《信息论基础》
《Fundamentals on Information Theory》

Li Chen (陈立)
Professor, School of Electronics and Information Technology (SEIT)
Sun Yat-sen University
Office: 631C, SEIT building
Email: chenli55@mail.sysu.edu.cn
Website: www.chencode.cn

《Fundamentals on Information Theory》

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	(3 W)
Chapter 2:	Channel Capacity	(3 W)
Chapter 3:	Source Coding	(2 W)
Chapter 4:	Channel Coding	(2 W)
Chapter 5:	Convolutional Codes	(4 W)
Chapter 6:	Reed-Solomon Codes	(4 W)

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

- What is information?
- How do we measure information?

Let us look at the following sentences:

1) I will be one year older next year.

   No information

   Boring!

2) I was born in 1993.

   Some information

   Being frank!

3) I was born in 1990s.

   More information

   Interesting, so which year?

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)}

Information should be *'additive'*.

§ 1.1 An Introduction of Information

Let us look at the following problem.

If there are 30 students in our class, and we would like to use binary bits to distinguish each of them, how many bits do we need?

Solution: 30 possibilities.

requires

$\log_2 30 = 4.907$ bits.

we need at least 5 bits to represent each of us.

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

We can use *'logarithm'* to scale down the a huge amount of possibilities.

Number (binary bit) 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

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

§ 1.1 An Introduction of Information

- How do we measure the overall event? (On average, how many bits do we need to represent an outcome?)

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

- The measure of information should be

$$\sum_{i=1}^{N}P_i\log_2 P_i^{-1}=-\sum_{i=1}^{N}P_i\log_2 P_i$$

  - P_i: probability of the ith possible event.
  - N: Total number of possible events.

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

§ 1.2 Entropy

- Information: knowledge not precisely known by the recipient, as it is a measure of unexpectedness.
- Amount of information $\propto$ (probability of occurance)$^{-1}$
- Messages: M_1 M_2 M_3 ... ... M_q

Prob of occur: P_1 P_2 P_3 ... ... P_q $(P_1 + P_2 + P_3 + \cdots + P_q = 1)$

Measure the amount of information carried by each message by

$I(M_i) = log_x P_i^{-1}$, $i = 1, 2, \ldots, q$

$x = 2$, $I(M_i)$ in bits

$x = e$, $I(M_i)$ in nats

$x = 10$, $I(M_i)$ in Hartley.

- Observations: … …

§ 1.2 Entropy

Observations:

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,
   $I(M_i \& M_j) = I(M_i) + I(M_j).$

§ 1.2 Entropy

Example 1.1: A source outputs five possible messages. The probabilities of these messages are:

$$P_1 = \frac{1}{2} \quad P_2 = \frac{1}{4} \quad P_3 = \frac{1}{8} \quad P_4 = \frac{1}{16} \quad P_5 = \frac{1}{16}.$$

Determine the information contained in each of these messages.

Solution:

$$I(M_1) = \log_2 \frac{1}{1/2} = 1 \text{ bit}$$

$$I(M_2) = \log_2 \frac{1}{1/4} = 2 \text{ bit}$$

$$I(M_3) = \log_2 \frac{1}{1/8} = 3 \text{ bit}$$

$$I(M_4) = \log_2 \frac{1}{1/16} = 4 \text{ bit}$$

$$I(M_5) = \log_2 \frac{1}{1/16} = 4 \text{ bit}$$

Total amount of information = 14 bits. Is it right?

§ 1.2 Entropy

Given a source vector of length N, and it has U possible symbols $S_1, S_2, \ldots S_U$, each of which has probability of $P_1, P_2, \ldots P_U$ of occurrence.

To represent the source vector, we need

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

So on average, how many information 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 number of information per source symbol.

§ 1.2 Entropy

Example 1.2: A source vector contains symbols of four possible outcomes A, B, C, D. They occur with probabilities of $\frac{1}{4}$, $\frac{1}{3}$, $\frac{1}{3}$ and $\frac{1}{12}$, respectively. Determine the entropy of the source vector.

$$H = \frac{1}{4}\log_2 \frac{1}{1/4} + \frac{2}{3}\log_2 \frac{1}{1/3} + \frac{1}{12}\log_2 \frac{1}{1/12}$$

$$= 1.856 \text{ bits/symbol}$$

§ 1.2 Entropy

Entropy of a binary source: The source 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

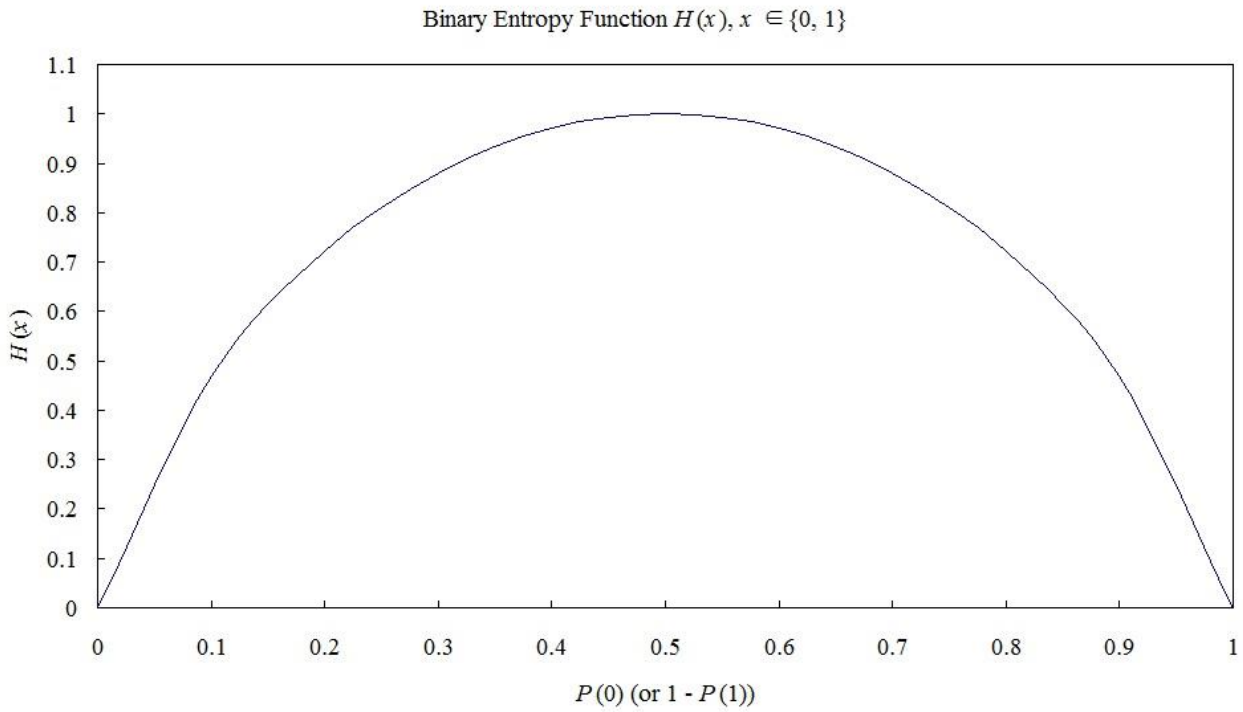
$$H = P(0) \cdot \log_2 P(0)^{-1} + P(1)\log_2 P(1)^{-1}$$

or

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

Binary Entropy Function

§ 1.2 Entropy


Binary Entropy Function $H(x)$, $x \in \{0, 1\}$
1.1
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
$H(x)$
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
$P(0)$ (or 1 - $P(1)$)

§ 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 $P(x, y)$,

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

Condition Entropy $H(Y|X)$:

$$H(Y|X) = \sum_{x\in X} P(x)H(Y|X=x)$$

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

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

§ 1.2 Entropy

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

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

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

Proof:

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

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

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

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

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

§ 1.3 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:

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

§ 1.3 Mutual Information

Mutual Information's Relationship with Entropy:

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

Proof:

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

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

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

Remark: 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

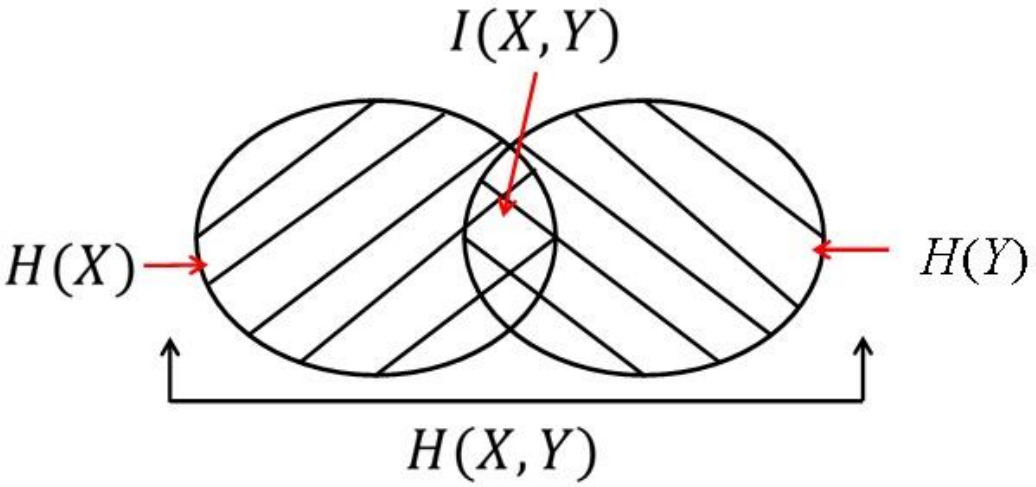


Fig. A Venn diagram

§ 1.3 Mutual Information

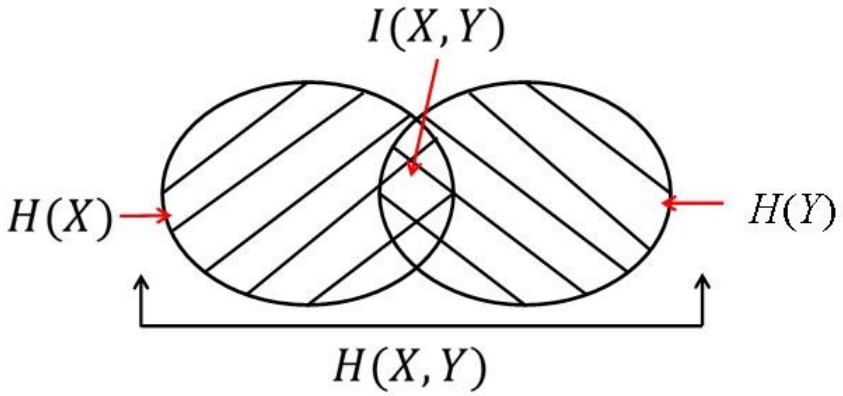


Fig. A Venn diagram

Corollary:

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

This can also be concluded using the Chain Rule.

Bounds on $I(X,Y)$

$$0 \le I(X,Y) \le \min\{H(X), H(Y)\}$$

§ 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.

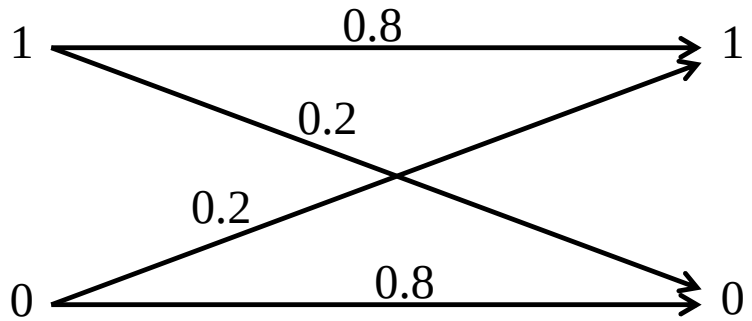
How much we still don't know AFTER the channel observations.

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

Remark: 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 1.3: 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 such a channel.

Solution:

- Entropy of the binary source is

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

§ 1.3 Mutual Information

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

$$P(y=1) = P(y=1|x=1)P(x=1) + P(y=1|x=0)P(x=0) = 0.62$$

$$P(y=0) = P(y=0|x=1)P(x=1) + P(y=0|x=0)P(X=0) = 0.38$$

$$P(x=0, y=0) = P(y=0|x=0) \cdot P(x=0) = 0.24$$

$$P(x=0|y=0) = \frac{P(x=0,y=0)}{P(y=0)} = 0.63$$

$$P(x=1, y=0) = P(y=0|x=1) \cdot P(x=1) = 0.14$$

$$P(x=1|y=0) = \frac{P(x=1,y=0)}{P(y=0)} = 0.37$$

$$P(x=0, y=1) = P(y=1|x=0)P(x=0) = 0.06$$

$$P(x=0|y=1) = \frac{P(x=0,y=1)}{P(y=1)} = 0.10$$

$$P(x=1, y=1) = P(y=1|x=1)P(x=1) = 0.56$$

$$P(x=1|y=1) = \frac{P(x=1, y=1)}{P(y=1)} = 0.90$$

§ 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:

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

§ 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

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

- 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

- **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:

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

$$\leq \sum_{x \in \operatorname{supp} P(x)} P(x) \left(\frac{P(\hat{x})}{P(x)} - 1 \right) \log_2 e$$

$$\leq \left(\sum_{x \in \operatorname{supp} P(x)} P(\hat{x}) - \sum_{x \in \operatorname{supp} P(x)} P(x) \right) \log_2 e$$

$$\leq (1 - 1) \log_2 e$$

$$= 0$$

§ 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

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

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

$$= \log_2 k - H(P(x))$$

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

§ 1.4 Further Results on Information Theory

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

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

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

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

$$f(p_1x_1 + p_2x_2) \leq p_1f(x_1) + p_2f(x_2).$$

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

$$f(p_1x_1 + \cdots + p_{k-1}x_{k-1}) \leq p_1f(x_1) + \cdots + p_{k-1}f(x_{k-1}).$$

Therefore, for k mass points, we have

$$\sum_{i=1}^{k} p_if(x_i) \geq p_kf(x_k) + f(p_1x_1 + \cdots + p_{k-1}x_{k-1}).$$

§ 1.4 Further Results on Information Theory

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

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

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

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

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

§ 1.4 Further Results on Information Theory

- Jensen's inequality can be applied to prove some properties on entropy.
- **Consequence 1:** $D\big(P(x), P(\hat{x})\big) \geq 0$

  Proof:

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

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

$$\leq \log_2 1 = 0$$

- **Consequence 2:** $I(X, Y) \geq 0$

  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

Fano's Inequality: Let X and Y are two random variables with realizations in $\{x_1, x_2, \ldots x_k\}$. 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

$$Z = 0, \text{if } X = Y.$$

$$Z = 1, \text{if } X \neq Y.$$

Hence, $H(Z) = H(P_e)$.

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

Note, with the knowledge of X and Y, Z is deterministic.

§ 1.4 Further Results on Information Theory

$$H(XZ|Y) = H(Z|Y) + H(X|YZ)$$
$$\leq H(Z) + H(X|YZ)$$

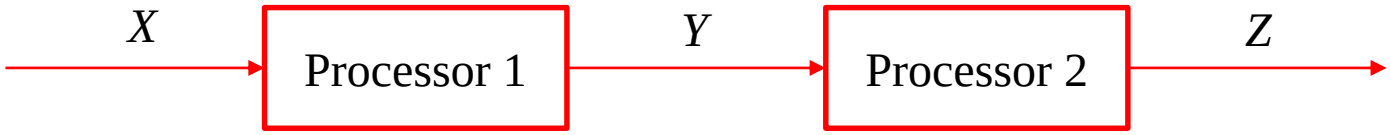
Note,

$$H(X|Y, Z = 0) = 0,$$
$$H(X|Y, Z = 1) = \log_2(k-1),$$
$$H(X|Y) \leq H(Z) + P_e \log_2(k-1).$$

- $H(X|Y)$ implies with the knowledge of Y, how much uncertainty is left about X (X and Y are related);
- $H(P_e)$: numbers of bits to describe X whenever $X = Y$;
- $\log_2(k-1)$: number of bits to describe X whenever $X \neq Y$.

§ 1.4 Further Results on Information Theory

Data Processing Inequality: Given a concatenated data processing system as



We have

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

§ 1.4 Further Results on Information Theory

Proof:

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

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

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.