Chapter 1

Shannon's Information Measures

Background

Probability

- Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets, also called *alphabets*.
- Let X and Y be discrete random variables taking values in $\mathcal{X}$ and $\mathcal{Y}$, respectively.
- Probability mass function: $p_X(x) = \Pr\{X = x\}$, $x \in \mathcal{X}$.
- We also denote the probability distribution by p rather than p_X when the random variable referred to is clear from context.
- Joint distribution: $p(x, y) = \Pr\{X = x, Y = y\}$.
- Conditional distribution: $p(x|y) = \frac{p(x,y)}{p(y)}$.
- If $(X, Y) \sim p(x, y)$ are independent, $p(x, y) = p(x)p(y)$ for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$.

1.1 Entropy

What is information

- Information is about uncertainty.
- Entropy is a measure of the uncertainty of a random variable.
- Entropy arises naturally as the fundamental limits of *source coding*.

Entropy

Definition 1.1 The *entropy* $H(X)$ of a discrete random variable X is defined by

$$H(X) = -\sum_x p(x) \log p(x).$$

1. The summation is over the support of X.
2. The log is to the base 2 and the unit of entropy is *bit*.
3. $H(X)$ depends only on $p(x)$, not on the actual values of x—entropy is independent of the alphabet $\mathcal{X}$. So we also write $H(X)$ as $H(p)$.

Other Forms

- Expectation form $H(X) = -\mathbb{E}\log(p(X))$
- Binary entropy function: $H(p) = -p \log p - (1 - p) \log(1 - p)$

Draw a plot of the binary entropy function.

The definiton of entropy matches some intuitions about information. First, fixed event has no information. Second, an event that is totally unexpected has the lagest information.

Properties

- $H(X) \geq 0$ where equality holds iff X is a deterministic.
- $H(X) \leq \log |\mathcal{X}|$ where $\mathcal{X}$ is the alphabet of X. The equality holds iff X is uniformly distributed on $\mathcal{X}$.

Joint Entropy

- The entropy of a pair of random variables (X, Y) with alphabets $\mathcal{X}$ and $\mathcal{Y}$ is also defined by considering (X, Y) as a single random variable over $\mathcal{X} \times \mathcal{Y}$. For convenience, we write $H(X, Y) = H((X, Y))$.
- The joint entropy $H(X, Y)$ of a pair of discrete random variable (X, Y) with a joint distribution $p(x, y)$ is defined as

$$H(X, Y) = -\sum_x \sum_y p(x, y) \log p(x, y) = -\mathbb{E} \log p(X, Y).$$

1.2 Conditional Entropy and Mutual Information

Conditional Entropy

- For random variables X and Y, the *conditional entropy* $H(Y|X)$ is defined as

$$H(Y|X) = -\sum_{x,y} p(x, y) \log p(y|x) = -\mathbb{E} \log p(Y|X).$$

- Denote

$$H(Y|X = x) = H(p_{Y|X}(\cdot|x)) = -\sum_y p(y|x) \log p(y|x).$$

- We can write

$$H(Y|X) = \sum_x p(x) H(Y|X = x).$$

- In other words, the conditional entropy is the expectation of the entropy of the conditional distribution of Y given $X = x$.

Basic Properties

- $H(Y|X) \geq 0$ with equality iff Y is a function of X (over the support of X).
- (Chain rule) $H(X, Y) = H(X) + H(Y|X)$.
- $H(Y|X) \leq H(Y)$ with equality iff X and Y are independent. In other words, conditioning reduces entropy.

Mutual Information

Definition 1.2 The *mutual information* between random variables X and Y is defined as

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

R

1. $I(X; Y)$ is symmetrical in X and Y.
2. $I(X; X) = H(X)$: observing X can get all the information of X.
3. $I(X; Y) \geq 0$ (Log-sum inequality).
4. $I(X; Y)$ only depends on the joint distribution $p_{X,Y}$, so we also write $I(X; Y) = I(p_{X,Y})$.

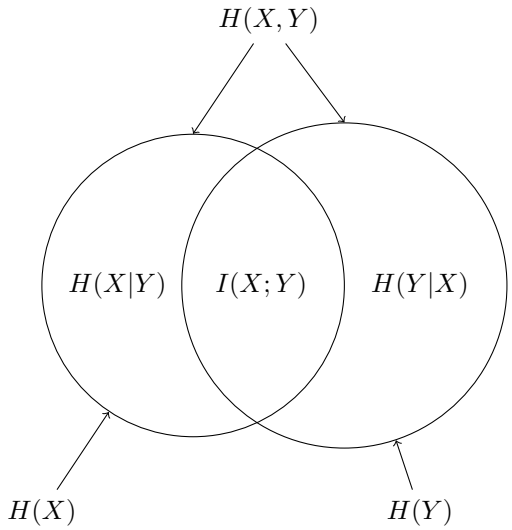
Relations

- We have the following equalities:

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

- If the alphabets are not finite, the above equalities hold provided that all the entropies and conditional entropies are finite.

Information Diagram of Two Random Variables



■ **Example 1.1** Let X and Y have the following joint distribution: Calculate $H(X)$, $H(Y)$, $H(X,Y)$,

Y \ X	1	2	3	4
1	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{32}$
2	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{1}{32}$	$\frac{1}{32}$
3	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$	$\frac{1}{16}$
4	$\frac{1}{4}$	0	0	0

$H(Y|X)$ and $H(X|Y)$. Verify that $H(X) + H(Y|X) = H(Y) + H(X|Y) = H(X,Y)$, and $H(X) - H(X|Y) = H(Y) - H(Y|X)$. ■

1.3 Chain Rules

Theorem 1.1

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

Proof.

$$\begin{aligned} H(X,Y|Z) &= -\mathbb{E}\log p(X,Y|Z) \\ &= -\mathbb{E}\log p(X|Z)p(Y|X,Z) \\ &= -\mathbb{E}[\log p(X|Z) + \log p(Y|X,Z)] \\ &= -\mathbb{E}\log p(X|Z) - \mathbb{E}\log p(Y|X,Z) \\ &= H(X|Z) + H(Y|X,Z). \end{aligned}$$

■

Recall the chain rule of probability:

$$p(x_1, x_2, \ldots, x_n) = p(x_1)p(x_2|x_1)\ldots p(x_i|x_1,\ldots,x_{i-1})\ldots p(x_n|x_1,\ldots,x_{n-1}).$$

Theorem 1.2 — Chain rule for entropy.

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

Proof.

$$H(X_1, X_2, \ldots, X_n) = H(X_1, X_2, \ldots, X_{n-1}) + H(X_n|X_1, X_2, \ldots, X_{n-1}).$$

■

Using chain rule for entropy and conditioning reduces entropy, we can obtain the independence bound on entropy (which will be formally proved later):

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

with equality iff X_i are independent.

Chain Rule for Conditional Entropy

Theorem 1.3

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

Conditional Mutual Information

Definition 1.3 The *conditional mutual information* of random variables X and Y given Z is defined by

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

Analogous to conditional, write

$$I(X;Y|Z) = \sum_{z} p(z)I(X;Y|Z=z)$$

where

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

Theorem 1.4 — Chain rule for mutual information.

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

Proof. Apply chain rules for entropy. ■

Theorem 1.5 — Chain rule for conditional mutual information.

$$I(X_1, X_2, \ldots, X_n; Y | Z) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1, Z).$$

1.4 Information Divergence

Relative Entropy

Definition 1.4 The *relative entropy* (*information divergence* or *Kullback-Leibler distance*) between two probability mass function $p(x)$ and $q(x)$ is defined as

$$D(p\|q) = \sum_x p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_p \log \frac{p(X)}{q(X)},$$

where we adopt the convention that $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$.

R

- $I(X;Y) = D(p(x,y)\|p(x)p(y))$.
- $D(p\|q) \geq 0$ with equality iff $p = q$.
- $D(p\|q)$ is not symmetric, i.e., we do not have $D(p\|q) = D(q\|p)$ in general.
- $D(p\|q)$ does not satisfy the triangle inequality.

Example

Consider two binary distributions p and q on $\{0, 1\}$. Let $p(1) = r$ and $q(1) = s$. Calculate $D(p||q)$ and $D(q||p)$. When they are the same?

Convexity

- $D(p\|q)$ is convex in the pair (p, q), which implies
- $H(p)$ is a concave function of p, and
- $I(X;Y)$ is 1) a concave function of $p(x)$ for fixed $p(y|x)$ and is 2) a convex function of $p(y|x)$ for fixed $p(x)$.

Theorem 1.6 — Chain rule for relative entropy.

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)).$$

Proof.

$$\begin{aligned} D(p(x,y)||q(x,y)) &= \mathbb{E}_p \log \frac{p(X,Y)}{q(X,Y)} \\ &= \mathbb{E}_p \log \frac{p(X)p(Y|X)}{q(X)q(Y|X)} \\ &= \mathbb{E}_p \left[\log \frac{p(X)}{q(X)} + \log \frac{p(Y|X)}{q(Y|X)} \right] \\ &= \mathbb{E}_p \log \frac{p(X)}{q(X)} + \mathbb{E}_p \log \frac{p(Y|X)}{q(Y|X)} \end{aligned}$$

■

1.5 Basic Inequalities

Lemma 1.7 — Fundamental Inequality. For any $a > 0$,

$$\ln a \leq a - 1$$

with equality if and only if $a = 1$.

Lemma 1.8 — Log-sum inequality. For arbitrary non-negative numbers a_i, b_i, $i = 1, \ldots, n$ we have

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b}$$

where $a = \sum_i a_i$ and $b = \sum_i b_i$. The equality holds iff $a_i b = b_i a$ for $i = 1, \ldots, n$.

Proof.
- We may assume that a_i are positive and b_i are positive.
- Further, it is sufficient to prove the lemma for $a = b$.
- For this case, the statement becomes $\sum_i a_i \log \frac{b_i}{a_i} \leq 0$ and follows from the inequality $\ln x \leq x - 1$.

■

Theorem 1.9 — Information inequality. Let p and q be two PMF over the same alphabet. Then

$$D(p||q) \geq 0$$

with equality iff $p(x) = q(x)$ for all x.

Corollary 1.10 For any two random variables X and Y,

$$I(X;Y) \geq 0$$

with equality iff X and Y are independent.

Corollary 1.11

$$I(X;Y|Z) \geq 0,$$

with equality iff X and Y are conditional independent given Z.

Theorem 1.12 $H(X) \leq \log |\mathcal{X}|$ where $\mathcal{X}$ is the alphabet of X. The equality holds iff X is uniformly distributed on $\mathcal{X}$.

Proof. Let u be the uniform distribution on $\mathcal{X}$, i.e., $u(x) = |\mathcal{X}|^{-1}$, $x \in \mathcal{X}$. Then

$$\begin{aligned}
\log |\mathcal{X}| - H(X) &= \sum_x p(x) \log |\mathcal{X}| + \sum_x p(x) \log p(x) \\
&= \sum_x p(x) \log 1/u(x) + \sum_x p(x) \log p(x) \\
&= \sum_x p(x) \log \frac{p(x)}{u(x)} \\
&\geq 1 \cdot \log \frac{1}{1} \qquad (1.1) \\
&= 0,
\end{aligned}$$

where inequality follows from the log-sum inequality. The equality in (1.1) holds if and only $p(x) = u(x)$ for all $x \in \mathcal{X}$. ■

Theorem 1.13 — Conditioning reduces entropy.

$$H(X|Y) \leq H(X)$$

with equality iff X and Y are independent.

Theorem 1.14 — Independence bound on entropy.

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

with equality iff X_i are independent.

1.6 Convextiy of Information Measures

Theorem 1.15 — Convexity of relative entropy. $D(p||q)$ is convex in the pair (p, q).

Hint: The theorem says that if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions, then

$$D(\lambda p_1 + \bar{\lambda} p_2 || \lambda q_1 + \bar{\lambda} q_2) \leq \lambda D(p_1||q_1) + \bar{\lambda} D(p_2||q_2)$$

for all $0 \leq \lambda \leq 1$.

Theorem 1.16 — Concavity of entropy. $H(p)$ is a concave function of p.

Proof. $H(p) = \log |\mathcal{X}| - D(p||u)$ ■

where u is the uniform distribution on $|\mathcal{X}|$.

Theorem 1.17 The mutual information $I(X;Y)$ is a concave function of $p(x)$ for fixed $p(y|x)$ and a convex function of $p(y|x)$ for fixed $p(x)$.

Proof. To prove the first part, we write

$$I(X;Y) = H(Y) - H(Y|X) = H(Y) - \sum_x p(x) H(Y|X = x).$$

To prove the second part, we use $I(X,Y) = D(p(x,y)||p(x)p(y))$, where the latter is a convex function of $(p(x,y), p(x)p(y)) = (p(x)p(y|x), p(x)p(y))$, which is a linear function of $p(y|x)$ when $p(x)$ is fixed. ■

1.7 Data-Processing Inequality

Markov Chain

Definition 1.5 — Markov chain. Random variables X, Y and Z form a Markov chain $X \to Y \to Z$ if

$$p(x, y, z)p(y) = p(x, y)p(y, z).$$

1. $X \to Y \to Z$ iff X and Z are conditional independent given Y.
2. $X \to Y \to Z$ iff $I(X; Z|Y) = 0$.
3. $X \to Y \to Z$ implies $Z \to Y \to X$.

Theorem 1.18

$$I(X; Y, Z) \geq I(X; Y)$$

with equality iff $X \to Y \to Z$ forms a Markov chain.

Theorem 1.19 — Data processing inequality. If $X \to Y \to Z$, then $I(X;Y) \geq I(X;Z)$.

Proof. Using the chain rule of mutual information to expend $I(X;Y|Z)$ in two different ways:

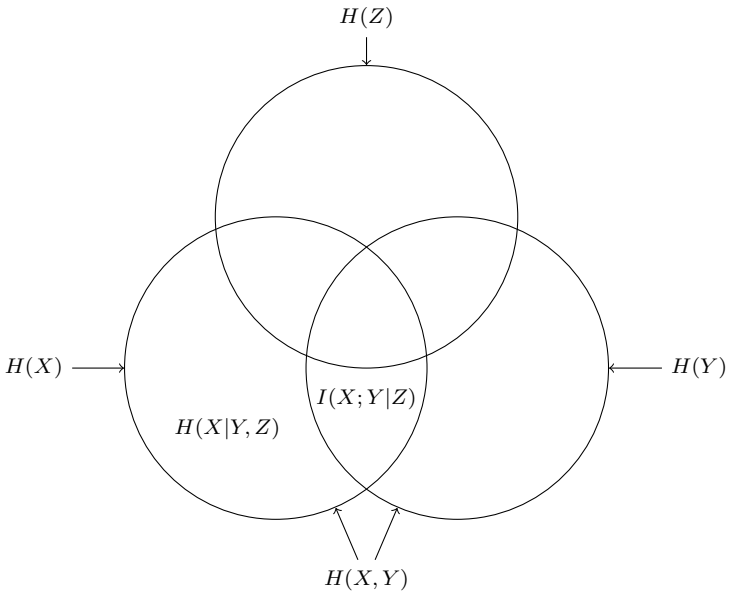
$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y). \end{aligned}$$

See that $I(X;Z|Y) = 0$ and $I(X;Y|Z) \geq 0$. ■

Corollary 1.20 $I(X;Y) \geq I(X;g(Y))$, where g is any function.

Corollary 1.21 If $X \to Y \to Z$, then $I(X;Y|Z) \leq I(X;Y)$.

Information Diagram for Three Random Variables



The information diagram of three random variables can help us to intuitively remember the information inequalities.