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 $[0, 1]$.

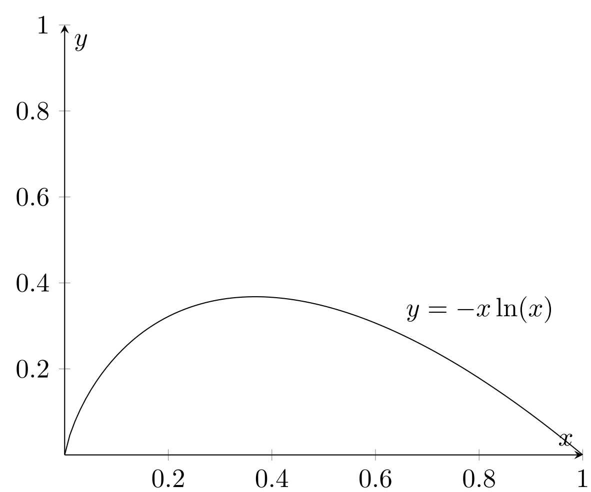

Figure 1: Plot of the function $y = -x \ln(x)$.

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 ϕ 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:

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

□

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 if X has uniform distribution. □

1.1.1 Joint Entropy and Conditional 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.*

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,*

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

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 is a probability vector, is concave on the probability simplex.*

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

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

□

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 1.4 (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:

$$D_{\mathrm{KL}}(P\|Q) = \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

Remark 1.3. If $P(x) = Q(x) = 0$, we set $P(x) \log \dfrac{P(x)}{Q(x)} := 0$.

Remark 1.4. 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:

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

□

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

Remark 1.5 (Asymmetry)**.** In general,

$$D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P) \quad .$$

To see this, let $\mathcal{X} = \{0, 1\}$, $P = (0.9, 0.1)$, $Q = (0.5, 0.5)$. Then:

$$D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P) \quad .$$

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 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
- Connects with entropy: $H(P) = \log n - D_{\mathrm{KL}}(P\|U)$

1.3 Mutual Information

Definition 1.5 (Mutual Information)**.** The *mutual information* between two discrete random variables X and Y is defined as:

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

Lemma 1.3. *Mutual information is symmetric: $I(X;Y) = I(Y;X)$.*

Lemma 1.4. *Mutual information is non-negative: $I(X;Y) \geq 0$.*

Proof. Follows from the non-negativity of KL divergence:

$$I(X;Y) = D_{\mathrm{KL}}(p(x,y) \,\|\, p(x)p(y)) \geq 0.$$

□

Theorem 1.3 (Chain Rule for Mutual Information)**.**

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

Definition 1.6 (Conditional Mutual Information)**.**

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

Theorem 1.4 (Relation Between Entropy and Mutual Information)**.**

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

Proof. Apply the chain rule in both directions:

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

□

2 Subadditivity of Mutual Information

Theorem 2.1 (Subadditivity over Pairwise Mutual Information)**.** *Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be discrete random variables. Then:*

$$\sum_{i=1}^{m} \sum_{j=1}^{n} I(X_i; Y_j) \leq I(X_1, \ldots, X_m; Y_1, \ldots, Y_n).$$

Proof. Each $I(X_i; Y_j) = D_{\mathrm{KL}}(p_{X_i,Y_j} \,\|\, p_{X_i} p_{Y_j})$, and since KL divergence is jointly convex and marginalizing reduces information, we have:

$$\sum_{i,j} D_{\mathrm{KL}}(p_{X_i,Y_j} \,\|\, p_{X_i} p_{Y_j}) \leq D_{\mathrm{KL}}(p_{\mathbf{X},\mathbf{Y}} \,\|\, p_{\mathbf{X}} p_{\mathbf{Y}}) = I(\mathbf{X}; \mathbf{Y}).$$

□