• Random processes

A random process $X(t)$ is wide-sense stationary (WSS) if

$$m_X(t) \triangleq E[X(t)] \equiv \text{constant}$$

$$R_X(t, t+\tau) \triangleq E[X(t)X(t+\tau)] = R_X(\tau)$$

The power spectral density of a WSS process $X(t)$ is

$$S_X(f) = \int_{-\infty}^{+\infty} R_X(\tau) e^{-j2\pi f\tau}\, d\tau$$

$$R_X(\tau) \rightleftharpoons S_X(f).$$

• WSS processes through linear filters

$$X(t) \longrightarrow \boxed{h(t)} \longrightarrow Y(t)$$

$X(t)$ is WSS $\Longrightarrow$ $Y(t)$ is WSS

$$m_X(t) \longrightarrow \boxed{h(t)} \longrightarrow m_Y(t)$$

$$R_X(\tau) \longrightarrow \boxed{h(\tau)} \longrightarrow \boxed{h(-\tau)} \longrightarrow R_Y(\tau).$$

$$S_Y(f) = S_X(f)\,|H(f)|^2, \qquad h(\tau) \rightleftharpoons H(f).$$

• **White Gaussian noise**: $W(t)$ is a white Gaussian noise process if (1) $W(t)$ is WSS
(2) $W(t)$ is Gaussian
(3) $m_W(t) \equiv 0$
and (4) $S_W(f) = N_0/2 \quad \forall f.$

Chapter 2: Source Coding

In this chapter, we consider the following diagram

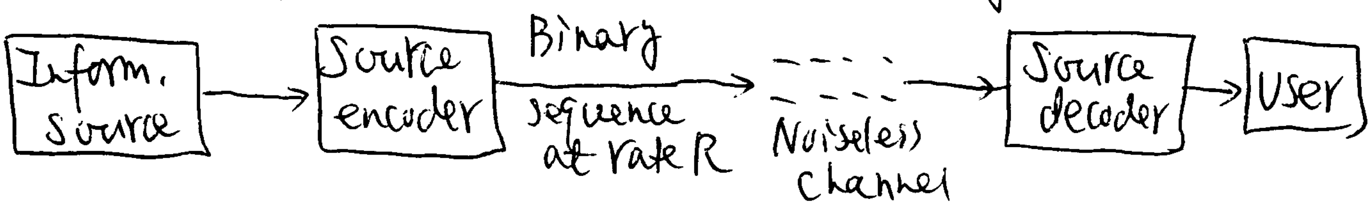


Fig-2.1: Source coding. Assume the channel is noiseless.

Our purpose is to provide answers to the following two questions.

Question 1: What is the minimum rate R? ~~To~~ In other words, what is the minimum number of information bits per symbol that must be conveyed in order for the receiver to recover the original source either perfectly or with some allowable distortion?

~~Questions~~

Question 2: The minimum rate R represents the ultimate compression rate. How to design source codes (or data compression algorithms) that can provide compression rates close or equal to the ultimate compression rate R?

The answers to the above questions, of course, depend on source models. As we shall see later, for a discrete source, the ultimate lossless compression rate is given by the entropy rate of the source. For an analog source, the ultimate compression rate is given by the rate distortion function of the source.

§2.1 Mathematical Models for Information Sources

As alluded above, there are two types of sources: discrete sources and analog sources.

• Discrete Sources

A discrete source is a signal in which both ~~alp~~ amplitude and time take on discrete values. Thus a discrete source can be represented by a sequence $\{X_i\}_{i=-\infty}^{+\infty}$ or $\{X_i\}_{i=1}^{+\infty}$, where each X_i takes values in a finite alphabet $\mathcal{X} = \{x_0, x_1, \cdots, x_{J-1}\}$. For example,

$$\{X_i\} = 0110110110001.$$

Usually, one assumes that $\{X_i\}$ is a random sequence with alphabet ~~$\mathcal{X}$~~ $\mathcal{X}$.

• Examples of Discrete Sources

• Example 1 (stationary sources) A discrete source $\{X_i\}$ with alphabet $\mathcal{X}$ is said to be (strictly) stationary if for any integer m and positive integer n, the random vectors $(X_1, X_2, \cdots, X_n)$ and $(X_{1+m}, X_{2+m}, \cdots, X_{n+m})$ have the same joint probability distribution, i.e.,

$$P(X_1 = u_1, X_2 = u_2, \cdots, X_n = u_n) = P(X_{1+m} = u_1, X_{2+m} = u_2, \cdots, X_{n+m} = u_n) \quad (2.1.1)$$

for any sequence $u_1, u_2, \cdots, u_n$, where $u_i \in \mathcal{X}$. In other words, the joint probability distribution of any random vector is invariant under a shift in the time origin.

• Example 2 (discrete memoryless sources) A discrete source $\{X_i\}$ with alphabet $\mathcal{X}$ is said to be memoryless if $\{X_i\}$ is an i.i.d random sequence, i.e., for any m and n, the random vector $(X_{m+1}, X_{m+2}, \cdots, X_{m+n})$ is i.i.d. ~~This~~ Thus a memoryless source $\{X_i\}$ is uniquely determined by its generic (or ~~marg~~ marginal) probability mass function:

$$p(x_j) = P(X_n = x_j) \quad \text{for any } 0 \le j \le J-1 \text{ and any } n \quad (2.1.2)$$

Furthermore, for any sequence $u_1, u_2, \cdots, u_n$, where $u_i \in \mathcal{X}$,

$$P(X_1=u_1, X_2=u_2, \ldots, X_n=u_n) = P(X_1=u_1)P(X_2=u_2)\cdots P(X_n=u_n)$$
$$= \prod_{i=1}^{n} p(u_i) \qquad (2.1.3)$$

The current output letter is statistically independent from all past and future outputs. Obviously, a discrete memoryless source is stationary.

• Example 3 (Markov Sources) A discrete source $\{X_i\}$ is said to be ~~Marko~~ a Markov (or first order Markov) source if the current source output statistically depends on only the previous output letter, that is, for any m and n

$$P(X_{n+m}=u_n \mid X_{n-1+m}=u_{n-1}, \ldots, X_{1+m}=u_1)$$
$$= P(X_{n+m}=u_n \mid X_{n-1+m}=u_{n-1})$$
$$\triangleq p(u_n \mid u_{n-1}) \qquad u_i \in \mathcal{X}$$

independent of m. (2.1.4)

The matrix $\mathbf{P}$

$$\mathbf{P} = \begin{bmatrix} p(x_0|x_0) & p(x_1|x_0) & \cdots & p(x_{J-1}|x_0) \\ p(x_0|x_1) & p(x_1|x_1) & \cdots & p(x_{J-1}|x_1) \\ \vdots & & & \\ p(x_0|x_{J-1}) & p(x_1|x_{J-1}) & \cdots & p(x_{J-1}|x_{J-1}) \end{bmatrix}$$

is called the ~~tran~~ probability transition matrix of the source.

~~Inf~~ Note that in the random process literature, the above definition corresponds to that of homogeneous Markov chain.

- Example 4 (other types of sources). Other types of sources include stationary, ergodic sources, and various mixing sources. Their exact definition needs some knowledge about measure theory which beyonds the scope of this course.

- Analog sources

An analog source is a signal in which both amplitude and time vary continuously over their respective intervals. Thus an analog source can be represented by a real function $X(t)$. Some examples of analog sources are audio and video signals. Usually, one assumes that $X(t)$ is a stochastic process. If $X(t)$ is a WSS process with power spectral density $S_X(f)$ band-limited to W (i.e., $S_X(f) = 0$ for any $|f| \geq W$), then one can convert $X(t)$ into a random sequence without loss of information. Consider the following diagram

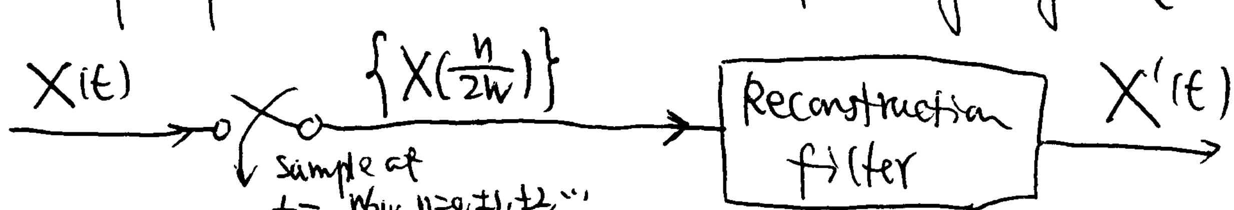


Fig. 2.2: Sampling and reconstruction of a WSS process.

In Fig 2.2, the reconstructed process $X'(t)$ is given by

$$X'(t) = \sum_{n=-\infty}^{+\infty} X\left(\frac{n}{2W}\right) \frac{\sin\left[2\pi W\left(t-\frac{n}{2W}\right)\right]}{2\pi W\left(t-\frac{n}{2W}\right)}$$

Under the ~~cas~~ condition that $S_X(f)$ is band-limited to W, the following theorem says that $X'(t)$ is equal to $X(t)$ in the <u>mean-square sense</u> for all ~~the~~ time.

<u>Theorem 2.1</u> Under the condition that $X(t)$ is WSS and $S_X(f)$ is band-limited to W,

$$E\left[\left(X(t)-X'(t)\right)^2\right] = 0 \quad \text{for all } t. \qquad (2.1.5)$$

Thus in this case, it suffices to consider the coding of the resulting random sequence $\{X(\frac{n}{2W})\}$.

§2.2 <u>Entropy and Mutual Information</u>

In this section, we define several basic information quantities: entropy, conditional entropy, and mutual information. Their meanings will be explained later in this course. <u>We begin with discrete random variables.</u>

Entropy: Given a discrete random variable $X \sim (\mathcal{X}, p(x))$, the entropy $H(X)$ of the discrete random variable X is defined by

$$H(X) \triangleq -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$
$$= -\sum_{j=0}^{J-1} p(x_j)\log p(x_j) \qquad (2.2.1)$$

$0\log 0 = 0$ since $y\log y \to 0$ as $y \to 0$

where $\mathcal{X} = \{x_0, x_1, \cdots, x_{J-1}\}$. For each letter $x_j \in \mathcal{X}$, the quantity $I(x_j) \triangleq -\log p(x_j)$ is called the self-information of the event $X = x_j$. Thus the entropy $H(X)$ is equal to the average self-information

$$H(X) = E[I(X)] \qquad (2.2.2)$$

Remarks

(1) The units of entropy and self-information are determined by the base of the logarithm, which is usually selected as either 2 or e. When the base of the logarithm is 2, the units of entropy and self-information are bits; when the base of the logarithm is e, the units of entropy and self-information are called nats (natural units).

(2) If X is the output of a discrete source, then the entropy $H(X)$ measured in terms of bits represents the average amount of information emitted by the source. In the context of lossless data compression, $H(X)$ represents the ultimate compression rate in bits/letter.

(3) The standard abbreviation for $\log_e$ is $\ln$. The natural logarithm is convenient for computation. Note that

$$\ln a = \ln 2 \log_2 a \qquad \forall a > 0 . \qquad (2.2.3)$$

Thus

$$1 \text{ nat} = \log_2 e \text{ bits}$$

and

$$1 \text{ bit} = \ln 2 = 0.69315 \text{ nats}.$$

In this course, ~~to~~ we shall write $\underline{\log_e a}$ as $\underline{\ln a}$ and $\underline{\log_2 a}$ as $\underline{\log a}$.

(4) Entropy, information, and uncertainty are all related. ~~Considere~~ Consider the event $X = x_j$. Before the ~~re~~ event $X = x_j$ occurs, $I(x_j)$ represents the amount of uncertainty of the event $X = x_j$, and $H(X)$ represents the average amount of uncertainty of the random variable X. After the event

$X = x_j$ occurs, the occurrence of this particular event gives us the amount of information $I(x_j)$. $H(X)$ represents the average amount of information obtained from the corresponding experiment.

(5) Note that $H(X)$ depends only on its probability mass function $p(x)$, $x \in \mathcal{X}$. ~~If we let denote $p(x_j)$ by P_j, then~~ Denote $p(x_j)$ by P_j. We sometimes also write $H(P_0, P_1, \cdots, P_{J-1})$ for $H(X)$. That is,

$$H(P_0, P_1, \cdots, P_{J-1}) = -\sum_{j=0}^{J-1} P_j \log P_j \qquad (2.2.4)$$

<u>Example 1</u> Let

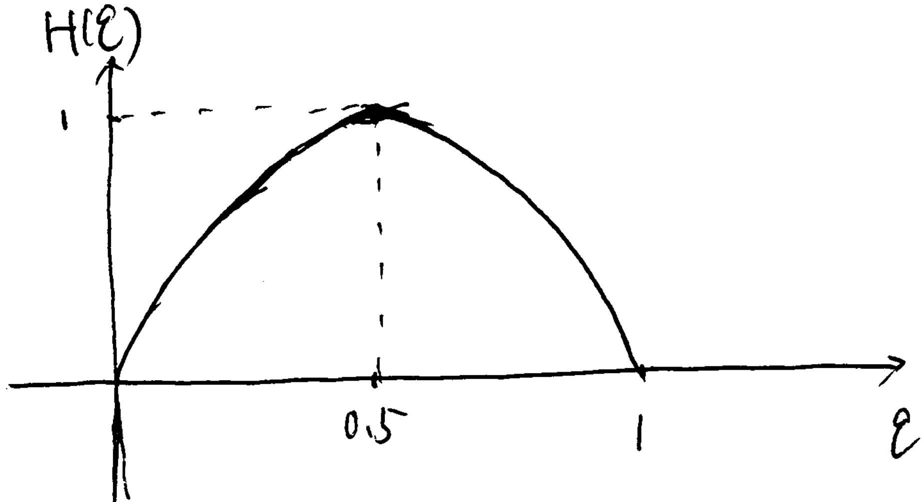
$$X = \begin{cases} 0 & \text{with prob. } \varepsilon \\ 1 & \text{with prob. } 1-\varepsilon. \end{cases}$$

Then

$$H(X) = \underbrace{-\varepsilon \log \varepsilon - (1-\varepsilon)\log(1-\varepsilon)}_{\text{also } \cancel{\text{deto}} \text{ denoted by } H(\varepsilon)} \qquad (2.2.5)$$

$$\triangleq H(\varepsilon)$$

In particular, $H(X) = 1$ bit when $\varepsilon = \frac{1}{2}$. (Thus two equally likely messages contain 1 bit information.) The graph of the function $H(\varepsilon)$ is shown in Fig 2.3.



~~the~~ Fig 2.3 Binary entropy function

The function $H(\epsilon)$ is refered to as the binary entropy function in information theory. It is concave and equal to 0 if $\epsilon = 0$ or 1. This makes sense, because when $\epsilon = 0$ or 1, the variable is not random and there is no uncertainty. Similarly, the uncertainty is maximized when $\epsilon = \frac{1}{2}$.

Example 2

$$X = \begin{cases} a & \text{with prob. } 1/2 \\ b & \text{with prob. } 1/4 \\ c & \text{with prob. } 1/8 \\ d & \text{with prob. } 1/8 \end{cases}$$

$$\begin{aligned} H(X) &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} \\ &= \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = \frac{7}{4} \text{ bits} \end{aligned}$$

$\boxed{\text{H(X) is also related to the problem of guessing the random output X.}}$

<u>Joint Entropy and Conditional Entropy</u>: Given a pair of random variables X and Y with a joint pmf $p(x,y)$, $x \in \mathcal{X} = \{x_0, x_1, \cdots, x_{J-1}\}$, and $y \in \mathcal{Y} = \{y_0, y_1, \cdots, y_{K-1}\}$, the <u>joint entropy</u> $H(X,Y)$ of (X,Y) is defined as

$$H(X,Y) \triangleq - \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log p(x,y)$$

$$= - \sum_{j=0}^{J-1} \sum_{i=0}^{K-1} p(x_j, y_i) \log p(x_j, y_i). \qquad (2.2.6)$$

The <u>conditional entropy</u> $H(X|Y)$ of X given Y is defined as

$$H(X|Y) \triangleq \sum_{y\in\mathcal{Y}} p(y) H(X|Y=y)$$

$$= \sum_{y\in\mathcal{Y}} p(y) \sum_{x\in\mathcal{X}} \left[-p(x|y) \log p(x|y)\right]$$

$$= - \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(y) p(x|y) \log p(x|y)$$

$$= - \sum_{y\in\mathcal{Y}} \sum_{x\in\mathcal{X}} p(x,y) \log p(x|y)$$

$$= - \sum_{i=0}^{K-1} \sum_{j=0}^{J-1} p(x_j, y_i) \log p(x_j|y_i), \qquad (2.2.7)$$

where $p(x|y)$ is the conditional probability of the event $X = x$ given the even $Y=y$. For any x_j and y_i, the quantity $I(x_j|y_i) = -\log p(x_j|y_i)$ is called the <u>conditional self-information</u> of the event $X = x_j$ given $Y = y_i$. Thus the conditional entropy $H(X|Y)$ is equal to the <u>average conditional self-information</u>

$$H(X|Y) = E[I(X|Y)]. \qquad (2.2.8)$$

- <u>Chain rule</u> Since for any $x_j \in \mathcal{X}$ and $y_i \in \mathcal{Y}$, $p(x_j, y_i) = p(y_i)\, p(x_j|y_i)$, it follows that

$$\begin{aligned} -\log p(x_j, y_i) &= -\log p(y_i) - \log p(x_j|y_i) \\ &= I(y_i) + I(x_j|y_i). \end{aligned}$$

Taking expectation on the both sides yields

$$\boxed{H(X,Y) = H(Y) + H(X|Y). \qquad (2.2.9)}$$

Similarly, one can show that

$$\boxed{\begin{aligned} H(X,Y) &= H(X) + H(Y|X) & (2.2.10) \\ H(X,Y|Z) &= H(X|Z) + H(Y|X,Z) & (2.2.11) \\ &= H(Y|Z) + H(X|Y,Z) & (2.2.12) \end{aligned}}$$

Example. Let (X,Y) have the following joint pmf

$Y \backslash X$	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

$\Rightarrow$ The marginal pdf of X is $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$ and the marginal pdf of Y is $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Thus

$$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} = \frac{7}{4} \text{ bits}$$

$$H(Y) = -\frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{4}\log\frac{1}{4} = 2 \text{ bits}$$

$$\begin{aligned} H(X,Y) = & -\frac{1}{8}\log\frac{1}{8} - \frac{1}{16}\log\frac{1}{16} - \frac{1}{32}\log\frac{1}{32} - \frac{1}{32}\log\frac{1}{32} \\ & -\frac{1}{16}\log\frac{1}{16} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{32}\log\frac{1}{32} - \frac{1}{32}\log\frac{1}{32} \\ & -\frac{1}{16}\log\frac{1}{16} - \frac{1}{16}\log\frac{1}{16} - \frac{1}{16}\log\frac{1}{16} - \frac{1}{16}\log\frac{1}{16} \\ & -\frac{1}{4}\log\frac{1}{4} \\ = & \frac{3}{4} + \frac{6\times 4}{16} + \frac{4\times 5}{32} + \frac{2}{4} = \frac{3}{4} + \frac{6}{4} + \frac{5}{8} + \frac{2}{4} \\ = & \frac{27}{8} \text{ bits} \end{aligned}$$

$$H(X|Y) = \frac{1}{4}\cdot H(\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{8}, \tfrac{1}{8}) + \frac{1}{4}\cdot H(\tfrac{1}{4}, \tfrac{1}{2}, \tfrac{1}{8}, \tfrac{1}{8}) + \frac{1}{4}\cdot H(\tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}, \tfrac{1}{4}) + \frac{1}{4} H(1,0,0,0)$$

$$= \frac{1}{4}\cdot\frac{7}{4} + \frac{1}{4}\cdot\frac{7}{4} + \frac{1}{4}\cdot 2 = \frac{11}{8} \text{ bits}$$

$$\underline{H(Y) + H(X|Y)} = 2 + \frac{11}{8} = \frac{27}{8} = \underline{H(X,Y)}.$$

Extension: Suppose now we have n random variables $X_1, X_2, \cdots, X_n$ with joint pdf $p(u_1, u_2, \cdots, u_n)$, $u_i \in \mathcal{X}_i$. The joint entropy $H(X_1, X_2, \cdots, X_n)$ is defined as

$$H(X_1, X_2, \cdots, X_n) = E[-\log p(X_1, X_2, \cdots, X_n)]$$

$$= -\sum_{u_1\in\mathcal{X}_1}\sum_{u_2\in\mathcal{X}_2}\cdots\sum_{u_n\in\mathcal{X}_n} p(u_1, u_2, \cdots, u_n)\log p(u_1, u_2, \cdots, u_n) \quad (2.2.13)$$

Since

$$p(u_1, u_2, \cdots, u_n) = p(u_1)\, p(u_2|u_1) \cdots p(u_n|u_1, u_2, \cdots, u_{n-1}),$$

it follows that

$$H(X_1, X_2, \cdots, X_n) = E[-(\log p(X_1) + \log p(X_2|X_1) + \cdots + p(X_n|X_1, \cdots, X_{n-1}))]$$

$$= H(X_1) + H(X_2|X_1) + \cdots + H(X_n|X_1, \cdots, X_{n-1}). \quad (2.2.14)$$

- Mutual Information. The self-information $I(x_j)$ is the amount of uncertainty ~~abouth~~ about the event $X = x_j$. The conditional self-information $I(x_j|y_i)$ is the amount of

~~information~~ uncertainty remaining about the event $X = x_j$ after we observe the outcome $Y = y_i$. Thus the difference $I(x_j) - I(x_j|y_i)$ is the amount of information that our observation $Y = y_i$ ~~provive~~ provides about the event $X = x_j$. We denote this difference by $I(x_j; y_i)$. That is

$$I(x_j; y_i) = I(x_j) - I(x_j|y_i).$$

Since

$$\begin{aligned} I(x_j; y_i) &= -\log p(x_j) + \log p(x_j|y_i) \\ &= \log \frac{p(x_j|y_i)}{p(x_j)} \\ &= \log \frac{p(x_j, y_i)}{p(x_j)p(y_i)} \qquad (2.2.15) \\ &= \log \frac{p(y_i|x_j)}{p(y_i)} = I(y_i) - I(y_i|x_j) \\ &= I(y_i; x_j) \qquad (2.2.15^+) \end{aligned}$$

The quantity $I(x_j; y_i)$ is called the mutual information between the events $X = x_j$ and $Y = y_i$. The average mutual information is denoted by $I(X; Y)$

$$\begin{aligned} I(X;Y) &= \sum_{j=0}^{J-1} \sum_{i=0}^{K-1} p(x_j, y_i) I(x_j; y_i) \\ &= \sum_{j=0}^{J-1} \sum_{i=0}^{K-1} p(x_j, y_i) \log \frac{p(x_j, y_i)}{p(x_j)p(y_i)} \qquad (2.2.16). \end{aligned}$$

So ~~the~~ $I(X;Y)$ is the average amount of information that Y provides about X and also the averge amount of information that X provides about Y.

~~• Property~~

• Properties of Entropy and Mutual Information

(1) $0 \le H(X) \le \log J$. Furthermore, $H(X)=0$ if and only if there exists a j_0 such that ~~$p(x_{j_0})=1$~~ $p(x_{j_0})=1$ and $p(x_j)=0$ for any $j \ne j_0$. $H(X)=\log J$ iff all outcomes are equally likely, i.e,

$$p(x_j) = \frac{1}{J} \quad \forall\, 0 \le j \le J-1.$$

(2) $I(X;Y) = I(Y;X) = H(X) - H(X|Y)$
$= H(Y) - H(Y|X)$

(3) $I(X;Y) \ge 0$ with equality iff X and Y are independent.

(4) $I(X;Y) = H(X) + H(Y) - H(X;Y)$

(5) The relationship between entropy, the joint entropy, conditional entropy, and mutual information can be represented by the following diagram.

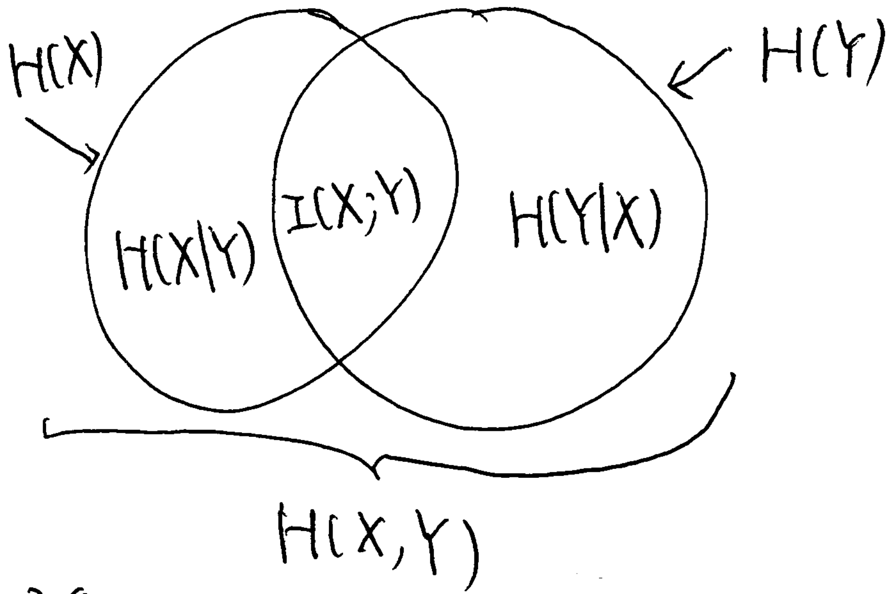


Fig.2.4. Relationship between entropy and mutual information

Properties (2) and (4) are straightforward. Properties (1) and (3) can be proved by applying the following log-sum inequality.

Theorem 2.2.1 (Log sum inequality) For non-negative numbers $a_1, a_2, \cdots, a_n$ and $b_1, b_2, \cdots, b_n$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \qquad (2.2.17)$$

with equality if and only if $\frac{a_i}{b_i}$ = constant.

Both the proofs of ~~Pro~~ Properties (1) and (3) and the proof of Theorem 2.2.1 are left for exercises.

Property (3) also implies that

Property (6): $H(X_1, X_2, \cdots, X_n) \leq \sum_{i=1}^{n} H(X_i)$ with equality iff X_i are independent

- Information Quantities for Continuous Random Variables

Let (X, Y) be a pair of real random variables with joint pdf $p(x, y)$ and marginal ~~p(x) and~~ pdfs $p(x)$ and $p(y)$. In analogy to the discrete case, we define

$$I(X;Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dx\,dy \qquad (2.2.18)$$

the (average) mutual information between X and Y

$$H(X) = -\int_{-\infty}^{+\infty} p(x)\log p(x)\,dx \qquad (2.2.19)$$

the differential entropy of X

$$H(Y) = -\int_{-\infty}^{+\infty} p(y)\log p(y)\,dy$$

the differential entropy of Y

$$H(X|Y) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p(x,y)\log p(x|y)\,dx\,dy \qquad (2.2.20)$$

the conditional differential entropy of X given Y.

Remarks.

(1) The quantity $I(X;Y)$ in the continuous case has the same physical meaning as in the discrete case.

(2) The concept of differential entropy, however, does not have

the physical meaning of the concept of entropy in the discrete case. To see this is the case, note that the differential entropy $H(X)$ may be negative.

(3) When both $H(X)$ and $H(X|Y)$ are finite, we still have

$$I(X;Y) = H(X) - H(X|Y).$$

• <u>Mixed Case</u>. Suppose that X is a discrete ~~ran~~ random variable with pmf $p(x)$, $x \in \mathcal{X} = \{x_0, x_1, \cdots, x_{J-1}\}$, and Y is a continuous random variable with pdf $p(y)$. We may express $p(y)$ as

$$p(y) = \sum_{j=0}^{J-1} p(x_j)\, p(y|x_j)$$

where $p(y|x_j)$ is the conditional pdf of Y given $X = x_j$. In this case, the average ~~information~~ mutual information $I(X;Y)$ between X and Y is given by

$$I(X;Y) = \sum_{j=0}^{J-1} \int_{-\infty}^{+\infty} p(x_j)\, p(y|x_j) \log \frac{p(y|x_j)}{p(y)}\, dy \qquad (2.2.21)$$