Information Theory and Communication

Ritwik Banerjee

rbanerjee@cs.stonybrook.edu
Upper Bound on Entropy

Theorem

\[ H(X) \leq \log |\mathcal{X}|, \text{ where } |\mathcal{X}| \text{ denotes the number of elements in the range of } X, \text{ with equality if and only if } X \text{ has a uniform distribution over } \mathcal{X}. \]

Proof.

Consider the uniform distribution over \( \mathcal{X} \), given by the probability mass function \( u(x) = \frac{1}{|\mathcal{X}|} \), and let \( p(x) \) denote the probability mass function for \( X \). Then,

\[
D(p||u) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{u(x)} = \log |\mathcal{X}| - H(X)
\]

Since \( H(X) \) is always non-negative, \( D(p||u) \leq \log |\mathcal{X}|. \) \( \square \)
More information can’t hurt!

Theorem

*Conditioning reduces entropy:* \( H(X|Y) \leq H(X) \).

Proof.

\[
H(X) - H(X|Y) = I(X;Y) \geq 0,
\]

and thus, \( H(X|Y) \leq H(X) \). Equality exists if and only if \( X \) and \( Y \) are independent.

- This theorem says that knowing another random variable \( Y \) can only reduce the uncertainty in \( X \).
- This is true only on the average. Mathematically, \( H(X|Y = y) \) may be greater than or less than or equal to \( H(X) \), but on the average

\[
H(X|Y) = \sum_y p(y)H(X|Y = y) \leq H(X).
\]

- For example, in a court case, specific new evidence might increase uncertainty, but on the average, evidence decreases uncertainty.
More information can’t hurt!

Example

Consider the following joint distribution of $X$ and $Y$:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>$\frac{3}{4}$</td>
</tr>
<tr>
<td>2</td>
<td>$\frac{1}{8}$</td>
<td>$\frac{1}{8}$</td>
</tr>
</tbody>
</table>

- $H(X) = H(\frac{1}{8}, \frac{7}{8}) = 0.544$ bits.
- $H(X|Y = 1) = 0$ bits.
- $H(X|Y = 2) = 1$ bit.
- $H(X|Y) = \frac{3}{4} H(X|Y = 1) + \frac{1}{4} H(X|Y = 2) = 0.25$ bits.

Thus, the uncertainty in $X$ is increased if $Y = 2$ is observed and decreased if $Y = 1$ is observed, but uncertainty decreases on the average.
Theorem

Let $X_1, X_2, \ldots, X_n$ be random variables drawn according to the joint distribution $p(x_1, x_2, \ldots, x_n)$. Then,

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i).$$

Proof.

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i | X_1, \ldots, X_{i-1})$$

$$\leq \sum_{i=1}^{n} H(X_i).$$

Equality exists if and only if the $X_i$ random variables are all independent. \qed
Log-Sum Inequality

Theorem

For non-negative numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$,

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

Again, we follow the convention that $0 \log 0 = 0$, $a \log \frac{a}{0} = \infty$ if $a > 0$ and $0 \log \frac{0}{0} = 0$.

Proof.

(Outline)

- We can assume without loss of generality that $a_i, b_i > 0$ for all $1 \leq i \leq n$.
- $f(t) = t \log t$ is a strictly convex function. Use this fact with Jensen’s inequality to get the log-sum inequality.
Consequences of log-sum inequality

- Non-negativity of KL divergence follows immediately. \textit{(left as exercise)}
- Other consequences that follow are:
  - Relative Entropy (i.e., KL divergence) is a convex function.
  - Entropy is a concave function.
  - Mutual information $I(X; Y)$ is a convex function of $p(y|x)$ for a fixed $p(x)$.
  - Mutual information $I(X; Y)$ is a concave function of $p(x)$ for a fixed $p(y|x)$. 
KL Divergence is a convex function

**Theorem**

$D(p||q)$ is a convex function of pairs of probability mass functions. That is, if $p_1, q_1$ and $p_2, q_2$ are two pairs of probability mass functions, then

$$D(p_3||q_3) \leq \lambda D(p_1||q_1) + (1 - \lambda) D(p_2||q_2) \forall 0 \leq \lambda \leq 1$$

where $p_3 = \lambda p_1 + (1 - \lambda) p_2$ and $q_3 = \lambda q_1 + (1 - \lambda) q_2$.

**Proof.**

Application of log-sum inequality. Left as exercise. \qed
Theorem

\( H(p) \) is a concave function of \( p \).

Proof.

(Outline)

- Define two random variables \( X_1 \) and \( X_2 \) with distributions \( p_1 \) and \( p_2 \), taking values on the same set \( \mathcal{X} \).
- Then consider a third random variable that is identical to \( X_1 \) with probability \( \lambda \) and identical to \( X_2 \) with probability \( 1 - \lambda \).
Theorem

Let $X$ and $Y$ have the joint distribution $p(x, y)$. Their mutual information $I(X; Y)$ is a concave function of $p(x)$ for a fixed $p(y|x)$, and a convex function of $p(y|x)$ for a fixed $p(x)$.

Proof.

- For the first part, expand the definition of mutual information in terms of the difference between entropy and conditional entropy.
- Skip the proof of the second part.
Markov Chain

Definition

Random variables $X, Y, Z$ are said to form a **Markov chain** in that order (denoted by $X \rightarrow Y \rightarrow Z$) if the conditional distribution of $Z$ depends only on $Y$ and is conditionally independent of $X$. Which means, $X, Y, Z$ form a Markov chain $X \rightarrow Y \rightarrow Z$ if

$$p(x, y, z) = p(x)p(y|x)p(z|y).$$
Data-Processing Inequality

This theorem tells us that no amount of processing of $Y$ can increase the amount of information it contains about $X$.

**Theorem**

*If $X \rightarrow Y \rightarrow Z$, then $I(X; Y) \geq I(X; Z)$.***