq \log |\mathcal{X}|$
- Chain rule: $H(X,Y) = H(X) + H(Y \mid X)$
- Subadditivity: $H(X,Y) \leq H(X) + H(Y)$
- Conditioning reduces entropy: $H(Y \mid X) \leq H(Y)$
- Concavity: $H(p)$ is concave in the distribution $p$

## 1.2 Kullback-Leibler Divergence

**Definition 1.5** (KL Divergence)**.** Let $P$ and $Q$ be two discrete probability distributions over the same finite set $\mathcal{X}$, with $P(x) > 0 \Rightarrow Q(x) > 0$. The Kullback-Leibler divergence (or relative entropy) from $P$ to $Q$ is defined as:

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &:= \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \\ &= -\sum_{x} P(x) \log Q(x) + \sum_{x} P(x) \log P(x) \\ &= H_Q(P) - H(P) \quad . \end{aligned}$$

**Remark 1.6.** If $P(x) = Q(x) = 0$, we set $P(x) \log \dfrac{P(x)}{Q(x)} := 0$.

**Remark 1.7.** KL divergence measures the inefficiency of assuming that the distribution is $Q$ when the true distribution is $P$. It is not a metric: it is not symmetric and does not satisfy the triangle inequality.

**Lemma 1.2** (Gibb's Inequality)**.** *Suppose that* $P = \{p_1, \ldots, p_n\}$ *and* $Q = \{q_1, \ldots, q_n\}$ *are discrete probability distributions. Then:*

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i \quad .$$

*Proof.* The claim is equivalent to $\sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log q_i \geq 0$. We have:

$$\begin{aligned} \sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log q_i &= \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \\ &= \sum_{i=1}^{n} p_i \left( -\log \frac{q_i}{p_i} \right) \\ &\underset{\text{Jensen's Inequality}}{\geq} -\log \left( \sum_{i=1}^{n} p_i \frac{q_i}{p_i} \right) \\ &= -\log(1) = 0 \quad . \end{aligned}$$

$\square$

**Corollary 1.4.** *It directly follows from the proof that* $D_{\mathrm{KL}}(P\|Q) \geq 0$ *and* $0 \leq H(P) \leq H_Q(P)$.

**Proposition 1.4** (Additivity)**.** *Let $P = P_1 \times P_2$, $Q = Q_1 \times Q_2$. Then:*

$$D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}(P_1\|Q_1) + D_{\mathrm{KL}}(P_2\|Q_2) \quad .$$

*Proof.*

$$\begin{aligned} D_{\mathrm{KL}}(P_1 \times P_2\|Q_1 \times Q_2) &= \sum_{x,y} P_1(x)P_2(y) \log \frac{P_1(x)P_2(y)}{Q_1(x)Q_2(y)} \\ &= \sum_{x,y} P_1(x)P_2(y) \left( \log \frac{P_1(x)}{Q_1(x)} + \log \frac{P_2(y)}{Q_2(y)} \right) \\ &= \sum_{x} P_1(x) \log \frac{P_1(x)}{Q_1(x)} + \sum_{y} P_2(y) \log \frac{P_2(y)}{Q_2(y)} \\ &= D_{\mathrm{KL}}(P_1\|Q_1) + D_{\mathrm{KL}}(P_2\|Q_2) \quad . \end{aligned}$$

□

**Proposition 1.5** (Entropy Representation via KL Divergence)**.** *Let $U$ be the uniform distribution over $\mathcal{X}$, where $|\mathcal{X}| = n$. Then for any distribution $P$,*

$$H(P) = \log n - D_{\mathrm{KL}}(P\|U) \quad .$$

*Proof.*

$$\begin{aligned} D_{\mathrm{KL}}(P\|U) &= \sum_{x} P(x) \log \frac{P(x)}{1/n} = \sum_{x} P(x) \log P(x) + \sum_{x} P(x) \log n \\ &= -H(P) + \log n \quad . \end{aligned}$$

□

**Summary of Key Properties**

- $D_{\mathrm{KL}}(P\|Q) \geq 0$
- $D_{\mathrm{KL}}(P\|Q) = 0 \iff P = Q$
- Asymmetric: $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$
- Additive over independent distributions
- Connection to entropy: $H(P) = \log n - D_{\mathrm{KL}}(P\|U)$

## 1.3 Mutual Information

**Definition 1.6** (Mutual Information)**.** Let $X$ and $Y$ be discrete random variables with joint distribution $p(x, y)$ and marginals $p(x)$, $p(y)$. The *mutual information* between $X$ and $Y$ is defined as:

$$I(X;Y) := \sum_{x,y} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right) \quad .$$

**Remark 1.8.** Mutual information quantifies how much knowing $X$ reduces uncertainty about $Y$, and vice versa. Per definition, it is symmetric: $I(X;Y) = I(Y;X)$.

**Proposition 1.6** (Equivalent Expressions)**.** *Mutual information can also be expressed as:*

$$\begin{aligned}
I(X;Y) &= D_{\mathrm{KL}}(p(x,y) \parallel p(x)p(y)) \\
&= H_{p(x)p(y)}(p(x,y)) - H(X,Y) \\
&= \left[ -\sum_{x,y} p(x,y) \log(p(x)p(y)) \right] - H(X,Y) \\
&= H(X) + H(Y) - H(X,Y) \\
&= H(X) - H(X \mid Y) \\
&= H(Y) - H(Y \mid X)
\end{aligned}$$

*Proof.* Each follows from basic entropy identities and the definition of KL divergence. □

**Corollary 1.5.** $I(X;Y) \geq 0$*, since* $I(X;Y) = D_{\mathrm{KL}}(p(x,y) \| p(x)p(y))$ *and KL divergence is always non-negative.*

**Definition 1.7** (Conditional Mutual Information)**.** Let $X, Y, Z$ be discrete random variables. The *conditional mutual information* of $X$ and $Y$ given $Z$ is defined as:

$$I(X;Y \mid Z) := \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)} \quad .$$

Equivalently, in terms of entropy:

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid (Y,Z)) \quad .$$

*Proof.*

$$
\begin{aligned}
&H(X \mid Z) - H(X \mid (Y,Z)) \\
&= \sum_z p(z) H(X \mid Z = z) - \sum_{y,z} p(y,z) H(X \mid Y = y, Z = z) \\
&= -\sum_z p(z) \sum_x p(x \mid z) \log p(x \mid z) \;+\; \sum_{y,z} p(y,z) \sum_x p(x \mid y,z) \log p(x \mid y,z) \\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x \mid y,z)}{p(x \mid z)} \\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \\
&= I(X;Y \mid Z) \quad .
\end{aligned}
$$

□

**Remark 1.9.** Conditional mutual information measures how much knowing $Y$ reduces the uncertainty of $X$, *given* that we already know $Z$.

**Proposition 1.7** (Chain Rule for Mutual Information)**.** *Let $X$, $Y$, and $Z$ be random variables. Then:*

$$I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z) \quad .$$

*Proof.* We use entropy-based expressions for mutual information:

$$
\begin{aligned}
I(X;Y,Z) &= H(X) - H(X \mid (Y,Z)) \\
&= I(X;Z) + H(X \mid Z) - H(X \mid (Y,Z)) \\
&= I(X;Z) + H(X \mid Z) - (H(X \mid Z) - I(X;Y \mid Z)) \\
&= I(X;Z) + I(X;Y \mid Z) \quad .
\end{aligned}
$$

□

**Proposition 1.8** (Non-Negativity of Conditional Mutual Information)**.** *It holds true that*

$$I(X;Y \mid Z) \geq 0 \quad .$$

*Proof.* We have:

$$
\begin{aligned}
I(X;Y \mid Z) &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \\
&= \sum_z p(z) \sum_{x,y} p(x,y \mid z) \log \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \\
&= \sum_z p(z) D_{\mathrm{KL}} \left( p(x,y \mid z) \,\|\, p(x \mid z) p(y \mid z) \right) \geq 0 \quad .
\end{aligned}
$$

□

**Corollary 1.6.** *As a direct consequence, we have*

$$I(X;Z) \leq I(X;Y,Z) \quad .$$

**Definition 1.8** (Conditional Independence)**.** Let $X, Y, Z$ be discrete random variables. We say that $X$ is *conditionally independent* of $Z$ given $Y$, and write:

$$X \perp Z \mid Y$$

if and only if

$$p(z \mid x, y) = p(z \mid y) \quad \text{for all } x, y, z \quad .$$

Equivalently:

$$p(x, z \mid y) = p(x \mid y)p(z \mid y) \quad .$$

**Proposition 1.9.** *If $X \perp Z \mid Y$, then the conditional mutual information between $X$ and $Z$ given $Y$ is zero:*

$$I(X;Z \mid Y) = 0 \quad .$$

*Proof.* By definition of conditional mutual information:

$$I(X;Z \mid Y) = \sum_{x,z,y} p(x,z,y) \log \frac{p(x,z \mid y)}{p(x \mid y)p(z \mid y)} \quad .$$

If $X \perp Z \mid Y$, then:

$$p(x, z \mid y) = p(x \mid y)p(z \mid y) \quad ,$$

so the logarithm becomes:

$$\log \frac{p(x \mid y)p(z \mid y)}{p(x \mid y)p(z \mid y)} = \log 1 = 0 \quad .$$

Hence, each term in the sum is zero, and:

$$I(X;Z \mid Y) = 0 \quad .$$

□

### 1.3.1 Data Processing Inequality

**Lemma 1.3** (Markov Chain)**.** *Let $X, Y, Z$ be random discrete random variables forming the Markov chain $X \to Y \to Z$. Then:*

$$X \perp Z \mid Y \quad .$$

*Proof.* Per definition from Markov chains, we have:

$$p(z \mid x, y) = p(z \mid y) \quad ,$$

and hence $X \perp Z \mid Y$. □

**Theorem 1.3** (Data Processing Inequality)**.** *If $X \to Y \to Z$ is a Markov chain, then:*

$$I(X;Z) \leq I(X;Y) \quad .$$

*Proof.* We use the chain rule and conditional independence:

$$\begin{aligned} I(X;Z) &= I(X;Z,Y) - I(X;Y \mid Z) \\ &= I(X;Y) + I(X;Z \mid Y) - I(X;Y \mid Z) \quad . \end{aligned}$$

Since $X \to Y \to Z$, we have $I(X;Z \mid Y) = 0$. Thus:

$$I(X;Z) = I(X;Y) - I(X;Y \mid Z) \leq I(X;Y) \quad ,$$

because $I(X;Y \mid Z) \geq 0$. □

**Corollary 1.7** (No Gain in Processing)**.** *Any function $f(Y)$ of $Y$ cannot increase information about $X$:*

$$I(X;f(Y)) \leq I(X;Y) \quad .$$

*Proof.* This follows by applying the DPI to the chain $X \to Y \to f(Y)$. □

#### Summary of Key Properties

- $I(X;Y) \geq 0$
- $I(X;Y) = 0$ if and only if $X \perp Y$
- $I(X;Y) = D_{\mathrm{KL}}(p(x,y) \| p(x)p(y))$
- Chain rule: $I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z)$
- Data Processing Inequality: $X \to Y \to Z \Rightarrow I(X;Z) \leq I(X;Y)$

## 1.4 Bounding Mutual Information via Matrix Rank of the Joint Distribution

**Theorem 1.4.** *Let $X, Y$ be random variables from finite sets $\mathcal{X}, \mathcal{Y}$, and let matrix $\boldsymbol{P}$ denote their joint probability distribution, i.e. $\boldsymbol{P}_{ij} = p(x_i, y_j)$. Let $r \coloneqq \operatorname{rank} \boldsymbol{P}$ denote the rank of matrix $\boldsymbol{P}$. Then we have*

$$I(X;Y) \leq \log r \quad .$$

*Proof.* Let $n \coloneqq |\mathcal{X}|$ and $m \coloneqq |\mathcal{Y}|$. If $\boldsymbol{P}$ has rank $r$, then so must the transition matrix $\boldsymbol{P}_{Y|X} \in \mathbb{R}^{m \times n}$ defined as $(\boldsymbol{P}_{Y|X})_{ij} \coloneqq p(y_i \mid x_j) = \frac{p(x_j, y_i)}{\sum_k p(x_k, y_i)}$, since $\boldsymbol{P}_{Y|X}$ is created from $\boldsymbol{P}$ by transposing and column scaling. If one column consisted of only zeros, i.e. $\sum_k p(x_k, y_i) = 0$, we may just copy a different scaled column vector to this column.

Now, let's analyze matrix $\boldsymbol{P}_{Y|X}$. First, note that it is a Markov chain transition matrix, and hence all its columns lie in the m-dimensional unit simplex. Consider the convex hull of the column vectors, it is a r-dimensional convex polytope in the m-dimensional unit simplex. Thus, we can find a r-dimensional simplex with corners collected by matrix $\boldsymbol{U}$ s.t. it is a superset of this polytope and still a subset of the (potentially) higher dimensional unit simplex.

Thus, every column vector in $\boldsymbol{P}_{Y|X}$ can be written as a convex combination of the column vectors in $\boldsymbol{U}$. It follows that $\boldsymbol{P}_{Y|X}$ can be decomposed as

$$\boldsymbol{P}_{Y|X} = \boldsymbol{U}\boldsymbol{V} \ , \quad \boldsymbol{U} \in \mathbb{R}^{m \times r}, \boldsymbol{V} \in \mathbb{R}^{r \times n} \quad ,$$

where both $\boldsymbol{U}$ and $\boldsymbol{V}$ are Markov chain transition matrices as well.

Hence, we can introduce a latent variable $Z \in \{1, \dots, r\}$, which forms the Markov chain

$$X \underset{\boldsymbol{V}}{\rightarrow} Z \underset{\boldsymbol{U}}{\rightarrow} Y \quad .$$

Finally, based on theorem 1.3 it follows that

$$I(X;Y) \leq I(X;Z) = H(Z) - H(Z \mid X) \leq H(Z) \leq \log r \quad .$$

$\square$