《信息论与编码》
《Information Theory and Coding》

Li Chen (陈立)
Professor, School of Electronics and Information Technology (SEIT)
Sun Yat-sen University
Office: 511，JIE Building
Email: chenli55@mail.sysu.edu.cn
Website: www.chencode.cn

《Information Theory and Coding》

Textbooks:

1.《Elements of Information Theory》, by T. Cover and J. Thomas, Wiley (and introduced by Tsinghua University Press), 2003.

2. 《Error Control Coding》, by S. Lin and D. Costello, Prentice Hall, 2004.

3. 《信息论与编码理论》, 王育民、李晖著，高等教育出版社，2013.

Outlines

Chapter 1:	Entropy and Mutual Information	(8L/4W)
Chapter 2:	Channel Capacity	(6L/3W)
Chapter 3:	Source Coding	(4L/2W)
Chapter 4:	Channel Coding	(4L/1W)
Chapter 5:	Convolutional Codes and TCM	(10L/2.5W)
Chapter 6:	Turbo Codes	(6L/1.5W)
Chapter 7:	Reed-Solomon Codes	(12L/3W)

L: Lectures / W: Weeks

Evolution of Communications

Analogue comm.

Late 80s to early 90s

Digital comm.

Information theory and coding techniques

1G

2G

2.5G

EE+CS

3G

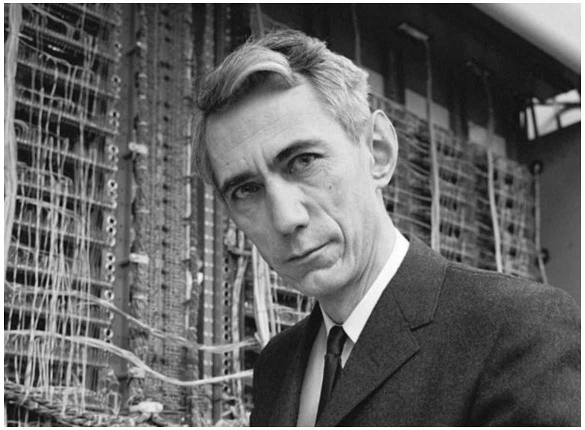
4G

Chapter 1 Entropy and Mutual Information

- 1.1 An Introduction of Information
- 1.2 Entropy
- 1.3 Mutual Information
- 1.4 Further Results on Information Theory

§ 1.1 An Introduction of Information

Information Theory, founded by Claude E. Shannon (1916-2001)



via "A Mathematical Theory of Communication," *Bell System Technical Journal*, 1948.

- What is information?
- How to measure information?
- How to represent information?
- How to transmit information and its limit?

§ 1.1 An Introduction of Information

What is information?

Let us look at the following sentences

1) I will be one year older next year.

   No information

2) I was born in 1993.

   Some information

3) I was born in 1990s.

   More information

Boring!

Being frank!

Interesting, so which year?

Observation 1: Information comes from uncertainty.

Observation 2: The number of *possibilities* should be linked to the information.

§ 1.1 An Introduction of Information

Let us do the following game

Throw a die once

You have 6 possible outcomes.
{1, 2, 3, 4, 5, 6}

Throw three dies

You have 6^3 possible outcomes.
{(1, 1, 1), (1, 1, 2), (1, 1, 3), (1, 1, 4)
……
(2, 1, 1), (2, 1, 2), (2, 1, 3), (2, 1, 4)
……
(6, 6, 3), (6, 6, 4), (6, 6, 5), (6, 6, 6)}

Observation 3: Information should be *‘additive’*.

§ 1.1 An Introduction of Information

Let us look at the following problem

Q: If there are 120 students in our class, and we would like to use bits to distinguish ea ch of them, how many bits do we need?

Solution: 120 possibilities requires

$$\log_2 120 = 6.907 \text{ bits}$$

We need at least 7 bits to represent each of us.

Q: There are 7 billion people on our planet, how many bits do we need?

Observation 4: We can use 'logarithm' to scale down the a huge amount of possibilities.

Observation 5: *Bit (=binary+digit)* permutations are used to represent all possibilities.

§ 1.1 An Introduction of Information

Finally, let us look into the following game

Pick one ball from the hat randomly,

The probability of picking up a white ball, $\frac{1}{4}$ (25%).

Representing the probability needs

$$\log_2 \frac{1}{1/4} = 2 \text{ bits}$$

The probability of picking up a black ball, $\frac{3}{4}$ (75%)

Representing the probability needs

$$\log_2 \frac{1}{3/4} = 0.415 \text{ bits}$$

On average, how many bits do we need to represent an outcome?

$$\frac{1}{4}\log_2 \frac{1}{1/4} + \frac{3}{4}\log_2 \frac{1}{3/4} = 0.811 \text{ bits}$$

Observation 6: Measure of information should consider the *probabilities of various possible events.*

§ 1.1 An Introduction of Information

Events: 1, 2, … , N

Probabilities: $P_1, P_2, \ldots, P_N$

$$P_1 \log_2 P_1^{-1} + P_2 \log_2 P_2^{-1} + \ \ldots + P_N \log_2 P_N^{-1}$$

§ 1.1 An Introduction of Information

- Information: knowledge not precisely known by the recipient, as it is a measure of uncertainty.
- Amount of information ∝ (probability of occurance)$^{-1}$
  E.g., given messages $M_1, M_2, \dots, M_q$ with prob. of occur. $P_1, P_2, \dots, P_q$ $(P_1 + P_2 + \cdots + P_q = 1)$, measure of amount of information carried by each message is
  $$I(M_i) = \log_x P_i^{-1}, \qquad i = 1,2, \dots, q$$
  $x = 2$, $I(M_i)$ in bits
  $x = e$, $I(M_i)$ in nats
  $x = 10$, $I(M_i)$ in Hartley.

- Properties of the measurement

  1) $I(M_i) \to 0, \quad if \quad P_i \to 1;$
  2) $I(M_i) \geq 0, \quad when \quad 0 \leq P_i \leq 1;$
  3) $I(M_i) > I(M_j), \quad if \quad P_j > P_i$
  4) Given M_i and M_j are statistically independent,
     $I(M_i \& M_j) = I(M_i) + I(M_j).$

§ 1.1 An Introduction of Information

Information ←→ 信息

《暮春怀故人》

李中（唐）

池馆寂寥三月尽，落花重叠盖莓苔。
惜春眷恋不忍扫，感物心情无计开。
梦断美人沈信息，目穿长路倚楼台。
琅玕绣段安可得，流水浮云共不回。

§ 1.2 Entropy

How to measure information?

Given a source vector of length N. It has U possible symbols $S_1, S_2, \dots, S_U$, with a probability of occurrence of $P_1, P_2, \dots, P_U$, respectively.

To represent the source vector, we need

$$I = \sum_{i=1}^{U} N P_i \log_2 P_i^{-1} \text{ bits}$$

On average, how many bits do we need for a source symbol?

$$H = \frac{I}{N} = \sum_{i=1}^{U} P_i \log_2 P_i^{-1} \text{ bits/symbol}$$

H is called the source entropy - average amount of information per source symbol. It can also be understood as the expectation of function $\log_2 P_i^{-1}$

$$H = \mathbb{E}\left[\log_2 P_i^{-1}\right] \text{ bits/symbol}$$

§ 1.2 Entropy

Example 1.1: A source vector contains symbols of four possible outcomes A, B, C, D. They occur with probabilities of $P(A) = \frac{1}{4}, P(B) = \frac{1}{3}$, $P(C) = \frac{1}{3}, P(D) = \frac{1}{12}$, respectively.
Entropy of the source vector can be determined as

$$H = \frac{1}{4}\log_2 \frac{1}{1/4} + \frac{2}{3}\log_2 \frac{1}{1/3} + \frac{1}{12}\log_2 \frac{1}{1/12}$$
$$= 1.856 \text{ bits/symbol}$$

Note: If $P(A) = P(B) = P(C) = P(D) = \frac{1}{4}$

$$H = 4 \cdot \frac{1}{4}\log_2 4 = 2 \text{ bits/symbol}$$

§ 1.2 Entropy

<u>Entropy of a binary source</u>: The vector has only two possible symbols, i.e., 0 and 1. Let $P(0)$ denote the probability of a source symbol being 0, and $P(1)$ denote the probability of a source symbol being 1, we have

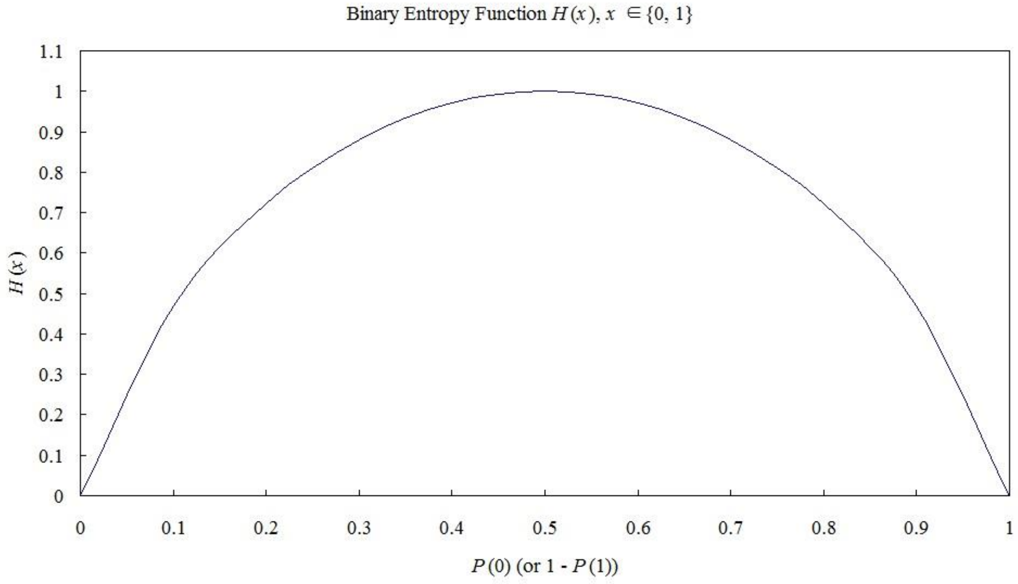
$$H = P(0) \cdot \log_2 P(0)^{-1} + P(1) \log_2 P(1)^{-1}$$

or

$$H = P(0) \cdot \log_2 P(0)^{-1} + \left(1 - P(0)\right) \cdot \log_2 \left(1 - P(0)\right)^{-1}$$

↑

Binary Entropy Function



§ 1.2 Entropy

Entropy of different bases can be interchanged by

$$H_b(x) = H_a(x)\log_b a$$

Proof:

$$H_a(x) = \mathbb{E}[-\log_a P(x)]$$

$$H_a(x)\log_b a = \frac{\lg a}{\lg b}\mathbb{E}\left[-\frac{\lg P(x)}{\lg a}\right]$$

$$= \mathbb{E}\left[-\frac{\lg P(x)}{\lg b}\right]$$

$$= \mathbb{E}[-\log_b P(x)]$$

$$= H_b(x)$$

§ 1.2 Entropy

- Entropy for two random variables X and Y.
- Realizations of X and Y are x and y.
- Distributions of X and Y are $P(x)$ and $P(y)$.

Joint Entropy $H(X, Y)$: Given a joint distribution $P(x, y)$,

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 P(x,y)$$
$$= -\mathbb{E}[\log_2 P(x,y)]$$

Condition Entropy $H(Y|X)$:

$$H(Y|X) = \sum_{x\in X} P(x)H(Y|X=x)$$
$$= -\sum_{x\in X}\sum_{y\in Y} P(x)P(y|x)\log_2 P(y|x)$$
$$= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 P(y|x) = -\mathbb{E}[\log_2 P(y|x)]$$

§ 1.2 Entropy

The Chain Rule (Relationship between Joint Entropy and Conditional Entropy)

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

If X and Y are independent,
$$H(X|Y) = H(X)$$
Hence,
$$H(X,Y) = H(X) + H(Y)$$

Proof:

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 P(x,y)$$
$$= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 (P(y|x)P(x))$$
$$= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 P(x) - \sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 P(y|x)$$
$$= -\sum_{x\in X} P(x)\log_2 P(x) - \sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 P(y|x)$$
$$= H(X) + H(Y|X)$$

§ 1.2 Entropy

The above chain rule can be extended to

$$(1)\quad H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$$
$$(2)\quad H(X_1,X_2,\dots,X_N) = \sum_{i=1}^{N} H(X_i|X_{i-1},X_{i-2},\dots,X_1)$$

Proof:

$$H(X_1,X_2) = H(X_1) + H(X_2|X_1)$$

$$\begin{aligned} H(X_1,X_2,X_3) &= H(X_1) + H(X_2,X_3|X_1) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_2,X_1) \end{aligned}$$

$$\vdots$$

$$H(X_1,X_2,\dots,X_N) = H(X_1) + H(X_2|X_1) + H(X_3|X_2,X_1) + \cdots + H(X_N|X_{N-1},X_{N-2},\dots,X_1)$$

§ 1.3 Mutual Information

- Two random variables X and Y.
- Realizations of X and Y are x and y.
- Distributions of X and Y are $P(x)$ and $P(y)$.
- Joint distribution of X and Y is $P(x, y)$.
- Conditional distribution of X is $P(x|y)$.

Mutual Information between X and Y:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \frac{P(x|y)}{P(x)}$$

$$= \mathbb{E}\left[\log_2 \frac{P(x|y)}{P(x)}\right]$$

§ 1.3 Mutual Information

$$\frac{P(x|y)}{P(x)} = \frac{P(x,y)}{P(x)P(y)}$$

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} P(x,y) \log_2 \frac{P(x,y)}{P(x)P(y)} = \mathbb{E}\left[\log_2 \frac{P(x,y)}{P(x)P(y)}\right]$$

Note: If X and Y are independent, $P(x)P(y) = P(x,y)$, $I(X,Y) = 0$.

§ 1.3 Mutual Information

Mutual Information's Relationship with Entropy:

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

Proof:

$$I(X,Y) = \sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)}$$

$$= \sum_{x\in X}\sum_{y\in Y} P(x,y)\log_2 P(x,y) - \sum_{x\in X} P(x)\log_2 P(x) - \sum_{y\in Y} P(y)\log_2 P(y)$$

$$= H(X) + H(Y) - H(X,Y)$$

Note: The above proof also shows the symmetry of mutual information as

$$I(X,Y) = I(Y,X)$$

§ 1.3 Mutual Information

Mutual Information's Relationship with Entropy:

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

This relationship can be visualized in the Venn diagram

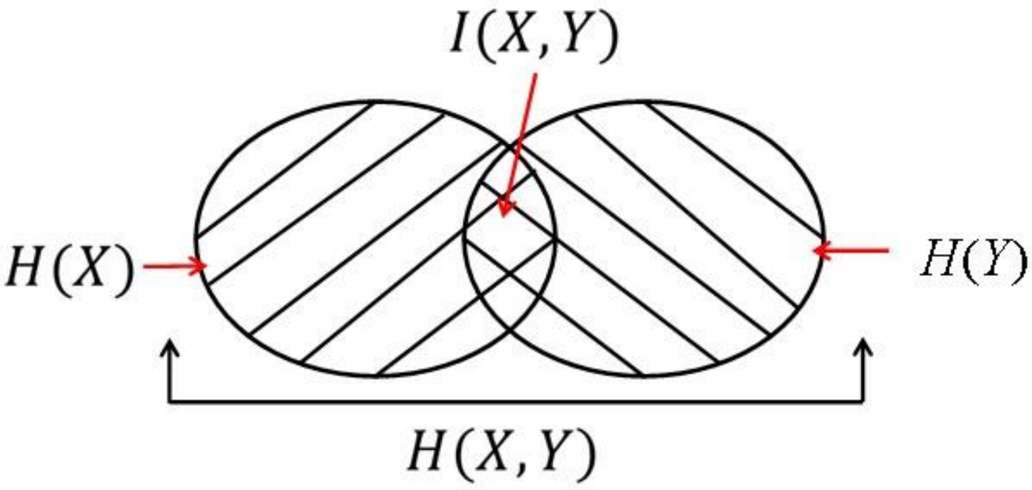


Fig. A Venn diagram

§ 1.3 Mutual Information

Corollary:

$$I(X,Y) = H(X) - H(X|Y)$$
$$= H(Y) - H(Y|X)$$

This can also be concluded using the chain rule.

Notes: 1) $0 \le I(X,Y) \le \min\{H(X), H(Y)\}$.

2) If $H(X) \sqsubset H(Y)$, $I(X,Y) = H(X)$.

Similarly if $H(Y) \sqsubset H(X)$, $I(X,Y) = H(Y)$.

3) $I(X,X) = H(X) - H(X|X) = H(X)$

<u>Entropy is also called self information</u>

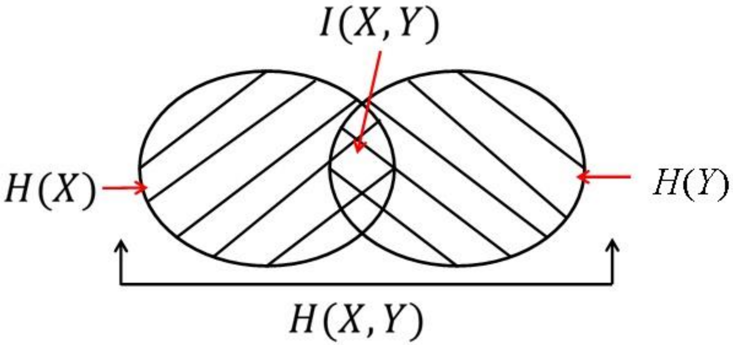


Fig. A Venn diagram

§ 1.3 Mutual Information

The chain rules for arbitrary number of variables

For entropy,

$$H(X_1, X_2, \ldots, X_N) = H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) + \cdots + H(X_N|X_{N-1}, X_{N-2}, \ldots, X_1)$$

$$= \sum_{i=1}^{N} H(X_i|X_{i-1}, X_{i-2}, \ldots, X_1)$$

For mutual information,

$$I(X_1, X_2, \ldots, X_N; Y) = H(X_1, X_2, \ldots, X_N) - H(X_1, X_2, \ldots, X_N|Y)$$

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

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

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

§ 1.3 Mutual Information

Mutual Information of a Channel

- Consider X is the transmitted signal, Y is the received signal.
- Y is a variant of X where the discrepancy is introduced by channel.

$$H(X) - H(X|Y)$$

How much we don't know BEFORE the channel observations.

−

How much we still don't know AFTER the channel observations.

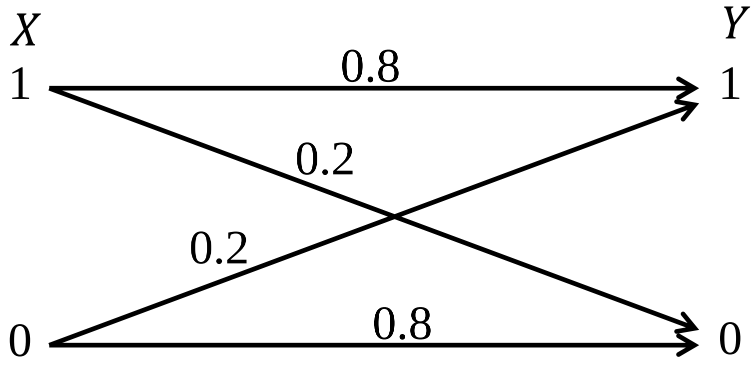
⇩

How much information is carried by the channel, and this is called the mutual information of the channel, denoted as $I(X,Y)$.

Note: Mutual information $I(X, Y)$ describes the amount of information one variable X contains about the other Y, or vice versa as in $I(Y, X)$.

§ 1.3 Mutual Information

Example 1.2: Given the binary symmetric channel shown as



We know $P(x=0)=0.3$, $P(x=1)=0.7$, $P(y=1|x=1)=0.8$,
$P(y=1|x=0)=0.2$, $P(y=0|x=1)=0.2$ and $P(y=0|x=0)=0.8$.

Please determine the mutual information of the channel.

Solution: We may determine the channel mutual information by $I(X,Y)=H(X)-H(X|Y)$

- Entropy of the binary source is

$$H(X) = -P(x=0)\log_2 P(x=0) - P(x=1)\log_2 P(x=1)$$
$$= 0.3 \cdot \log_2 \frac{1}{0.3} + 0.7 \cdot \log_2 \frac{1}{0.7}$$
$$= 0.881 \text{ bits/symbol}$$

§ 1.3 Mutual Information

- With $P(x)$ and $P(y|x)$, we know

$$P(y=1) = P(y=1|x=1)P(x=1) + P(y=1|x=0)P(x=0)$$
$$= 0.62$$
$$P(y=0) = P(y=0|x=1)P(x=1) + P(y=0|x=0)P(x=0)$$
$$= 0.38$$
$$P(x=0, y=0) = P(y=0|x=0)P(x=0) = 0.24$$
$$P(x=0|y=0) = \frac{P(x=0,y=0)}{P(y=0)} = 0.63$$
$$P(x=1, y=0) = P(y=0|x=1)P(x=1) = 0.14$$
$$P(x=1|y=0) = \frac{P(x=1,y=0)}{P(y=0)} = 0.37$$
$$P(x=0, y=1) = P(y=1|x=0)P(x=0) = 0.06$$
$$P(x=0|y=1) = \frac{P(x=0,y=1)}{P(y=1)} = 0.10$$
$$P(x=1, y=1) = P(y=1|x=1)P(x=1) = 0.56$$
$$P(x=1|y=1) = \frac{P(x=1,y=1)}{P(y=1)} = 0.90$$

§ 1.3 Mutual Information

- Hence, the conditional entropy is:

$$H(X|Y) = P(x=0,y=0)\log_2\frac{1}{P(x=0|y=0)} + P(x=1,y=0)\log_2\frac{1}{P(x=1|y=0)}$$

$$+P(x=0,y=1)\log_2\frac{1}{P(x=0|y=1)} + P(x=1,y=1)\log_2\frac{1}{P(x=1|y=1)}$$

$$= 0.24\log_2\frac{1}{0.63} + 0.14\log_2\frac{1}{0.37} + 0.06\log_2\frac{1}{0.10} + 0.56\log_2\frac{1}{0.90}$$

$$= 0.644 \text{ bits/sym.}$$

- The mutual information is:

$$I(X,Y) = H(X) - H(X|Y) = 0.237 \text{ bits}$$

Note: You may try to solve the same problem through

$$I(X,Y) = H(Y) - H(Y|X)$$

§ 1.4 Further Results on Information Theory

Relative Entropy: Assume X and $\hat{X}$ are two random variables with realizations of x and $\hat{x}$, respectively. They aim to describe the same event, with probability mass functions of $P(x)$ and $P(\hat{x})$, respectively. Their relative entropy is

$$D\big(P(x), P(\hat{x})\big) = \sum_{x \in \text{supp}\, P(x)} P(x) \log_2 \frac{P(x)}{P(\hat{x})}$$

$$= \mathbb{E}\left[\log_2 \frac{P(x)}{P(\hat{x})}\right]$$

- It is often called the **Kullback-Leibler distance** between two distributions $P(x)$ and $P(\hat{x})$.
- It is a measure of inefficiency by assuming a distribution $P(\hat{x})$ when the true distribution is $P(x)$. E.g., an event can be described by an average length of $H(P(x))$ bits. However, if we assume its distribution is $P(\hat{x})$, we will need an average length of $H\big(P(x)\big) + D\big(P(x), P(\hat{x})\big)$ bits to describe it.
- It is not symmetric as $D\big(P(x), P(\hat{x})\big) \neq D\big(P(\hat{x}), P(x)\big)$.

§ 1.4 Further Results on Information Theory

***Example 1.3*:**

Let	X	:	**A**	**B**	**C**	**D**
	$P(x)$	:	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{8}$
	$P(\hat{x})$	:	$\frac{3}{8}$	$\frac{2}{5}$	$\frac{1}{10}$	$\frac{1}{8}$

$H\big(P(x)\big) = 1.75$ bits/symbol $\qquad$ $H\big(P(\hat{x})\big) = 1.805$ bits/symbol

$$D\big(P(x), P(\hat{x})\big) = \frac{1}{4}\log_2\frac{1/4}{3/8} + \frac{1}{2}\log_2\frac{1/2}{2/5} + \frac{1}{8}\log_2\frac{1/8}{1/10} + \frac{1}{8}\log_2\frac{1/8}{1/8}$$

If $P(x_i) = P(\hat{x_i})$, no extra bits;
If $P(x_i) < P(\hat{x_i})$, less bits;
If $P(x_i) > P(\hat{x_i})$, more bits.

§ 1.4 Further Results on Information Theory

- **Corollary 1:** When $P(x) = P(\hat{x})$, $D\big(P(x), P(\hat{x})\big) = 0$.
- **Corollary 2:** $D\big(P(x), P(\hat{x})\big) \geq 0$.

Proof:

$$
\begin{aligned}
-D\big(P(x), P(\hat{x})\big) &= \sum_{x \in \operatorname{supp} P(x)} P(x) \log_2 \frac{P(\hat{x})}{P(x)} \\
&\leq \sum_{x \in \operatorname{supp} P(x)} P(x) \left(\frac{P(\hat{x})}{P(x)} - 1 \right) \log_2 e \\
&= \left(\sum_{x \in \operatorname{supp} P(x)} P(\hat{x}) - \sum_{x \in \operatorname{supp} P(x)} P(x) \right) \log_2 e \\
&\leq (1 - 1) \log_2 e \\
&= 0
\end{aligned}
$$

IT Inequality: Given $b > 1$ and $\varepsilon > 0$

$$\left(1 - \frac{1}{\varepsilon}\right) \log_b e \leq \log_b \varepsilon \leq (\varepsilon - 1) \log_b e$$

§ 1.4 Further Results on Information Theory

***Example 1.4*:** The true distribution $P(x)$ is given. If we assume a distribution of $P(\hat{x}_i) = \frac{1}{k}$ for $i = 1,2, \dots, k$ to describe the same event, then

$$D\big(P(x), P(\hat{x})\big) = \mathbb{E}\left[\log_2 \frac{P(x)}{P(\hat{x})}\right] = \mathbb{E}[\log_2 kP(x)]$$

$$= \mathbb{E}[\log_2 k] + \mathbb{E}[\log_2 P(x)]$$

$$= \mathbb{E}[\log_2 P(\hat{x})^{-1}] - \mathbb{E}[\log_2 P(x)^{-1}]$$

$$= H\big(P(\hat{x})\big) - H(P(x))$$

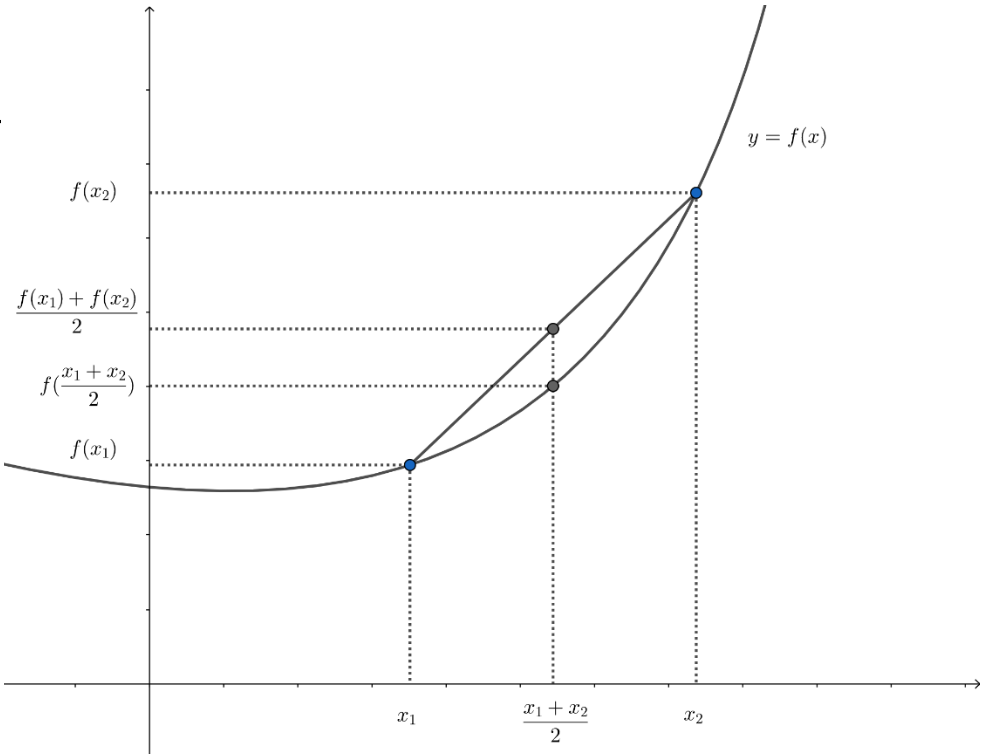
§ 1.4 Further Results on Information Theory

Convex Function: A function $f(x)$ is convex (凸) over the interval (a, b) if $\forall x_1, x_2 \in (a, b)$ and $0 \le \lambda \le 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2).$$

It is strictly convex if the equality holds when $\lambda = 0$ or $\lambda = 1$.

- If $f(x)$ is convex, $-f(x)$ is concave (凹).



§ 1.4 Further Results on Information Theory

- ***Example 1.5*:** $\log_2 \frac{1}{x}$ is strictly convex over $(0, \infty)$.

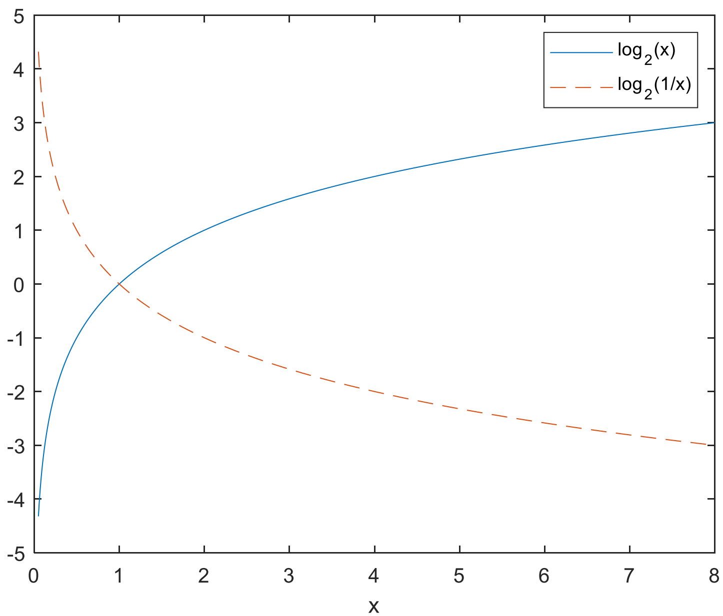
  Let $x_1 = 2$, $x_2 = 5$ and $\lambda = 0.5$,

$$\log_2 \frac{1}{0.5 \times 2 + 0.5 \times 5} = -1.81$$

$$0.5 \times \log_2 \frac{1}{2} + 0.5 \times \log_2 \frac{1}{5} = -1.66$$

  When $\lambda = 0$ or $\lambda = 1$, the equality holds.

  Note that $\log_2 x$ is concave.



§ 1.4 Further Results on Information Theory

Jensen's Inequality: If function $f(x)$ is convex, then

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)].$$

Proof: With two mass points x_1 and x_2 and distributions of p_1 and p_2, the convexity implies

$$f(p_1 x_1 + p_2 x_2) \leq p_1 f(x_1) + p_2 f(x_2).$$

Assume this is also true for $k-1$ mass points that

$$f(p_1 x_1 + \cdots + p_{k-1} x_{k-1}) \leq p_1 f(x_1) + \cdots + p_{k-1} f(x_{k-1}).$$

For k mass points that substantiate $\sum_{i=1}^{k-1} p_i + p_k = 1$, we have

$$f(p_1 x_1 + \cdots + p_{k-1} x_{k-1}) + p_k f(x_k) \leq p_1 f(x_1) + \cdots + p_k f(x_k) = \sum_{i=1}^{k} p_i f(x_i)$$

§ 1.4 Further Results on Information Theory

Let $p_i' = \frac{p_i}{1-p_k}$, for $i = 1,2,\dots,k-1$.

$$\sum_{i=1}^{k} p_i f(x_i) = \sum_{i=1}^{k-1}(1-p_k)p_i' f(x_i) + p_k f(x_k)$$

$$\geq (1-p_k) f\left(\sum_{i=1}^{k-1} p_i' x_i\right) + p_k f(x_k)$$

$$\geq f\left(\sum_{i=1}^{k-1}(1-p_k)p_i' x_i + p_k x_k\right)$$

$$= f\left(\sum_{i=1}^{k} p_i x_i\right)$$

Note: If function $f(x)$ is concave, $\mathbb{E}[f(x)] \leq f(\mathbb{E}[x])$.

§ 1.4 Further Results on Information Theory

- Jensen's inequality can be applied to prove some properties on entropy.
- **Corollary 2:** $D\big(P(x), P(\hat{x})\big) \geq 0$

Proof:

$$-D\big(P(x), P(\hat{x})\big) = \sum_{x \in \operatorname{supp} P(x)} P(x) \log_2 \frac{P(\hat{x})}{P(x)}$$

$$\leq \log_2 \sum_{x \in \operatorname{supp} P(x)} P(\hat{x})$$

$$\leq \log_2 1 = 0$$

- **Corollary 3:** $I(X, Y) \geq 0$

Proof:

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} P(x, y) \log_2 \frac{P(x, y)}{P(x)P(y)}$$

$$= D(P(x, y), P(x)P(y)) \geq 0$$

$I(X, Y) = 0$ only if $P(x, y) = P(x)P(y)$, i.e., X and Y are independent.

§ 1.4 Further Results on Information Theory

- **Corollary 4 (Maximum Entropy Distribution):**

Given variable $X \in \{x_1, x_2, \dots, x_U\}$, with a distribution of $P_1, P_2, \dots, P_U$. We have

$$H(X) \leq \log_2 U$$

Proof:

$$H(X) = \sum_{i=1}^{U} P_i \log_2 P_i^{-1}$$

Since $\log_2(\cdot)$ is a concave function, based on Jensen's inequality, we have

$$H(X) \leq \log_2 \left(\sum_{i=1}^{U} P_i P_i^{-1} \right)$$

$$= \log_2 U$$

Note: If X is uniformly distributed over $x_1, x_2, \dots, x_U$, i.e., $P_1 = P_2 = \cdots = P_U = \frac{1}{U}$,

$$H(X) = \log_2 U$$

§ 1.4 Further Results on Information Theory

Fano's Inequality: Let X and Y be two random variables with realizations in $\{x_1, x_2, \dots x_k\}$. Let $P_e = \Pr[X \neq Y]$, then

$$H(X|Y) \leq H(P_e) + P_e \log_2(k-1).$$

Proof: Let us create a binary variable Z such that

$$\begin{matrix} Z = 0, \text{if } X = Y \\ Z = 1, \text{if } X \neq Y \end{matrix} \Rightarrow \begin{matrix} \Pr(Z=0) = 1 - P_e \\ \Pr(Z=1) = P_e \end{matrix}$$

Hence, $H(Z) = H(P_e)$. Base on the chain rule for entropy,

$$H(XZ|Y) = H(X|Y) + H(Z|XY) = H(X|Y)$$

Note, with the knowledge of X and Y, Z is deterministic and $H(Z|XY) = 0$.
Also based on the chain rule,

$$\begin{aligned} H(XZ|Y) &= H(Z|Y) + H(X|YZ) \\ &\leq H(Z) + H(X|YZ) \end{aligned}$$

§ 1.4 Further Results on Information Theory

Therefore, $H(X|Y) \leq H(Z) + H(X|YZ)$.

- $H(Z) + H(X|YZ)$.
- $H(X|YZ) = \Pr(Z = 0)\, H(X|Y, Z = 0) + \Pr(Z = 1)\, H(X|Y, Z = 1)$.

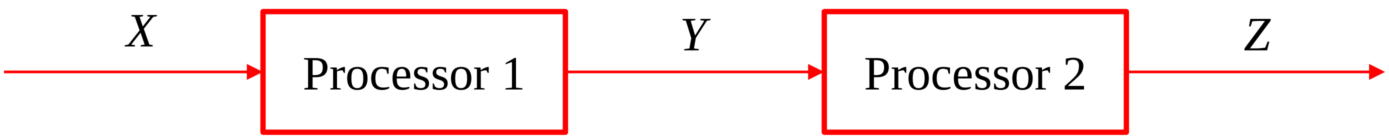
$$= (1 - P_e) \cdot 0 + P_e \log_2(k - 1)$$

$$= P_e \log_2(k - 1)$$

Note: $H(P_e)$ is the number of bits required to describe X whenever $X = Y$;
$\log_2(k-1)$ is the number of bits required to describe X whenever $X \neq Y$.
The equality is reached when X is uniformly distributed over all $k - 1$ values.

§ 1.4 Further Results on Information Theory

Data Processing Inequality: Given a concatenated data processing system as



$X \to Y \to Z$ form a Markov chain that holds

$$P(x,y,z) = P(x,y) \cdot P(z|y) = P(x)P(y|x)P(z|y)$$
$$P(z|x,y) = P(z|y)$$
$$P(x|y,z) = P(x|y)$$

We have

$$I(X,Z) \leq \begin{cases} I(X,Y) \\ I(Y,Z) \end{cases}.$$

§ 1.4 Further Results on Information Theory

Proof: Since $P(z|x,y) = P(z|y)$ holds,

$$H(Z|XY) = \mathbb{E}[-\log_2 P(z|xy)] = \mathbb{E}[-\log_2 P(z|y)] = H(Z|Y)$$

Similarly, since $P(x|y,z) = P(x|y)$ holds,

$$H(X|ZY) = H(X|Y)$$

$$\begin{aligned} I(X,Z) &= H(X) - H(X|Z) \\ &\le H(X) - H(X|ZY) \\ &= H(X) - H(X|Y) \\ &= I(X,Y) \end{aligned}$$

$$\begin{aligned} I(X,Z) &= H(Z) - H(Z|X) \\ &\le H(Z) - H(Z|XY) \\ &= H(Z) - H(Z|Y) \\ &= I(Y,Z) \end{aligned}$$

Remark: Information cannot be increased by data processing.

References:

[1] Elements of Information Theory, by T. Cover and J. Thomas.

[2] Scriptum for the lectures, Applied Information Theory, by M. Bossert.