Chapter 3
Shannon Theory

3.1 Information Space

According to Shannon, a message x is a random event. Let $p(x)$ be the probability of occurrence of event x. If $p(x) = 0$, this event does not occur; If $p(x) = 1$, this event must occur. When $p(x) = 0$ or $p(x) = 1$, information x can be called trivial information or spam information. Therefore, the real mathematical significance of information x lies in its uncertainty, that is $0 < p(x) < 1$. Quantitative research on the uncertainty of nontrivial information constitutes all the starting point of Shannon's theory; this starting point is now called information quantity or information entropy, or entropy for short. Shannon and his colleagues at Bell laboratory considered "bit" as the basic quantitative unit of information. What is "bit"? We can simply understand it as the number of bits in the binary system. However, according to Shannon, the binary system with n digits can express up to 2^n numbers. From the point of view of probability and statistics, the probability of occurrence of these 2^n numbers is $\frac{1}{2^n}$. Therefore, a bit is the amount of information contained in event x with probability $\frac{1}{2}$. Taking this as the starting point, Shannon defined the self-information $I(x)$ contained in an information x as

$$I(x) = -\log_2 p(x). \tag{3.1}$$

Therefore, one piece of information x contains $I(x)$-bit information, when $p(x) = \frac{1}{2}$, then $I(x) = 1$. Equation (3.1) is Shannon's first extraordinary progress in information quantification. On the other hand, with the emergence of Telegraph and telephone, binary is widely used in the conversion and transmission of information. Therefore, we can assert that without binary, there would be no Shannon's theory, let alone the current informatics and information age. The purpose of this section is to strictly mathematically deduce and simplify the most basic and important conclusions in Shannon's theory. First, we start with the rationality of the definition of formula (3.1).

If $I(x)$ is used to represent the self-information of a random event x, the greater the probability of occurrence $p(x)$, the smaller the uncertainty. Therefore, $I(x)$ should be a monotonic decreasing function of probability $p(x)$. If xy is a joint event and

Z. Zheng, *Modern Cryptography Volume 1*, Financial Mathematics and Fintech,
https://doi.org/10.1007/978-981-19-0920-7_3

is statistically independent, that is, $p(xy) = p(x)p(y)$, then the self-information amount is $I(xy) = I(x) + I(y)$. Of course, the self-information amount $I(x)$ is nonnegative, that is $I(x) \geq 0$. Shannon prove, the self-information $I(x)$ satisfying the above three assumptions must be

$$I(x) = -c \log p(x),$$

where c is a constant. This conclusion can be derived directly from the following mathematical theorems.

Lemma 3.1 *If the real function $f(x)$ satisfies the following conditions in interval $[1, +\infty)$:*

(i) $f(x) \geq 0$,
(ii) *If* $x < y \Rightarrow f(x) < f(y)$,
(iii) $f(xy) = f(x) + f(y)$.

Then $f(x) = c \log x$, where c is a constant.

Proof Repeated use condition (iii), then there is

$$f(x^k) = kf(x),\ k \geq 1$$

for any positive integer k. Take $x = 1$, then the above formula holds if and only if $f(1) = 0$. It can be seen from (ii) that $f(x) > 0$ when $x > 1$. Let $x > 1$, $y > 1$ and $k \geq 1$ given, you can always find a nonnegative integer n to satisfy

$$y^n \leq x^k < y^{n+1},$$

Take logarithms on both sides to get

$$\frac{n}{k} \leq \frac{\log x}{\log y} < \frac{n+1}{k},$$

On the other hand, we have

$$nf(y) \leq kf(x) < (n+1)f(y),$$

thus

$$\left|\frac{f(x)}{f(y)} - \frac{\log x}{\log y}\right| \leq \frac{1}{k},$$

when $k \to \infty$, we have

$$\frac{f(x)}{f(y)} = \frac{\log x}{\log y},\ \forall\, x, y \in (1, +\infty).$$

Therefore,

$$\frac{f(x)}{\log x} = \frac{f(y)}{\log y} = c,\ \forall\, x, y \in (1, +\infty).$$

That is $f(x) = c \log x$. The Lemma holds.

In Lemma 3.1, let $I(x) = f(\frac{1}{p(x)})$, then $f(x)$ satisfies the condition (i), (ii) and (iii), thus $I(x) = -c \log p(x)$. That is (3.1) holds.

In order to introduce the definition of information space, we use X to represent a finite set of original information, or a countable and additive information set, which is called source state set. It can be an alphabet, a finite number of symbols or a set of numbers. For example, 26 letters in English and 2-element finite field $\mathbb{F}_2$ are commonly used source state sets. Elements in X can be called messages, events, etc., or characters. We often use English capital letters such as X, Y, Z to represent a source state set, and lowercase Greek letters $\xi, \eta, \ldots$ to represent a random variable in a given probability space.

Definition 3.1 The value space of a random variable ξ is a source state set X; the probability distribution of characters on X as events is defined as

$$p(x) = P\{\xi = x\},\ \forall\, x \in X. \tag{3.2}$$

We call (X, ξ) an information space in a given probability space, when the random variable ξ is clear, we usually record the information space (X, ξ) as X. If η is another random variable valued on X, and ξ and η obey the same probability distribution, that is

$$P\{\xi = x\} = P\{\eta = x\},\ \forall\, x \in X.$$

Call two information spaces $(X, \xi) = (X, \eta)$, usually recorded as X.

As can be seen from Definition 3.1, an information space X constitutes a finite complete event group, that is, we have

$$\sum_{x \in X} p(x) = 1,\ 0 \le p(x) \le 1,\ x \in X. \tag{3.3}$$

It should be noted that if there are two random variables ξ and η with values on X, when the probability distributions obeyed by ξ and η are not equal, then (X, ξ) and (X, η) are two different information spaces; at this point, we must distinguish the two different information spaces with $X_1 = (X, \xi)$ and $X_2 = (X, \eta)$.

Definition 3.2 X and Y are two source state sets, and the random variables ξ and η are taken on X and Y, respectively; if ξ and η are compatible random variables, the probability distribution of joint event $xy (x \in X, y \in Y)$ is defined as

$$p(xy) = P\{\xi = x, \eta = y\},\ \forall\, x \in X,\ y \in Y. \tag{3.4}$$

Then, we call the joint event set

$$XY = \{xy | x \in X, y \in Y\}$$

Together with the corresponding random variables ξ and η, it is called the product space of information space (X, ξ) and (Y, η), denote as (XY, ξ, η), when ξ and η are clear, they can be abbreviated as $XY = (XY, \xi, \eta)$. If $X = Y$ are two identical source state sets, ξ and η have the same probability distribution, then the product space XY is denoted as X^2 and is called a power space.

Since the information space is a complete set of events, defined by the product information space, we have the following full probability formula and probability product formula:

$$\begin{cases} \sum\limits_{x \in X} p(yx) = p(y), \ \forall\, y \in Y \\ \sum\limits_{y \in Y} p(xy) = p(x), \ \forall\, x \in X. \end{cases} \tag{3.5}$$

And

$$p(x)p(y|x) = p(xy), \ \forall\, x \in X, y \in Y.$$

Where $p(y|x)$ is the conditional probability of y under the condition of x.

Definition 3.3 Let $X_1, X_2, \ldots, X_n (n \geq 2)$ be n source state sets, $\xi_1, \xi_2, \ldots, \xi_n$ be n compatible random variables with values, respectively, in X_i, the probability distribution of joint event $x_1 x_2 \cdots x_n$ is

$$p(x_1 x_2 \cdots x_n) = P\{\xi_1 = x_1, \xi_2 = x_2, \ldots, \xi_n = x_n\}. \tag{3.6}$$

Then called

$$X_1 X_2 \cdots X_n = \{x_1 x_2 \cdots x_n | x_i \in X_i, \ 1 \leq i \leq n\}$$

are the product of n information spaces, especially when $X_1 = X_2 = \cdots = X_n = X$, and each ξ_i has the same probability distribution on X, define $X^n = X_1 X_2 \cdots X_n$, called the n-th power space of information space X.

Let us give some classic examples of information space.

Example 3.1 (*Two point information space with parameter* λ) Let $X = \{0, 1\} = \mathbb{F}_2$ be a binary finite field, the random variable ξ taken on X is subject to the two-point distribution with parameter λ, that is

$$\begin{cases} p(0) = P\{\xi = 0\} = \lambda, \\ p(1) = P\{\xi = 1\} = 1 - \lambda. \end{cases}$$

where $0 < \lambda < 1$, then (X, ξ) is called a two-point information space with parameter λ, still denote as X.

Example 3.2 (*Equal probability information space*) Let $X = \{x_1, x_2, \ldots, x_n\}$ be a source state sets, the random variable ξ on X obeys the equal probability distribution, that is

$$p(x) = P\{\xi = x\} = \frac{1}{|X|}, \ \forall\, x \in X.$$

Then (X, ξ) is called equal probability information space, still denote as X.

Example 3.3 (*Bernoulli information space*) Let $X_0 = \{0, 1\} = \mathbb{F}_2$. Let the random variable ξ_i be the i-th Bernoulli test; therefore, $\{\xi_i\}_{i=1}^n$ is a set of independent and identically distributed random variables. We let the product space

$$X = (X_0, \xi_1)(X_0, \xi_2)\cdots(X_0, \xi_n) = X_0^n \subset \mathbb{F}_2^n,$$

the power space X is called Bernoulli information space, also alled memoryless binary information space. The probability function $p(x)$ in X is

$$p(x) = p(x_1 x_2 \cdots x_n) = \prod_{i=1}^{n} p(x_i), \ \ x_i = 0 \text{ or } 1. \tag{3.7}$$

where $p(0) = \lambda, \ \ p(1) = 1 - \lambda$.

Example 3.4 (*Degenerate information space*) If $X = \{x\}$, it contains only one character. X is called a degenerate information space, or trivial information space. The random variable ξ takes the value x of probability 1, that is $P\{\xi = x\} = 1$. At this time, ξ is a random variable with degenerate distribution in probability.

Definition 3.4 Let $X = \{x_1, x_2, \ldots, x_n\}$ be a source state sets, if X is an information space, the information entropy $H(X)$ of X is defined as

$$H(X) = -\sum_{x \in X} p(x) \log p(x) = -\sum_{i=1}^{n} p(x_i) \log p(x_i), \tag{3.8}$$

if $p(x_i) = 0$ in the above formula, we agreed that $p(x_i) \log p(x_i) = 0$, the base of logarithm can be selected arbitrarily; if the base of the logarithm is $D (D \geq 2)$, then $H(X)$ is called D-ary entropy, sometimes denote as $H_D(X)$.

Theorem 3.1 *For any information space X, always have*

$$0 \leq H(X) \leq \log |X|. \tag{3.9}$$

And $H(X) = 0$ if and only if X is a degenerate information space, $H(X) = \log |X|$ if and only if X is a equal probability information space.

Proof $H(X) \geq 0$ is trivial. We only prove the inequality on the right of Eq. (3.9). Because $f(x) = \log x$ is a strictly convex real value, from the Lemma 1.7 in Chap. 1, thake $g(x) = \frac{1}{p(x)}$ is a positive function, $p(x) > 0$, thus let $X = \{x_1, x_2, \ldots, x_m\}$,

$$H(X) = \sum_{i=1}^{m} p(x_i) \log \frac{1}{p(x_i)} \leq \log \sum_{i=1}^{m} \frac{p(x_i)}{p(x_i)} = \log m.$$

The above equal sign holds if and only if $p(x_1) = p(x_2) = \cdots = p(x_m) = \frac{1}{m}$, that is, X is equal probability information space. If $X = \{x\}$ is a degenerate information space, because $p(x) = 1$, so $H(X) = 0$. Conversely, if $H(X) = 0$, let $X = \{x_1, x_2, \ldots, x_m\}$, suppose $\exists\, x_i \in X$, such that $0 < p(x_i) < 1$, then

$$0 < p(x_i) \log \frac{1}{p(x_i)} \leq H(X).$$

So there is $p(x_i) = 1$, but $p(x_j) = 0 (j \neq i)$; at this time, X degenerates into $X = \{x_i\}$, which is a trivial information space, the Lemma holds.

An information space is a dynamic code (which changes with the change of the random variable on it). For "dynamic code", that is, the code rate of information space X, Shannon replaces $\frac{1}{n} H(X)$ with information entropy, so information entropy $H(X)$ becomes the first mathematical quantity to describe dynamic code. From Theorem 3.1, when the code is degenerate, the minimum rate of a dynamic code is 0, when the code is equal probability, the maximum rate is the rate of the usual static code.

Next, we discuss the information entropy of several typical information spaces.

Example 3.5 (i) Let X be the two-point information space of parameter λ, then

$$H(X) = -\lambda \log \lambda - (1 - \lambda) \log(1 - \lambda) = H(\lambda).$$

$H(\lambda)$ we defined it in Chap. 1, it was called binary information entropy function at that time. Now we know why it is called entropy function

(ii) $X = \{x\}$ is degraded information space, then $H(X) = 0$.
(iii) When X is equal overview information space, then $H(X) = \log |X|$.

Remark Most authors directly regard a random variable as an information space. Mathematically, it is convenient to do so and call it the information measurement of random variables. However, from the perspective of information, using the concept of information space can better understand and simplify Shannon's theory; the core idea of this theory is the random measurement of information, not the information measurement of random variables.

3.2 Joint Entropy, Conditional Entropy, Mutual Information

Definition 3.5 Let X, Y be two information spaces, and ξ, η be random variables with corresponding values, respectively. If ξ and η are independent random variables, that is

$$P\{\xi = x, \eta = y\} = P\{\xi = x\} \cdot P\{\eta = y\},\ \forall\, x \in X, y \in Y.$$

X and Y are called independent information space, and the probability distribution of joint events is

$$p(xy) = p(x)p(y),\ \forall x \in X, y \in Y.$$

Definition 3.6 Let X, Y be two information spaces, the information entropy $H(XY)$ of the product space XY is called the joint entropy of X and Y, that is

$$H(XY) = -\sum_{x\in X}\sum_{y\in Y} p(xy)\log p(xy). \tag{3.10}$$

The conditional entropy $H(X|Y)$ of X versus Y is defined as

$$H(X|Y) = -\sum_{x\in X}\sum_{y\in Y} p(xy)\log p(x|y). \tag{3.11}$$

Lemma 3.2 (Addition formula of entropy) *For any two information spaces X and Y, then we have*

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

Generally, for n information spaces $X_1, X_2, \ldots, X_n$, we have

$$H(X_1X_2\cdots X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}X_{i-2}\cdots X_1). \tag{3.12}$$

Proof By (3.10) and probability multiplication formula,

$$\begin{aligned} H(XY) &= -\sum_{x\in X}\sum_{y\in Y} p(xy)\log p(xy) \\ &= -\sum_{x\in X}\sum_{y\in Y} p(xy)(\log p(x) + \log p(y|x)) \\ &= -\sum_{x\in X} p(x)\log p(x) + H(Y|X) \\ &= H(X) + H(Y|X). \end{aligned}$$

The same can be proved

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

We prove (3.12) by induction, when $n = 2$,

$$H(X_1X_2) = H(X_1) + H(X_2|X_1).$$

The proposition is true, and for general n, we have

$$\begin{aligned} H(X_1X_2\cdots X_n) &= H(X_1X_2\cdots X_{n-1}) + H(X_n|X_1X_2\cdots X_{n-1}) \\ &= \sum_{i=1}^{n-1} H(X_i|X_{i-1}X_{i-2}\cdots X_1) + H(X_n|X_1X_2\cdots X_{n-1}) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1}X_{i-2}\cdots X_1). \end{aligned}$$

The Lemma 3.2 holds.

Theorem 3.2 *We have*

$$H(XY) \leq H(X) + H(Y). \tag{3.13}$$

If and only if X and Y are statistically independent information spaces,

$$H(XY) = H(X) + H(Y). \tag{3.14}$$

Generally, we have

$$H(X_1X_2\cdots X_n) \leq H(X_1) + H(X_2) + \cdots + H(X_n). \tag{3.15}$$

If and only if $X_1, X_2, \ldots, X_n$ is an independent random process,

$$H(X_1X_2\cdots X_n) = H(X_1) + H(X_2) + \cdots + H(X_n). \tag{3.16}$$

Proof By definition and Jensen inequality, we have

$$\begin{aligned} H(XY) - H(X) - H(Y) &= \sum_{x\in X}\sum_{y\in Y} p(xy)\log\frac{p(x)p(y)}{p(xy)} \\ &\leq \log\sum_{x\in X}\sum_{y\in Y} p(x)p(y) \\ &= 0. \end{aligned}$$

The above equal sign holds, if and only if for all $x \in X$, $y \in Y$, $\frac{p(x)p(y)}{p(xy)} = c$(where c is a constant), thus $p(x)p(y) = cp(xy)$. Both sides sum at the same time, we have

$$1 = \sum_{x\in X} p(x)\sum_{y\in Y} p(y) = c\sum_{x\in X}\sum_{y\in Y} p(xy),$$

thus $c = 1$, $p(xy) = p(x)p(y)$. So if and only if X and Y are independent information spaces, (3.14) holds. By induction, we have (3.15) and (3.16). Theorem 3.2 holds.

By (3.15), we have the following direct corollary; for any information space X and $n \geq 1$, we have

$$H(X^n) \leq nH(X). \tag{3.17}$$

Definition 3.7 Let X and Y be two information spaces, and say that X is completely determined by Y, if there is always a subset $N_x \subset Y$ of Y for any given $x \in X$, satisfies

$$\begin{cases} p(x|y) = 1, & \text{if } y \in N_x; \\ p(x|y) = 0, & \text{if } y \notin N_x. \end{cases} \tag{3.18}$$

With regard to conditional information entropy $H(X|Y)$, we have the following two important special cases.

Lemma 3.3 *(i)* $0 \leq H(X|Y) \leq H(X)$.
(ii) If the information space X is completely determined by Y, then

$$H(X|Y) = 0. \tag{3.19}$$

(iii) If X and Y are two separate information spaces,

$$H(X|Y) = H(X). \tag{3.20}$$

Proof (i) is trivial. Let us prove (3.19) first. By Definition 3.7 and (3.18), for given $x \in X$, we have

$$p(xy) = p(y)p(x|y) = 0, \ y \notin N_x.$$

Thus

$$\begin{aligned} H(X|Y) &= -\sum_{x\in X}\sum_{y\in Y} p(xy)\log p(x|y) \\ &= -\sum_{x\in X}\sum_{y\in N_x} p(xy)\log p(x|y) = 0. \end{aligned}$$

The proof of the formula (3.20) is obvious. Because X and Y are independent, the conditional probability

$$p(x|y) = p(x), \ \forall\, x \in X, y \in Y.$$

Thus

$$\begin{aligned} H(X|Y) &= -\sum_{x\in X}\sum_{y\in Y} p(x)p(y)\log p(x) \\ &= -\sum_{x\in X} p(x)\log p(x) = H(X). \end{aligned}$$

The Lemma 3.3 holds.

Next, we define the mutual information $I(X, Y)$ of two information spaces X and Y.

Definition 3.8 Let X and Y be two information spaces, and then their mutual information $I(X, Y)$ is defined as

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(xy) \log \frac{p(x|y)}{p(x)}. \tag{3.21}$$

From the multiplication formula of probability, for all $x \in X$, $y \in Y$,

$$p(x)p(y|x) = p(y)p(x|y) = p(xy).$$

We have

$$\frac{p(x|y)}{p(x)} = \frac{p(y|x)}{p(y)}.$$

Therefore, there is a direct conclusion from the definition of mutual information $I(X, Y)$

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

Lemma 3.4

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

Proof By definition,

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

The same can be proved

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

Lemma 3.5 *Assuming that X and Y are two information spaces, $I(X, Y)$ is the amount of mutual information, then*

$$H(XY) = H(X) + H(Y) - I(X, Y). \tag{3.22}$$

Further, we have $I(X,Y) \geq 0$, if and only if X and Y are independent, $I(X,Y) = 0$.

Proof By the addition formula of Lemma 3.2,

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

The conclusion about $I(X,Y) \geq 0$ can be deduced directly from Theorem 3.2.

Let us prove an equation about entropy commonly used in the statistical analysis of cryptography.

Theorem 3.3 *If X, Y, Z are three information spaces, then*

$$\begin{aligned} H(XY|Z) &= H(X|Z) + H(Y|XZ) \\ &= H(Y|Z) + H(X|YZ). \end{aligned} \tag{3.23}$$

Proof By the definition, we have

$$H(XY|Z) = -\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(xyz)\log p(xy|z).$$

By probability product formula,

$$p(xyz) = p(z)p(xy|z) = p(xz)p(y|xz).$$

Thus

$$p(xy|z) = \frac{p(xz)p(y|xz)}{p(z)} = p(x|z)p(y|xz).$$

So we have

$$\begin{aligned} H(XY|Z) &= -\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(xyz)\log p(x|z)p(y|xz) \\ &= -\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(xyz)(\log p(x|z) + \log p(y|xz)) \\ &= -\sum_{x\in X}\sum_{z\in Z} p(xz)\log p(x|z) - \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(xyz)\log p(y|xz) \\ &= H(X|Z) + H(Y|XZ). \end{aligned}$$

Similarly, the second formula can be proved.

Finally, we extend the formula (3.15) to conditional entropy.

Lemma 3.6 *Let $X_1, X_2, \ldots, X_n, Y$ be information spaces, then we have*

$$H(X_1X_2\cdots X_n|Y) \le H(X_1|Y) + \cdots + H(X_n|Y). \tag{3.24}$$

Specially, when $X_1 = X_2 = \cdots = X_n = X$,

$$H(X^n|Y) \le nH(X|Y). \tag{3.25}$$

Proof We make an induction of n. The proposition is trivial when $n = 1$. Let the proposition be true when n, i.e.,

$$H(X_1X_2\cdots X_n|Y) \le H(X_1|Y) + \cdots + H(X_n|Y).$$

Then when $n+1$, we let $X = X_1X_2\cdots X_n$, then

$$\begin{aligned} &H(X_1X_2\cdots X_{n+1}|Y) = H(XX_{n+1}|Y) \\ &= -\sum_{x\in X}\sum_{z\in X_{n+1}}\sum_{y\in Y} p(xzy)\log p(xz|y). \end{aligned}$$

From the full probability formula,

$$H(X|Y) + H(X_{n+1}|Y) = -\sum_{x\in X}\sum_{z\in X_{n+1}}\sum_{y\in Y} p(xzy)\log p(x|y)p(z|y).$$

So by Jensen inequality,

$$\begin{aligned} &H(XX_{n+1}|Y) - H(X|Y) - H(X_{n+1}|Y) \\ &= \sum_{x\in X}\sum_{z\in X_{n+1}}\sum_{y\in Y} p(xzy)\log\frac{p(x|y)p(z|y)}{p(xz|y)} \\ &\le \log\sum_{x\in X}\sum_{z\in X_{n+1}}\sum_{y\in Y} p(y)p(x|y)p(z|y). \end{aligned}$$

By product formula

$$\begin{aligned} &\sum_{x\in X}\sum_{z\in X_{n+1}}\sum_{y\in Y} p(y)p(x|y)p(z|y) \\ &= \sum_{x\in X}\sum_{y\in Y} p(x|y)p(y) \\ &= \sum_{x\in X} p(x) = 1. \end{aligned}$$

So by the inductive hypothesis,

$$\begin{aligned} H(XX_{n+1}|Y) &\leq H(X_{n+1}|Y) + H(X|Y) \\ &\leq H(X_1|Y) + H(X_2|Y) + \cdots + H(X_{n+1}|Y). \end{aligned}$$

The proposition holds for $n+1$. So the Lemma holds.

3.3 Redundancy

Select a alphabet $\mathbb{F}_q$ or a remaining class ring $\mathbb{Z}_m$ of module m, each element in the alphabet is called character, and in the field of communication, alphabet is also called source state, and character is also called transmission signal. If the length of a q-ary code is increased, redundant transmission signals or characters will appear in each codeword. The digital measurement of "redundant characters" is called redundancy, which is a technical means to improve the accuracy of codeword transmission, and redundancy is an important mathematical quantity to describe this technical means. Therefore, we start by proving the following lemma.

Lemma 3.7 *Let X, Y, Z be three information spaces, then*

$$H(X|YZ) \leq H(X|Z). \tag{3.26}$$

Proof By total probability formula,

$$\begin{aligned} H(X|Z) &= -\sum_{x\in X}\sum_{z\in Z} p(xz)\log p(x|z) \\ &= -\sum_{x\in X}\sum_{z\in Z}\sum_{y\in Y} p(xyz)\log p(x|z). \end{aligned}$$

So

$$\begin{aligned} &H(X|YZ) - H(X|Z) \\ &= \sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(xyz)\log\frac{p(x|z)}{p(x|zy)} \\ &\leq \log\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z}\frac{p(xyz)p(x|z)}{p(x|zy)} \\ &= \log\sum_{x\in X}\sum_{y\in Y}\sum_{z\in Z} p(yz)p(x|z) \\ &= \log\sum_{x\in X}\sum_{z\in Z} p(z)p(x|z) \\ &= 0. \end{aligned}$$

Thus $H(X|YZ) \leq H(X|Z)$. The Lemma holds.

Let X be a source state set and randomly select codewords to enter the channel of information transmission is a discrete random process. This mathematical model can be constructed and studied on X by taking the value of a group of random variables $\{\xi_i\}_{i\geq 1}$. Firstly, we assume that $\{\xi_i\}_{i\geq 1}$ obeys the same probability distribution when taking value on X, and we get a set of information spaces $\{X^i\}_{i\geq 1}$, let $H_0 = \log|X|$ be the entropy of X as the equal probability information space, for $n \geq 1$, we let

$$H_n = H(X|X^{n-1}),\ H_1 = H(X).$$

By Lemma 3.7, then $\{H_n\}$ constitutes a number sequence with monotonic descent and lower bound, so that its limit exists, that is

$$\lim_{n\to\infty} H_n = a\ (a \geq 0). \tag{3.27}$$

We will extend the above observation to the general case: Let $\{\xi_i\}_{i\geq 1}$ be any set of random variables valued on X, for any $n \geq 1$, we let

$$X_n = (X, \xi_n),\ n \geq 1.$$

Definition 3.9 A source state set X has a set of random variables $\{\xi_i\}_{i\geq 1}$ valued on X, then X is called a source.

(i) If $\{\xi_i\}_{i\geq 1}$ is a group of independent and identically distributed random variables, X is called a memoryless source.

(ii) If for any integers $k, t_1, t_2, \ldots, t_k$ and h, random vector

$$(\xi_{t_1}, \xi_{t_2}, \ldots, \xi_{t_k})(\xi_{t_1+h}, \xi_{t_2+h}, \ldots, \xi_{t_k+h})$$

obey the same joint probability distribution, then X is called a stationary source.

(iii) If $\{\xi_i\}_{i\geq 1}$ is a k-order Markov process, that is, for $\forall\ m > k \geq 1$,

$$\begin{aligned} &p(x_m|x_{m-1}x_{m-2}\cdots x_1) \\ &= p(x_m|x_{m-1}x_{m-2}\cdots x_{m-k}),\ \forall\ x_1, x_2, \ldots, x_m \in X, \end{aligned}$$

Then X is called k-order Markov source, specially, $k = 1$, i.e.,

$$p(x_m|x_{m-1}x_{m-2}\cdots x_1) = p(x_m|x_{m-1}),\ \forall\ x_1, x_2, \ldots, x_m \in X,$$

call X Markov source.

The concept from information space to source changes from a single random variable taking value on X to an infinite dimensional random vector, so that the transmission process of code X constitutes a discrete random process. By definition, we have

Lemma 3.8 *Let X be a source state set, and $\{\xi_i\}_{i\geq 1}$ be a set of random variables valued on X, we write*

$$X_i = (X, \xi_i),\ i \geq 1. \tag{3.28}$$

(i) If X is a memoryless source, the joint probability distribution on X satisfies

$$p(x_1 x_2 \cdots x_n) = \prod_{i=1}^{n} p(x_i),\ x_i \in X_i,\ n \geq 1. \tag{3.29}$$

(ii) If X is a stationary source, then for all integers $t_1, t_2, \ldots, t_k (k \geq 1)$ and h, there is the following joint probability distribution,

$$p(x_{t_1} x_{t_2} \cdots x_{t_k}) = p(x_{t_1+h} x_{t_2+h} \cdots x_{t_k+h}), \tag{3.30}$$

where $x_i \in X_i,\ i \geq 1$.

(iii) If X is a stationary Markov source, then the conditional probability distribution on X satisfies for any $m \geq 1$ and $x_1 x_2 \cdots x_m \in X_1 X_2 \cdots X_m$, we have

$$\begin{aligned} &p(x_m | x_1 \cdots x_{m-1}) = p(x_m | x_{m-1}) \\ &= P\{\xi_{i+1} = x_m | \xi_i = x_{m-1}\},\ \forall\ 1 \leq i \leq m-1. \end{aligned} \tag{3.31}$$

Proof (i) and (ii) can be derived directly from the definition. We only prove (iii). By (ii) of the definition 3.9, for $\forall\ i \geq 1$, we have

$$P\{\xi_i = x_{m-1}, \xi_{i+1} = x_m\} = P\{\xi_{m-1} = x_{m-1}, \xi_m = x_m\}$$

and

$$P\{\xi_i = x_{m-1}\} = P\{\xi_{m-1} = x_{m-1}\}.$$

Thus

$$\begin{aligned} &P\{\xi_i = x_{m-1}\} P\{\xi_{i+1} = x_m | \xi_i = x_{m-1}\} \\ &= P\{\xi_{m-1} = x_{m-1}\} P\{\xi_m = x_m | \xi_{m-1} = x_{m-1}\}. \end{aligned}$$

We have

$$P\{\xi_{i+1} = x_m | \xi_i = x_{m-1}\} = p(x_m | x_{m-1}).$$

The Lemma holds.

Corollary 3.1 *A memoryless source X must be a stationary source.*

Proof Derived directly from Definition 3.9.

Next, we extend the limit formula in memoryless sources revealed by formula (3.27) to general stationary sources. For this purpose, we first prove two lemmas.

Lemma 3.9 *Let $\{f(n)\}_{n\geqslant 1}$ be a sequence of real numbers, which satisfies the following semi countable additivity,*

$$f(n+m) \leqslant f(n) + f(m), \quad \forall\, n \geqslant 1,\ m \geqslant 1.$$

Then $\lim\limits_{n\to\infty} \dfrac{1}{n} f(n)$ *exists, and*

$$\lim_{n\to\infty} \frac{1}{n} f(n) = \inf\left\{\frac{1}{n} f(n) | n \geqslant 1\right\}. \tag{3.32}$$

Proof Let

$$\delta = \inf\left\{\frac{1}{n} f(n) | n \geqslant 1\right\},\ \delta \neq -\infty.$$

For any $\varepsilon > 0$, select a sufficiently large positive integer m so that

$$\frac{1}{m} f(m) < \delta + \frac{\varepsilon}{2}.$$

Let $n = am + b$, where a is an integer, $0 \leqslant b < m$, by semi countable additivity, we have

$$f(n) \leqslant a f(m) + (n - am) f(1).$$

Divide n on both sides, we have

$$\frac{1}{n} f(n) \leqslant \frac{a}{am+b} f(m) + \frac{b}{am+b} f(1).$$

For given b, when m is large enough, we have

$$\frac{bf(1)}{am+b} < \frac{1}{2}\varepsilon.$$

So there is

$$\frac{1}{n} f(n) < \frac{1}{m} f(m) + \frac{1}{2}\varepsilon < \varepsilon + \delta. \tag{3.33}$$

Thus we have

$$\delta \leqslant \underline{\lim_{n\to\infty}} \frac{1}{n} f(n) \leqslant \overline{\lim_{n\to\infty}} \frac{1}{n} f(n) < \delta + \varepsilon.$$

So

$$\lim_{n\to\infty} \frac{1}{n} f(n) = \delta.$$

If $\delta = -\infty$, by (3.33),

$$\overline{\lim_{n\to\infty}}\frac{1}{n}f(n) = -\infty,$$

so we still have

$$\lim_{n\to\infty}\frac{1}{n}f(n) = \delta = -\infty.$$

The Lemma holds.

Lemma 3.10 *Let $\{a_n\}_{n\geqslant 1}$ be a sequence of real numbers, and the limit $\lim_{n\to\infty} a_n = a$, then*

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} a_i = a.$$

Proof

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n} a_i - a\right| &= \left|\frac{1}{n}\sum_{i=1}^{n}(a_i - a)\right| \leqslant \frac{1}{n}\sum_{i=1}^{n}|(a_i - a)| \\ &= \frac{1}{n}\sum_{i=1}^{N}|a_i - a| + \frac{1}{n}\sum_{i=N+1}^{n}|a_i - a| \\ &< \frac{1}{n}\sum_{i=1}^{N}|a_i - a| + \frac{n-N}{n}\varepsilon \\ &< \frac{1}{n}\sum_{i=1}^{N}|a_i - a| + \varepsilon.\end{aligned}$$

When $\varepsilon > 0$ is given, $N$is also given accordingly, the first item of the above formula tends to 0, when $n \to \infty$. So for any $\varepsilon > 0$, when $n > N_0$,

$$\left|\frac{1}{n}\sum_{i=1}^{n} a_i - a\right| < 2\varepsilon.$$

Thus there is

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} a_i = a.$$

The Lemma holds.

With the above preparations, we now give the main results of this section.

Theorem 3.4 *Let X be any source, $\{\xi_i\}_{i\geqslant 1}$ is a set of random variables valued on X. For any positive integer $n \geqslant 1$, let*

$$X_n = (X, \xi_n), \quad n \geqslant 1.$$

Then when X is a stationary source, we have the following two limits that exist and are equal, that is

$$\lim_{n\to\infty}\frac{1}{n}H(X_1X_2\ldots X_n)=\lim_{n\to\infty}H(X_n|X_1X_2\ldots X_{n-1}).$$

We denote the above common limit as $H_\infty(X)$.

Proof Because X is a stationary source, for any $n\geqslant 1$, $m\geqslant 1$, then the joint event probability distribution of random vector $\{\xi_{n+1},\xi_{n+2},\ldots,\xi_{n+m}\}$ on X is equal to the joint probability distribution of random vector $(\xi_1,\xi_2,\ldots,\xi_m)$; therefore, we have

$$H(X_1X_2\cdots X_m)=H(X_{n+1}X_{n+2}\cdots X_{n+m}). \tag{3.34}$$

By Theorem 3.2, then

$$\begin{aligned}H(X_1X_2\cdots X_nX_{n+1}\cdots X_{n+m})&\leqslant H(X_1\cdots X_n)+H(X_{n+1}\cdots X_{n+m})\\&=H(X_1\cdots X_n)+H(X_1\cdots X_m).\end{aligned}$$

Let $f(n)=H(X_1\cdots X_n)$, then $f(n+m)\leqslant f(n)+f(m)$, so $\{f(n)\}_{n\geqslant 1}$ is a non-negative real number sequence with semi countable additive property, by Lemma 3.9, we have

$$\lim_{n\to\infty}\frac{1}{n}H(X_1X_2\cdots X_n)=\inf\left\{\frac{1}{n}H(X_1X_2\cdots X_n)|n\geqslant 1\right\}\geqslant 0.$$

Next, we prove that there is a second limit, that is

$$\lim_{n\to\infty}H(X_n|X_1X_2\cdots X_{n-1})\text{exist.}$$

Firstly, we prove that the sequence is monotonically decreasing, because X is a stationary source, so

$$H(X_1X_2\cdots X_{n-1})=H(X_2X_3\cdots X_n)$$

and

$$H(X_2X_3\cdots X_nX_{n+1})=H(X_1X_2\cdots X_n).$$

So we have

$$H(X_{n+1}|X_2X_3\cdots X_n)=H(X_n|X_1X_2\cdots X_{n-1}). \tag{3.35}$$

By Lemma 3.7,

$$\begin{aligned}H(X_{n+1}|X_1X_2\cdots X_n)&\leqslant H(X_{n+1}|X_2X_3\cdots X_n)\\&=H(X_n|X_1X_2\cdots X_{n-1}).\end{aligned}$$

So $\{H(X_n|X_1X_2\cdots X_{n-1})\}_{n\geqslant 1}$ is a monotonically decreasing sequence and has a lower bound, so $\lim_{n\to\infty} H(X_n|X_1X_2\cdots X_{n-1})$ exist. Further, by the addition formula of Lemma 3.2,

$$\frac{1}{n}H(X_1X_2\cdots X_n)=\frac{1}{n}\sum_{i=1}^{n}H(X_i|X_1X_2\cdots X_{i-1}).$$

By Lemma 3.10, finally we have

$$\lim_{n\to\infty}\frac{1}{n}H(X_1X_2\cdots X_n)=\lim_{n\to\infty}H(X_n|X_1X_2\cdots X_{n-1})=H_\infty(X).$$

We completed the proof of the Theorem.

We call $H_\infty(X)$ the entropy rate of source X. obviously, there is the following corollary.

Corollary 3.2 *(i) For any stationary source X, we have*

$$H_\infty(X)\leqslant H(X_1)\leqslant \log|X|.$$

(ii) If X is a memoryless source, then

$$H_\infty(X)=H(X_1).$$

(iii) If X is a stationary Markov source, then

$$H_\infty(X)=H(X_2|X_1).$$

Proof Since $\{H(X_n|X_1\cdots X_{n-1})\}_{n\geqslant 1}$ is a monotonically decreasing sequence, then

$$H_\infty(X)\leqslant H(X_1).$$

That is, (i) holds. If X is a memoryless source, then

$$\begin{aligned}H(X_1\cdots X_n)&=-\sum_{x_1\in X_1}\cdots\sum_{x_n\in X_n}p(x_1x_2\cdots x_n)\log p(x_1x_2\cdots x_n)\\&=-\sum_{x_1\in X_1}\cdots\sum_{x_n\in X_n}p(x_1\ldots x_n)\{\log p(x_1)+\cdots+\log p(x_n)\}\\&=nH(X_1).\end{aligned}$$

So we have

$$H_\infty(X)=H(X_1).$$

Similarly, we can prove (iii).

Definition 3.10 Let X be a stationary source, we define

$$\delta = \log|X| - H_\infty(X),\ r = 1 - \frac{H_\infty X}{\log|X|}, \tag{3.36}$$

δ is the redundancy of information space X, and r is the relative redundancy of X.

We write

$$H_0 = \log|X|,\ H_n = H(X_n|X_1X_2\cdots X_{n-1}),\ \forall\, n \geq 1.$$

By Theorem 3.4, we have $H_\infty(X) = H_0 \leq H_n$, so

$$H_n \geq (1-r)H_0,\ \forall\, n \geq 1. \tag{3.37}$$

In information theory, redundancy is used to describe the effectiveness of the information carried by the source output symbol. The smaller the redundancy, the higher the effectiveness of the information carried by the source output symbol, and vice versa.

3.4 Markov Chain

Let X, Y, Z be three information spaces, if there is the following conditional probability formula

$$p(xy|z) = p(x|z)p(y|z). \tag{3.38}$$

Say that X and Y are statistically independent under the given condition of Z.

Definition 3.11 If the information space X and Y are statistically independent under condition Z, X, Y, Z is called a Markov chain, denote as $X \to Z \to Y$.

Theorem 3.5 *$X \to Z \to Y$ is a Markov chain if and only if the probability of occurrence of the joint event xzy is*

$$p(xzy) = p(x)p(z|x)p(y|z), \tag{3.39}$$

if and only if

$$p(xzy) = p(y)p(z|y)p(x|z). \tag{3.40}$$

Proof If $X \to Z \to Y$ is a Markov chain, then $p(xy|z) = p(x|z)p(y|z)$, thus

$$\begin{aligned} p(xzy) &= p(z)p(xy|z) \\ &= p(z)p(x|z)p(y|z) \\ &= p(x)p(z|x)p(y|z). \end{aligned}$$

Similarly,

$$\begin{aligned} p(xzy) &= p(z)p(y|z)p(x|z) \\ &= p(y)p(z|y)p(x|z). \end{aligned}$$

That is (3.39) and (3.40) holds. Conversely, if (3.39) holds, then

$$\begin{aligned} p(xzy) &= p(x)p(z|x)p(y|z) \\ &= p(z)p(x|z)p(y|z). \end{aligned}$$

On the other hand, the product formula

$$p(xzy) = p(z)p(xy|z).$$

So we have

$$p(xy|z) = p(x|z)p(y|z).$$

That is $X \to Z \to Y$ is a Markov chain. Similarly, if (3.40) holds, then $X \to Z \to Y$ also is a Markov chain. The Theorem holds.

According to the above Theorem, or by Definition 3.11, obviously, if $X \to Z \to Y$ is a Markov chain, then $Y \to Z \to X$ is also a Markov chain.

Definition 3.12 Let U, X, Z, Y be four information spaces, and the probability of joint event $uxzy$ is

$$p(uxzy) = p(u)p(x|u)p(z|x)p(y|z), \tag{3.41}$$

Call U, X, Z, Y a Markov chain, denote as $U \to X \to Z \to Y$.

Theorem 3.6 *If $U \to X \to Z \to Y$ is a Markov chain, then $U \to X \to Z$ and $U \to Z \to Y$ are also Markov chains.*

Proof Assuming that $U \to X \to Z \to Y$ is a Markov chain, then

$$p(uxzy) = p(u)p(x|u)p(z|x)p(y|z),$$

Both sides sum $y \in Y$ at the same time, and notice that $\sum_{y\in Y} p(y|z) = 1$, then

$$p(uxz) = p(u)p(x|u)p(z|x).$$

By Theorem 3.5, $U \to X \to Z$ is a Markov chain. The left side of the above formula can be expressed as

$$p(uxz) = p(ux)p(z|ux).$$

So we have

$$p(z|ux) = p(z|x).$$

Because $U \to X \to Z \to Y$ is a Markov chain, then

$$\begin{aligned} p(uxzy) &= p(u)p(x|u)p(z|x)p(y|z) \\ &= p(ux)p(z|ux)p(y|z) \\ &= p(uxz)p(y|z). \end{aligned}$$

Both sides sum $x \in X$ at the same time, then we have

$$\begin{aligned} p(uzy) &= p(uz)p(y|z) \\ &= p(u)p(z|u)p(y|z). \end{aligned}$$

Thus $U \to Z \to Y$ is also a Markov chain. The Theorem holds.

In the previous section, we defined the mutual information $I(X, Y)$ of two information spaces X and Y as

$$I(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(xy) \log \frac{p(xy)}{p(x)p(y)}.$$

Now we define the mutual information $I(X, Y|Z)$ of X and Y under condition Z as

$$I(X, Y|Z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(xyz) \log \frac{p(xy|z)}{p(x|z)p(y|z)}. \tag{3.42}$$

By definition, we have

$$I(X, Y|Z) = I(Y, X|Z). \tag{3.43}$$

$I(X, Y|Z)$ is called the conditional mutual information of X and Y.

For conditional mutual information, we first prove the following formula.

Theorem 3.7 *Let X, Y, Z be three information spaces, then*

$$I(X, Y|Z) = H(X|Z) - H(X|YZ) \tag{3.44}$$

and

$$I(X, Y|Z) = H(Y|Z) - H(Y|XZ). \tag{3.45}$$

Proof We only prove (3.44), the same is true for equation (3.45). Because

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

So (3.44) holds.

Corollary 3.3 *We have $I(X, Y|Z) \geq 0$, if and only if $X \to Z \to Y$ is a Markov chain $I(X, Y|Z) = 0$.*

Proof By Theorem 3.7,

$$I(X, Y|Z) = H(X|Z) - H(X|YZ) \geq 0.$$

If $X \to Z \to Y$ is a Markov chain, by (3.42),

$$\log \frac{p(xy|z)}{p(x|z)p(y|z)} = \log 1 = 0,$$

that is $I(X, Y|Z) = 0$. Vice versa.

Conditional mutual information can be used to establish the addition formula of mutual information.

Corollary 3.4 (Addition formula of mutual information) *If $X_1, X_2, \ldots, X_n, Y$ are information spaces, then*

$$I(X_1X_2 \cdots X_n, Y) = \sum_{i=1}^{n} I(X_i, Y|X_{i-1} \cdots X_1). \tag{3.46}$$

Specially, when $n = 2$, we have

$$I(X_1X_2, Y) = I(X_1, Y) + I(X_2, Y|X_1). \tag{3.47}$$

Proof By Lemma 3.4, we have

$$\begin{aligned} I(X_1X_2 \cdots X_n, Y) &= H(X_1X_2 \cdots X_n) - H(X_1X_2 \cdots X_n|Y) \\ &= \sum_{i=1}^{n} H(X_i|X_{i-1} \cdots X_1) - \sum_{i=1}^{n} H(X_i|X_{i-1} \cdots X_1Y). \end{aligned}$$

Again by the chain rule of conditional entropy to get

$$I(X_1X_2 \cdots X_n, Y) = \sum_{i=1}^{n} I(X_i, Y|X_1X_2 \cdots X_{i-1}).$$

Therefore, the corollary holds.

Finally, we use Markov chain to prove the inequality of mutual information.

Theorem 3.8 *Suppose $X \to Z \to Y$ is a Markov chain, then we have*

$$I(X, Y) \leq I(X, Z) \tag{3.48}$$

and

$$I(X, Y) \leq I(Y, Z). \tag{3.49}$$

Proof We only prove (3.48), the same is true for equation (3.49). From equation (3.47) and corollary 3.3:

$$I(YZ, X) = I(Y, X) + I(X, Z|Y).$$

Thus we have

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

In the last step, we use the Markov chain condition, thus $I(X, Y|Z) = 0$. The Theorem holds.

Theorem 3.9 (Data processing inequality) *Suppose* $U \to X \to Y \to V$ *is a Markov chain, then we have*

$$I(U, V) \leq I(X, Y).$$

Proof According to the conditions, $U \to X \to Y$ and $U \to Y \to V$ is a Markov chain, respectively, by Theorem 3.8,

$$I(U, Y) \leq I(X, Y)$$

and

$$I(U, V) \leq I(U, Y).$$

Thus

$$I(U, V) \leq I(X, Y).$$

The Theorem holds.

3.5 Source Coding Theorem

The information coding theory is usually divided into two parts: channel coding and source coding. The so-called channel coding is to ensure the success rate of decoding by increasing the length of codewords. Channel coding, also known as error correction code, is discussed in detail in Chap. 2. Source coding is to compress the data with redundant information to improve the success rate of decoding and recovery after information or data is stored. Another important result of Shannon's theory is that there are so-called good codes in source coding, which is characterized

by fewer codewords as much as possible. To improve the storage space efficiency, and the error of decoding and restoration can be arbitrarily small. Source coding is also called typical code. Shannon first proved the asymptotic bisection property of 'block code' for memoryless source, and drew the statistical characteristics of typical code from now on. At Shannon's suggestion, McMillan (1953) and Breiman (1957) also proved a similar asymptotic bisection property for stationary ergodic sources. This is the very famous Shannon–McMillan–Breiman theorem in source coding, which constitutes the core content of modern typical code theory. The main purpose of this section is to strictly prove the asymptotic bisection of memoryless sources, so as to derive the source coding theorem for data compression (see Theorem 3.10). For the more general Shannon–McMillan–Breiman theorem, Chap. 2 of Ye Zhongxing's fundamentals of information theory (see Zhongxing, 2003 in reference 3) gives a proof under the condition of stationary ergodic Markov source, interested readers can refer to it or refer to more original documents (see McMillan, 1953; Moy, 1961; Shannon, 1959 in reference 3).

Firstly, let $X = (X, \xi)$ be an information space, and the entropy $H(X)$ of X essentially depends only on the probability function $p(x)(x \in X)$ of random variable ξ. We can define the random variable taking value on X according to $p(x)$.

$$\eta_1 = p(X), \ \eta_2 = \log p(X). \tag{3.50}$$

The probability function is

$$P\{\eta_1 \text{ value } x\} = P\{\eta_2 \text{ value } x\} = p(x). \tag{3.51}$$

It is easy to see the expected value of η_2

$$\begin{aligned} -E(\eta_2) &= -E(\log p(X)) \\ &= -\sum_{x \in X} p(x) \log p(x) = H(X). \end{aligned} \tag{3.52}$$

Therefore, we can regard the entropy $H(X)$ of X as the mathematical expectation of random variable $\log \frac{1}{p(X)}$.

Lemma 3.11 *Let X be a memoryless source, $p(X^n)$ and $\log p(X^n)$ be two random variables whose values are on the power space X^n, then $-\frac{1}{n} \log p(X^n)$ converges to $H(X)$ according to probability, that is*

$$-\frac{1}{n} \log p(X^n) \xrightarrow{P} H(X).$$

Proof Since X is a memoryless source, $\{\xi_i\}_{i \geq 1}$ is a group of independent and identically distributed random variables, $X_i = (X, \xi_i)(i \geq 1)$, $X^n = X_1 X_2 \cdots X_n (n \geq 1)$ is a power space, then there is

$$\begin{cases} p(X^n) = p(X_1)p(X_2)\cdots p(X_n) \\ \log p(X^n) = \sum_{i=1}^{n} \log p(X_i). \end{cases}$$

Because $\{\xi_i\}_{i\geq 1}$ is independent and identically distributed, $\{p(X^n)\}$ and $\{\log p(X^n\}$ is also a group of independent and identically distributed random variables. According to Chebyshev's law of large numbers (see Theorem 1.3 of Chap. 1),

$$-\frac{1}{n}\log p(X^n) = \frac{1}{n}\sum_{i=1}^{n} \log \frac{1}{p(X_i)}$$

converges to the common expected value $H(X)$, that is

$$E\left(\log \frac{1}{p(X_i)}\right) = E\left(\log \frac{1}{p(X)}\right) = H(X).$$

For any $\varepsilon > 0$, for any codeword $x = x_1 x_2 \cdots x_n \in X^n$, there is

$$P\{|-\frac{1}{n}\log p(X^n) - H(X)| < \varepsilon\} > 1 - \varepsilon. \tag{3.53}$$

The proof is completed.

Definition 3.13 Let X be a memoryless source, power space X^n, also known as block code,

$$X^n = \{x = x_1 \cdots x_n | x_i \in X,\ 1 \leq i \leq n\}, n \geq 1. \tag{3.54}$$

For any given $\varepsilon > 0$, $n \geq 1$, we define a typical code or a typical sequence $W_\varepsilon^{(n)}$ in the power space X^n as

$$W_\varepsilon^{(n)} = \{x = x_1 \cdots x_n \mid |-\frac{1}{n}\log p(x) - H(X)| < \varepsilon\}. \tag{3.55}$$

Because the definition, and $\varepsilon > 0$, $n \geq 1$, we have

$$W_\varepsilon^{(n)} \subset X^n,\ |X^n| = |X|^n. \tag{3.56}$$

Lemma 3.12 (Progressive bisection) $|W_\varepsilon^{(n)}|$ *represents the number of codewords in typical code* $W_\varepsilon^{(n)}$*, then for any* $\varepsilon > 0$*, in binary channels, we have*

$$(1-\varepsilon)2^{n(H(X)-\varepsilon)} \leq |W_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}. \tag{3.57}$$

Proof By Lemma 3.11 and (3.53), then for any $x \in X^n$, we have

$$P\{|-\frac{1}{n}\log p(x) - H(X)| < \varepsilon\} > 1 - \varepsilon.$$

In other words, for all codewords $x = x_1 x_2 \cdots x_n \in W_\varepsilon^{(n)}$, we have

$$H(X) - \varepsilon < -\frac{1}{n} \log p(x) < H(X) + \varepsilon.$$

Equivalent in binary channel,

$$2^{-n(H(X)+\varepsilon)} \leq p(x) \leq 2^{-n(H(X)-\varepsilon)}, \tag{3.58}$$

Denote the probability of occurrence of $W_\varepsilon^{(n)}$ as $P\{W_\varepsilon^{(n)}\}$, then

$$P\{W_\varepsilon^{(n)}\} = P\{x \in X^n : x \in W_\varepsilon^{(n)}\} > 1 - \varepsilon.$$

On the other hand,

$$P\{W_\varepsilon^{(n)}\} = \sum_{x \in W_\varepsilon^{(n)}} p(x),$$

by (3.58),

$$|W_\varepsilon^{(n)}| \cdot 2^{-n(H(X)+\varepsilon)} \leq P\{W_\varepsilon^{(n)}\} \leq 1.$$

So

$$|W_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}.$$

Again by (3.58), there is

$$|W_\varepsilon^{(n)}| \cdot 2^{-n(H(X)-\varepsilon)} \geq P\{W_\varepsilon^{(n)}\} > 1 - \varepsilon.$$

So we have

$$|W_\varepsilon^{(n)}| > (1 - \varepsilon) 2^{n(H(X)-\varepsilon)}.$$

Combined with the above inequalities on both sides, we have

$$(1 - \varepsilon) 2^{n(H(X)-\varepsilon)} \leq |W_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}.$$

We completed the proof.

By Lemma 3.12, for memoryless source X, the probability distribution $p(x)$ of its power space X^n is approximate to

$$p(x) \sim 2^{-nH(X)}, \ \forall\, x \in X^n.$$

The number of codewords $|W_\varepsilon^{(n)}|$ in typical code $W_\varepsilon^{(n)}$ is approximately

$$|W_\varepsilon^{(n)}| \sim 2^{nH(X)}.$$

Further analysis shows that the proportion of typical code $W_\varepsilon^{(n)}$ in block code X^n is very small, which can be summarized as the following Lemma.

Lemma 3.13 *For a sufficiently small $\varepsilon > 0$ given, when X is not an equal probability information space, we have*

$$\lim_{n\to\infty} \frac{|W_\varepsilon^{(n)}|}{|X|^n} = 0.$$

Proof By Lemma 3.12, we have

$$\frac{|W_\varepsilon^{(n)}|}{|X|^n} \le \frac{2^{n(H(X)+\varepsilon)}}{|X|^n}.$$

So

$$\frac{|W_\varepsilon^{(n)}|}{|X|^n} \le 2^{-n(\log|X|-H(X)-\varepsilon)}.$$

By Theorem 3.1, since X is not an equal probability information space, when ε is sufficient, we have

$$H(X) + \varepsilon < \log|X|.$$

Therefore, when n is sufficiently large, the ratio of $\frac{|W_\varepsilon^{(n)}|}{|X|^n}$ can be arbitrarily small. The Lemma 3.13 holds.

Combining Lemmas 3.11, 3.12 and 3.13, we can describe that the typical codes in block codes have the following statistical characteristics.

Corollary 3.5 *Assuming that X is a memoryless source and the typical sequence (or typical code) $W_\varepsilon^{(n)}$ in block code X^n is defined by formula (3.55), then for any $\varepsilon > 0$, $n \ge 1$, we have*

(i) (*Progressive bisection*)

$$(1-\varepsilon)2^{n(H(X)-\varepsilon)} \le |W_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}.$$

(ii) The occurrence probability $P\{W_\varepsilon^{(n)}\}$ of $W_\varepsilon^{(n)}$ is infinitely close to 1, that is

$$P\{W_\varepsilon^{(n)}\} = P\{x \in X^n : x \in W_\varepsilon^{(n)}\} > 1-\varepsilon.$$

(iii) When X is not equal to almost information space, the proportion of $W_\varepsilon^{(n)}$ in block code X^n is any smaller, that is,

$$\lim_{n\to\infty} \frac{|W_\varepsilon^{(n)}|}{|X|^n} = 0.$$

The above description of the statistical characteristics of typical codes is an important theoretical basis for source coding or data compression. Therefore, we find an

effective way to compress the packet code information, so that the rearranged codewords are as few as possible, and the error probability of decoding and recovery is as small as possible. An effective method is to divide the codeword in block code X^n into two parts; the codeword of typical code $W_\varepsilon^{(n)}$ is uniformly numbered from 1 to M. That is, the codeword in $W_\varepsilon^{(n)}$ forms one-to-one correspondence with the following positive integer set I,

$$I = \{1, 2, \ldots, M\},\ M = |W_\varepsilon^{(n)}|.$$

For codewords that do not belong to $W_\varepsilon^{(n)}$, we uniformly number them as 1: Obviously, for $i, i \neq 1, 1 \leq i \leq n$, there is a unique codeword $x^{(i)} \in W_\varepsilon^{(n)}$ in $W_\varepsilon^{(n)}$, so we can accurately restore i to $x^{(i)}$, that is $i \xrightarrow{\text{decode}} x^{(i)}$ is the correct decoding. For $i = 1$, we will not be able to decode correctly, resulting in decoding recovery error. We denote the code rate of the typical code $W_\varepsilon^{(n)}$ as $\frac{1}{n}\log M$, by Lemma 3.12,

$$(1-\varepsilon)2^{n(H(X)-\varepsilon)} \leq M \leq 2^{n(H(X)+\varepsilon)}.$$

Equivalently,

$$\log(1-\varepsilon) + n(H(X)-\varepsilon) \leq \log M \leq n(H(X)+\varepsilon),$$

Therefore, the bit rate of typical code $W_\varepsilon^{(n)}$ is estimated as follows

$$\frac{1}{n}\log(1-\varepsilon) + H(X) - \varepsilon \leq \frac{1}{n}\log M \leq H(X) + \varepsilon, \tag{3.59}$$

when $0 < \varepsilon < 1$ given, we have

$$H(X) - \varepsilon \leq \lim_{n\to\infty} \frac{1}{n}\log M \leq H(X) + \varepsilon.$$

In other words, the code rate is typically close to $H(X)$. Let us look at the decoding error probability P_e after this number, where

$$P_e = P\{x \in X^n : x \notin W_\varepsilon^{(n)}\}.$$

Because

$$P_e + P\{W_\varepsilon^{(n)}\} = 1,$$

According to the statistical characteristics (ii) of the typical code $W_\varepsilon^{(n)}$,

$$P_e = 1 - P\{W_\varepsilon^{(n)}\} < \varepsilon. \tag{3.60}$$

From this, we derive the main result of this section, the so-called source coding theorem.

Theorem 3.10 (*Shannon, 1948*) *Assuming that X is a memoryless source, then*

(i) When the code rate $R = \frac{1}{n}\log M_1 > H(X)$, there is an encoding with the code rate of R, so that when $n \to \infty$, the error probability of decoding recovery is $P_e \to 0$.

(ii) When the code rate $R = \frac{1}{n}\log M_1 < H(X) - \delta, \delta > 0$ and does not change with $n \to \infty$, then any coding with R as the code rate has $\lim\limits_{n\to\infty} P_e = 1$.

Proof The above analysis has given the proof of (i). In fact, if

$$R = \frac{1}{n}\log M_1 > H(X),$$

then when ε is sufficiently small, by (3.59). Typical codes in block code X_n are

$$R > \frac{1}{n}\log |W_\varepsilon^{(n)}|,\ M_1 > |W_\varepsilon^{(n)}|.$$

Therefore, we construct a code $C \subset X^n$, which satisfies

$$W_\varepsilon^{(n)} \subset C, |C| = M_1.$$

Thus, the code rate of C is just equal to R, and the decoding error probability $P_e(C)$ after compression coding satisfies $P_e(C) < \varepsilon$. Because the probability of occurrence of C

$$P\{C\} + P_e(C) = 1.$$

But

$$P\{C\} \geq P\{W_\varepsilon^{(n)}\} > 1 - \varepsilon,$$

(i) holds. To prove (ii), we note that, $\forall\, x \in W_\varepsilon^{(n)}$, then

$$|-\frac{1}{n}\log p(x) - H(X)| < \varepsilon.$$

The above formula contains $\forall\, x \in W_\varepsilon^{(n)}$,

$$p(x) < 2^{-n(H(X)-\varepsilon)}.$$

Thus, the probability of occurrence of $W_\varepsilon^{(n)}$ satisfies

$$P\{W_\varepsilon^{(n)}\} = \sum_{x\in W_\varepsilon^{(n)}} p(x) \leq |W_\varepsilon^{(n)}| \cdot 2^{-n(H(X)-\varepsilon)}. \tag{3.61}$$

If we use R as the bit rate, because

$$R = \frac{1}{n}\log M < H(X) - \delta,$$

then we have

$$|W_\varepsilon^{(n)}| < M = 2^{n(H(X)-\delta)}.$$

By (3.61),

$$P\{W_\varepsilon^{(n)}\} < 2^{-n(\delta-\varepsilon)}, \tag{3.62}$$

when $0 < \varepsilon < \delta$, we have

$$1 - P_e = P\{W_\varepsilon^{(n)}\} < \varepsilon.$$

Thus

$$\lim_{n\to\infty} P_e = 1,$$

Thus the theorem holds.

3.6 Optimal Code Theory

Let X be a source state set, $x = x_1x_2\cdots x_n \in X^n$ be a message sequence, and x be output as a codeword $u = u_1u_2\cdots u_k \in \mathbb{Z}_D^k$ of length k after compression coding, where $D \geq 1$ is a positive integer, $\mathbb{Z}_D$ is the remaining class ring of mod D, $u = u_1u_2\cdots u_k \in \mathbb{Z}_D^k$ is called a D- ary codeword of length k. u is decoded and translated into message x, that is $u \to x$. The purpose of source coding is to find a good coding scheme to make the code rate as small as possible under the requirement of sufficiently small decoding error. Below, we give the strict mathematical definitions of equal length code and variable length code.

Definition 3.14 Let X be a source state set, $\mathbb{Z}_D$ is the remaining class ring of mod D, n, k are positive integers. The mapping $f : X^n \to \mathbb{Z}_D^k$ is called equal length code coding function; $\mathbb{Z}_D^k \xrightarrow{\psi} X^n$ is called the corresponding decoding function. For $\forall\ x = x_1\cdots x_n \in X^n$, $f(x) = u = u_1\cdots u_k \in \mathbb{Z}_D^k$, $u = u_1\cdots u_k$ is called a codeword of length k.

$$C = \{f(x) \in \mathbb{Z}_D^k | x \in X^n\}, \tag{3.63}$$

call

Call C is the code coded by f, and $R = \frac{k}{n}logD$ is the coding rate of f, also known as the code rate of C. C is called equal length code; it is sometimes called a block code with a packet length of k.

By Definition 3.14, the error probability of an equal length code coding scheme (f, ψ) is

$$P_e = P\{\psi(f(x)) \neq x, \quad x \in X^n\}. \tag{3.64}$$

Let us first consider error free coding, that is $P_e = 0$. Obviously, $P_e = 0$ if and only if f is a injection, $\psi = f^{-1}$ is the left inverse mapping of f. select a coding function $f : X^n \to \mathbb{Z}_D^k$ as a injection if and only if $|\mathbb{Z}_D^k| \geq |X^n|$, that is $D^k \geq N^n$, where $N = |X|$, take logarithms on both sides,

$$R = \frac{k}{n} log D \geq log N = \log |X|. \tag{3.65}$$

Therefore, the code rate of error free compression coding f is at least $\log_2 |X|$ bits or $\ln |X|$ naits.

We consider progressive error free coding, that is, for any given $\varepsilon > 0$, required decoding error probability $P_e \leq \varepsilon$. By Theorem 3.10, only the code rate $R \geq H(X)$ is needed. In fact, take X as an information space and encode the n-lengthen message column $x = x_1 x_2 \cdots x_n \in X^n$, if $x \in W_\varepsilon^{(n)}$ is a typical sequence (typical code), x corresponds to a number in $M = |W_\varepsilon^{(n)}|$, if $x \notin W_\varepsilon^{(n)}$, uniformly code x as 1. If the M codewords in $W_\varepsilon^{(n)}$ are represented by D-ary digits, let $D^k = M$ (the insufficient part can be supplemented), and the code rate R is

$$R = \frac{1}{n} \log M = \frac{k}{n} \log D.$$

Since M is approximately $2^{nH(X)}$, R is approximately $H(X)$, that is $R = \frac{1}{n} \log M \sim H(X)$. From the asymptotic bisection, the error probability of such coding is

$$P_e = P\{x = x_1 \cdots x_n \notin W_\varepsilon^{(n)}\} < \varepsilon, \text{ When } n \text{ is sufficiently large.}$$

However, in practical application, n cannot increase infinitely, which requires us to find the best coding scheme when given a finite n, so that the code rate is as close as possible to the theoretical value $H(X)$. However, in application, we find that equal length code is not an efficient coding scheme, while variable length code is more practical. For example,

Example 3.6 Let $X = \{1, 2, 3, 4\}$ be an information space, and the probability distribution of random variable ξ taking value on X is

$$\xi \sim \begin{pmatrix} 1 & 2 & 3 & 4 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{pmatrix}.$$

The entropy $H(X)$ of information space X is

$$H(X) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{8} \log_2 \frac{1}{8} = 1.75\text{bits}.$$

If equal length code is used for coding, the code length is 2, and the code is

Source letter	Codeword
1	00
2	01
3	10
4	11

Then the code rate $R(k = 2, n = 1)$ is

$$R = 2\log_2 2 = 2 > 1.75\text{bits}.$$

Obviously, the use efficiency of equal length codes is not high. If the above codes are replaced with unequal length codes, such as

Source letter	Codeword
1	0
2	10
3	110
4	111

We use $l(x)$ to represent the code length after the source letter x is encoded, then the average code length L required for X encoding is

$$L = \sum_{i=1}^{4} p(x_i)l(x_i) = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = 1.75 \text{ bits} = H(X).$$

It can be seen that using unequal length code to compile X has higher efficiency. This example also explains the following compression coding principle: for characters with high probability of occurrence, a shorter codeword is prepared, and for characters with low probability of occurrence, a longer codeword is prepared to ensure that the average coding length is as small as possible.

Next, we give the mathematical definition of variable length coding. For this purpose, let X^* and $\mathbb{Z}_D^*$ be the set of finite length sequences, respectively. That is $X^* = \bigcup_{1 \le k < \infty} X^k$.

Definition 3.15 (i) $X^n \xrightarrow{f} \mathbb{Z}_D^*$ is called a variable length code function, if any $x \in X^n$, $f(x) \in \mathbb{Z}_D^*$, When x is different, the code length of $f(x)$ is not necessarily the same. We use $l(x)$ to table the length of $f(x)$, which is called the coding length of x. $C = \{f(x) \in \mathbb{Z}_D^* | x \in X^n\}$ is called variable length codeword set.

(ii) Let $f : X^* \longrightarrow \mathbb{Z}_D^*$ be a amapping, call f is a coding mapping, $f(X^*)$ is called a code.

(iii) $f : X^* \longrightarrow \mathbb{Z}_D^*$ is called a block code mapping, if there is a mapping $g : X \longrightarrow \mathbb{Z}_D^*$, so that for any $x \in X^n (n \ge 1)$, write $x = x_1 x_2 \cdots x_n$, there is $f(x) = g(x_1)g(x_2) \cdots g(x_n)$.

(iv) $f : X^* \longrightarrow \mathbb{Z}_D^*$ is called a uniquely decodable map, if f is a block code mapping and f is a injection.

(v) $f : X^* \longrightarrow \mathbb{Z}_D^*$ is called a real-time code mapping. If f is a block code mapping, and for any $x, y \in X^*$, $f(x)$ and $f(y)$ cannot be prefixes to each other.

Remark 3.1 $a = a_1 a_2 \cdots a_n \in \mathbb{Z}_D^n$, $b = b_1 b_2 \cdots b_m \in \mathbb{Z}_D^m$, call codeword a the prefix of b, if $m \geq n$, and for any $1 \leq i \leq n$, there is $a_i = b_i$.

Lemma 3.14 *Block code mapping $f : X^* \longrightarrow \mathbb{Z}_D^*$ is called a uniquely decodable mapping if and only if for $\forall\, n \geq 1$, $X^n \longrightarrow \mathbb{Z}_D^*$, f is restricted to a injection on X^n.*

Proof The necessity is obvious and the adequacy is proved. That is to prove for $\forall\, x = x_1 x_2 \cdots x_n \in X^n$, $y = y_1 y_2 \cdots y_m \in X^m$, $x \neq y$, there is $f(x) \neq f(y)$. Suppose there is $f(x) = f(y)$, because f is a block code mapping, there is a mapping $g : X \longrightarrow \mathbb{Z}_D^*$, we have

$$f(x) = g(x_1)g(x_2)\cdots g(x_n) = g(y_1)g(y_2)\cdots g(y_m) = f(y).$$

Then

$$\begin{aligned} f(xy) &= g(x_1)g(x_2)\cdots g(x_n)g(y_1)g(y_2)\cdots g(y_m) \\ &= g(y_1)g(y_2)\cdots g(y_m)g(x_1)g(x_2)\cdots g(x_n) \\ &= f(yx). \end{aligned}$$

But $xy \neq yx$, this contradicts the fact that f is restricted to a injection on X^{n+m}.

Lemma 3.15 *A real-time code is uniquely decodable, and vice versa.*

Proof Suppose $f : X^* \longrightarrow \mathbb{Z}_D^*$ as an instant code mapping, and for $x, y \in X^*$, $x \neq y$, there is $f(x) = a_1 a_2 \cdots a_n \in \mathbb{Z}_D^n$, $f(y) = b_1 b_2 \cdots b_m \in \mathbb{Z}_D^m (m \geq n)$. Because $f(x)$ is not a prefix of $f(y)$, it exists $i (1 \leq i \leq n)$, there is $a_i \neq b_i$, thus $f(x) \neq f(y)$, that is f is an injection. In turn, let us take a counter example,

Source letter	Codeword
1	0
2	01
3	011
4	111

where $X = \{1, 2, 3, 4\}$ is the information space and $f : X \to \mathbb{Z}_2^*$ is a variable length code. $f(1)$ is the prefix of $f(2)$, that is, f is not a real-time code map, but obviously f is the only decodeable map. The Lemma holds.

What are the conditions for the code length of a real-time code? The following Kraft inequality gives a satisfactory answer.

Lemma 3.16 *For the uniquely decodable code C value in $\mathbb{Z}_D^*$, $|C| = m$, the code lengths are $l_1, l_2, \ldots, l_m$, then there is the following McMillan–Kraft inequality.*

$$\sum_{i=1}^{m} D^{-l_i} \leq 1. \tag{3.66}$$

On the contrary, if l_i satisfies the above conditions, there is a code length set of real-time code C such that $\{l_1, l_2, \ldots, l_m\}$ is C.

Proof Consider

$$\left(\sum_{i=1}^{m} D^{-l_i})^n = (D^{-l_1} + D^{-l_2} + \cdots + D^{-l_m}\right)^n,$$

the form of each item is $D^{-l_{i_1}-l_{i_2}-\cdots-l_{i_n}} = D^{-k}$, where $l_{i_1} + l_{i_2} + \cdots + l_{i_n} = k$. Suppose $l = \max\{l_1, l_2, \ldots, l_m\}$, then the range of k is from n to nl. Define the number of items where N_k is D^{-k}, then

$$\left(\sum_{i=1}^{m} D^{-l_i}\right)^n = \sum_{k=n}^{nl} N_k D^{-k}.$$

Note that N_k can be regarded as the number of codeword sequences with a total length of k just assembled by n codewords in C, i.e.,

$$N_k = |\{(c_1, c_2, \ldots, c_n) \mid |c_1 c_2 \cdots c_n| = k, c_i \in C\}|.$$

The codeword is still in $\mathbb{Z}_D^*$, and because $f : X^* \longrightarrow \mathbb{Z}_D^*$ is an injection, so $N_k \leq D^k$. then we have

$$\left(\sum_{i=1}^{m} D^{-l_i}\right)^n = \sum_{k=n}^{nl} N_k D^{-k} \leq \sum_{k=n}^{nl} D^k D^{-k} = nl - n + 1 \leq nl.$$

If $x \geq 1$, and when n Is Sufficiently Large, $x^n > nl$. But the above formula holds for all arbitrary n. That is $\sum_{i=1}^{m} D^{-l_i} \leq 1$.

On the contrary, assuming that Kraft inequality exists, that is, there is a given length $l_i (1 \leq i \leq m)$ satisfying formula (3.66), now we need to construct a real-time code with these lengths, and $l_i (1 \leq i \leq m)$ may not be completely different. Definition n_j is the number of codewords with length j, if $l = \max\{l_1, l_2, \ldots, l_m\}$, then

$$\sum_{j=1}^{l} n_j = m.$$

(3.66) equivalent to

$$\sum_{j=1}^{l} n_j D^{-j} \leq 1.$$

Multiply both sides by D^l, then $\sum_{j=1}^{l} n_j D^{l-j} \leq D^l$. There is

$$n_l \leq D^l - n_1 D^{l-1} - n_2 D^{l-2} - \cdots - n_{l-1} D,$$

$$n_{l-1} \leq D^{l-1} - n_1 D^{l-2} - n_2 D^{l-3} - \cdots - n_{l-2} D,$$

$$\cdots$$

$$n_3 \leq D^3 - n_1 D^2 - n_2 D,$$

$$n_2 \leq D^2 - n_1 D,$$

$$n_1 \leq D.$$

Because $n_1 \leq D$, we can choose these n_1 codes arbitrarily, and the remaining $D - n_1$ codes with length 1 can be used as the prefix of other codewords. Therefore, there are $(D - n_1)D$ options for codewords with length of 2. That is $n_2 \leq D^2 - n_1 D$. Similarly, $(D - n_1)D - n_2$ codewords can be used as prefixes of subsequent codewords. Therefore, there are at most $((D - n_1)D - n_2)D$ options for codewords with length of 3. That is $n_3 \leq D^3 - n_1 D^2 - n_2 D$. ..., in this way, we can always construct a real-time code with length $\{l_1, l_2, \ldots, l_m\}$. The Lemma holds!

Let us give an example that is not the only one that can be decoded.

Example 3.7 Let $X = \{1, 2, 3, 4\}$, $\mathbb{Z}_D = \mathbb{F}_2$, the coding scheme is

Source letter	Codeword
1	0=f(1)
2	1=f(2)
3	00=f(3)
4	11=f(4)

Because the encoder inputs and the decoder receives continuous codeword symbols, if the character received by the decoder is 001101, there may be two decoding results, 112212 and 3412. This shows that f^* is not an injection, that is, the code written by f is not uniquely decodable.

By Lemma 3.16, real-time codes or, more generally, uniquely decodable codes must satisfy Kraft inequality. However, the variable length code compiled according to kraft inequality is not the optimal code, because from the perspective of random coding, an optimal code not only requires the accuracy of decoding, but also ensures the efficiency, that is, the average random code length requires the shortest. We summarize the strict mathematical definition of the optimal code as.

Definition 3.16 Let $X = \{x_1, x_2, \ldots, x_m\}$ is an information space, a real-time code $C = \{f(x_1), f(x_2), \ldots, f(x_m)\}$ is called an optimal code if its average random code length

$$L = \sum_{i=1}^{m} p_i l_i \tag{3.67}$$

is the smallest, where $p_i = p(x_i)$ is the occurrence probability of x_i and l_i is the code length of x_i.

For a source state set X, when its statistical characteristics are determined, that is, after X becomes an information space, the probability distribution $\{p(x)|x \in X\}$ is given. Therefore, to find the optimal compression coding scheme for an information space X is to find the optimal solution $\{l_1, l_2, \ldots, l_m\}$ of (3.67) under the condition of kraft inequality. Usually, we use the Lagrange multiplier method to find the optimal solution. Let

$$J = \sum_{i=1}^{m} p_i l_i + \lambda \left(\sum_{i=1}^{m} D^{-l_i} \right),$$

Find the partial derivative of l_i

$$\frac{\partial J}{\partial l_i} = p_i - \lambda D^{-l_i} \log D.$$

Thus

$$D^{-l_i} = \frac{p_i}{\lambda \log D}.$$

By Kraft inequality, that is

$$\sum_{i=1}^{m} D^{-l_i} \leq 1.$$

We get

$$1 \geq \sum_{i=1}^{m} D^{-l_i} = \frac{1}{\lambda \log D} \sum_{i=1}^{m} p_i \Rightarrow \lambda \geq \frac{1}{\log D}.$$

Thus, the optimal code length l_i is

$$l_i \geq -\log_D p_i, \; p_i \geq D^{-l_i}. \tag{3.68}$$

The corresponding optimal average code length L is

$$L = \sum_{i=1}^{m} p_i l_i \geq -\sum_{i=1}^{m} p_i \log_D p_i = H_D(X). \tag{3.69}$$

That is, L is the D-ary information entropy $H_D(X)$ of X. from this, we get the main results of this section.

Theorem 3.11 *The average length L of any D-ary real-time code in an information space X shall satisfies*

$$L \geq H_D(X).$$

The equal sign holds if and only if $p_i = D^{-l_i}$.

Next, we will give another proof of Theorem 3.11. Therefore, we consider that there are two random variables ξ and η on a source state set X, and their probability distributions are

$$p(x) = P\{\xi = x\},\ q(x) = P\{\eta = x\},\ \forall\, x \in X.$$

The relative entropy of random variables is defined as

$$D(p||q) = \sum_{x\in X} p(x) \log \frac{p(x)}{q(x)}. \tag{3.70}$$

Lemma 3.17 *The relative entropy* $D(p||q)$ *of two random variables on* X *satisfies*

$$D(p||q) \geq 0,\ \textit{and}\ D(p||q) = 0 \iff p(x) = q(x), \forall\, x \in X.$$

Proof If the real number $x > 0$ is expanded by the power series of e^x, it can be obtained

$$e^{x-1} = 1 + (x-1) + \frac{1}{2}(x-1)^2 + \cdots.$$

Thus $e^{x-1} \geq x$, there is $\log x \leq x - 1$, by (3.70), then

$$\begin{aligned} -D(p||q) &= \sum_{x\in X} p(x) \log \frac{q(x)}{p(x)} \\ &\leq \sum_{x\in X} p(x)(\frac{q(x)}{p(x)} - 1) = 0. \end{aligned}$$

Thus, there is $D(p||q) \geq 0$, $D(p||q) = 0$'s conclusion is obvious.

Proof (Another proof of theorem 3.11) Investigate $L - H_D(X)$,

$$\begin{aligned} L - H_D(X) &= \sum_{i=1}^{m} p_i l_i - \sum_{i=1}^{m} p_i \log_D \frac{1}{p_i} \\ &= -\sum_{i=1}^{m} p_i \log_D D^{-l_i} + \sum_{i=1}^{m} p_i \log_D p_i. \end{aligned} \tag{3.71}$$

Define

$$r_i = \frac{D^{-l_i}}{c},\ c = \sum_{j=1}^{m} D^{-l_i}.$$

By Kraft inequality, we have

$$c \leq 1, \text{ and } \sum_{i=1}^{m} r_i = 1.$$

Therefore, $\{r_i, 1 \leq i \leq m\}$ is a probability distribution on X, by (3.71),

$$L - H_D(X) = -\sum_{i=1}^{m} p_i \log_D cr_i + \sum_{i=1}^{m} p_i \log_D p_i = \sum_{i=1}^{m} p_i \left(\log_D \frac{p_i}{r_i} + \log_D \frac{1}{c} \right).$$

By Lemma 3.17 and $c \leq 1$, we have

$$L - H_D(X) \geq 0, \text{and } L = H_D(X) \text{ if and only if } c = 1 \text{ and } r_i = p_i,$$

that is

$$p_i = D^{-l_i}, \text{ or } l_i = \log_D \frac{1}{p_i}.$$

We complete the proof of theorem 3.11.

By Theorem 3.11, coding according to probability, then the code length of D-ary optimal code is

$$l_i = \log_D \frac{1}{p_i}, \ 1 \leq i \leq m.$$

But in general, $\log_D \frac{1}{p_i}$ is not an integer, we use $\lceil a \rceil$ to represent the smallest integer not less than the real number a. Take

$$l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil, \ 1 \leq i \leq m. \tag{3.72}$$

Then

$$\sum_{i=1}^{m} D^{-l_i} \leq \sum_{i=1}^{m} D^{-\log_D \frac{1}{p_i}} = \sum_{i=1}^{m} p_i = 1.$$

Then the code length defined by formula (3.72) is $\{l_1, l_2, \ldots, l_m\}$ and satisfies Kraft inequality. From Lemma 3.16, we can define the corresponding real-time code.

Definition 3.17 Let $X = \{x_1, x_2, \ldots, x_m\}$ be an information space, $p_i = p(x_i)$,

$$l(f(x_i)) = l_i = \left\lceil \log_D \frac{1}{p_i} \right\rceil, \ 1 \leq i \leq m.$$

Then the real-time code corresponding to $\{l_1, l_2, \ldots, l_m\}$ is called Shannon code.

Corollary 3.6 *The code length $l(f(x_i))$ of a Shannon code $C = \{f(x_i) | 1 \leq i \leq m\}$ satisfies*

$$l_i = \left\lceil \log_D \frac{1}{p(x_i)} \right\rceil, \log_D \frac{1}{p(x_i)} \leq l_i < \log_D \frac{1}{p(x_i)} + 1 \tag{3.73}$$

and

$$H_D(X) \leq L < H_D(X) + 1.$$

Where L is the average code length of C.

Proof According to the definition of $\lceil a \rceil$, $a \leq \lceil a \rceil < a + 1$, thus

$$\log_D \frac{1}{p(x_i)} \leq l_i < \log_D \frac{1}{p(x_i)} + 1.$$

So both sides multiply by $p(x_i)$ and sum $1 \leq i \leq m$, then there is

$$\sum_{i=1}^{m} p(x_i) \log_D \frac{1}{p(x_i)} \leq \sum_{i=1}^{m} p(x_i) l_i < \sum_{i=1}^{m} p(x_i) \left(\log_D \frac{1}{p(x_i)} + 1 \right).$$

That is

$$H_D(X) \leq L < H_D(X) + 1.$$

The Corollary holds.

3.7 Several Examples of Compression Coding

3.7.1 Morse Codes

In variable length codes, in order to make the average code length as close to the source entropy as possible, the code length should match the occurrence probability of the corresponding coded characters as much as possible. The principle of probabilistic coding is that the characters with high occurrence probability are configured with short codewords, and the characters with low occurrence probability are configured with long codewords, So as to make the average code length as close to the source entropy as possible. This idea has existed long before Shannon theory. For example, Morse code invented in 1838 uses three symbols of dot, dash and space to encode 26 letters in English. It is expressed in binary, one dot is 10, a total of 2 bits, one dash is 1110, a total of 4 bits and the space is 000. There are three bits in total. For example, the commonly used English letter E is represented by a dot, while the infrequently used letter Q is represented by two dashes, one dot and one dash, which can make the average length of the codeword of the English text shorter. However, Morse code does not completely match the occurrence probability, so it is not the optimal code, and it is basically not used now. The following table is the coding table of Morse code (Fig. 3.1)

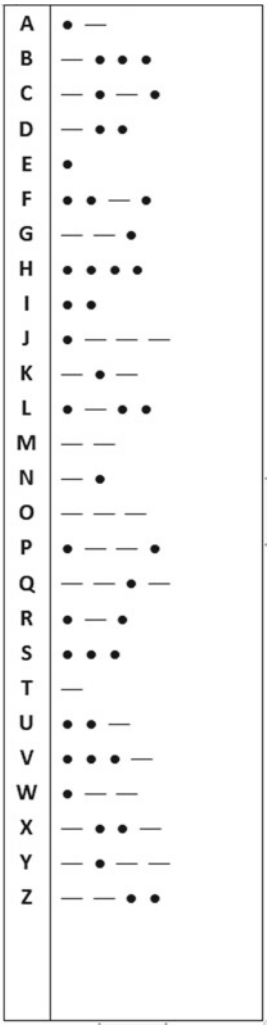

Fig. 3.1 The coding table of Morse code

It is worth noting that Morse code appeared as a kind of password in the early stage, which is widely used in the transmission and storage of sensitive politics (such as military intelligence). The early cryptosystem compilers were also manufactured based on the principle of Morse code, which quickly mechanized the compilation and translation of passwords. In this sense, Morse code has played an important role in promoting the development of cryptography.

3.7.2 Huffman Codes

Shannon, Fano and Huffman have all studied the coding methods of variable length codes, among which Huffman codes have the highest coding efficiency. We focus on the coding methods of Huffman binary and ternary codes.

Let $X = \{x_1, x_2, \ldots, x_m\}$ be the source letter set of m symbols, arrange the m symbols in the order of occurrence probability, take the two letters with the lowest probability to prepare the numbers "0" and "1," respectively, then add their probabilities as a new letter and rearrange them in the order of probability with the source letters without binary numbers. Then take the two letters with the lowest probability to prepare the numbers "0" and "1," respectively, add the probabilities of the two letters as the probability of a new letter, and re queue; continue the above process until the probability of the remaining letters is added to 1. At this time, all source letters correspond to a string of "0" and "1," and we get a variable length code, which is called Huffman code. Taking $X = \{1, 2, 3, 4, 5\}$ as the information space as an example, the corresponding probability distribution is

$$\xi \sim \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0.25 & 0.25 & 0.2 & 0.15 & 0.15 \end{pmatrix}.$$

Binary information entropy $H_2(X)$ and ternary information entropy $H_3(X)$ are

$$\begin{aligned} H_2(X) &= -0.25\log_2 0.25 - 0.25\log_2 0.25 - 0.2\log_2 0.2 \\ &\quad - 0.15\log_2 0.15 - 0.15\log_2 0.15 \\ &= 2.28 \text{ bits}, \\ H_3(X) &= -0.25\log_3 0.25 - 0.25\log_3 0.25 - 0.2\log_3 0.2 \\ &\quad - 0.15\log_3 0.15 - 0.15\log_3 0.15 \\ &= 1.44 \text{ bits}, \end{aligned}$$

respectively. The binary Huffman coding diagram of X is (Fig. 3.2).

The ternary Huffman coding diagram of X is (Fig. 3.3).

In summary, Huffman code has the following characteristics. Assuming that the occurrence probability of the i-th source letter is p_i and the corresponding code length is l_i, then

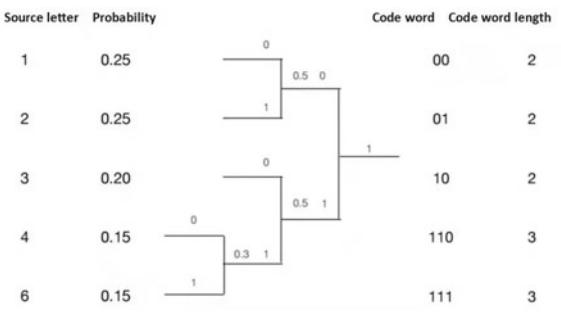

Fig. 3.2 The binary Huffman coding

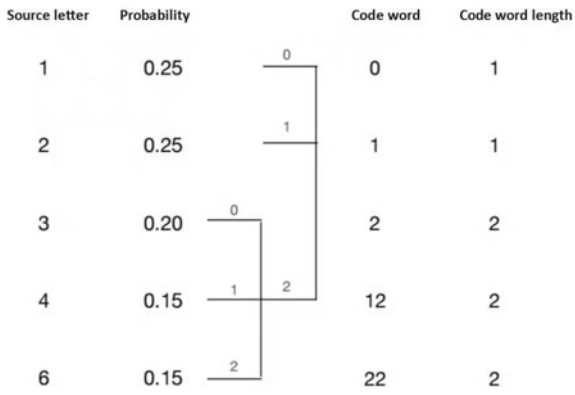

Fig. 3.3 The ternary Huffman coding

(1) If $p_i > p_j$, then $l_i \leq l_j$, that is, the source letter with low probability has a longer codeword;
(2) The longest two codewords have the same code length;
(3) The codeword letters of the two longest codewords are only different from the last letter, and the front ones are the same;
(4) In real-time codes, the average code length of Huffman code is the smallest. In this sense, Huffman code is the optimal code.

Huffman code has been applied in practice, which is mainly used in the compression standard of fax image. However, in the actual data compression, the statistical characteristics of some sources change before and after. In order to make the statistical characteristics based on the coding adapt to the changes of the actual statistical characteristics of the source, an adaptive coding technology has been developed. In each step of coding, the coding of a new message is based on the statistical characteristics of previous messages. For example, R. G. Gallager first proposed the step-by-step updating technology of Huffman code in 1978, and D.E. Knuth made this technology a practical algorithm in 1985. Adaptive Huffman coding technology requires complex data structure and continuous updating of codeword set according to the statistical characteristics of source, We would not go into details here.

3.7.3 Shannon–Fano Codes

Shannon–Fano code is an arithmetic code. Let X be an information space. It can be inferred from Corollary 3.6 in the previous section that the code length of Shannon code on X is

$$l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil, \forall\, x \in X.$$

Here, we introduce a constructive coding method using cumulative distribution function to allocate codewords, commonly known as Shannon–Fano coding method. Without losing generality, let each letter x in X, there is $p(x) > 0$, and define the cumulative distribution function $F(x)$ and the modified distribution function $\bar{F}(x)$ as

$$F(x) = \sum_{a \le x} p(a), \ \bar{F}(x) = \sum_{a < x} p(a) + \frac{1}{2} p(x), \tag{3.74}$$

where $X = \{1, 2, \ldots, m\}$ is a given information space. Without losing generality, let $p(1) \le p(2) \le \cdots \le p(m)$.

As can be seen from the definition, if $x \in X$, then $p(x) = F(x) - F(x-1)$, specially, if $x, y \in X$, then we have

$$\bar{F}(x) \ne \bar{F}(y).$$

So when we know $\bar{F}(x)$, we can find the corresponding x. The basic idea of Shannon–Fano arithmetic code is to use $\bar{F}(x)$ to encode x. Because $\bar{F}(x)$ is a real number, its binary decimal represents the first $l(x)$ bits, denote as $\{\bar{F}(x)\}_{l(x)}$, there is

$$\bar{F}(x) - \{\bar{F}(x)\}_{l(x)} < 2^{-l(x)}. \tag{3.75}$$

Take $l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil + 1$, then we have

$$\frac{1}{2^{l(x)}} = \frac{1}{2 \cdot 2^{\left\lceil \log \frac{1}{p(x)} \right\rceil}} < \frac{p(x)}{2} = \bar{F}(x) - F(x-1), \tag{3.76}$$

Now let the binary decimal of $\bar{F}(x)$ be expressed as

$$\bar{F}(x) = 0.a_1 a_2 \cdots a_{l(x)} a_{l(x)+1} \cdots, \ \forall\, a_i \in \mathbb{F}_2.$$

Then Shannon–Fano code is

$$f(x) = a_1 a_2 \cdots a_{l(x)}, \text{ that is } x \xrightarrow{\textbf{encode}} a_1 a_2 \cdots a_{l(x)} \in \mathbb{F}_2^{l(x)}. \tag{3.77}$$

Lemma 3.18 *The binary Shannon Fano code is a real-time code, and its average length L is at most two bits different from the theoretical optimal value $H(X)$.*

Proof By (3.76),

$$2^{-l(x)} < \frac{1}{2} p(x) = \bar{F}(x) - F(x-1).$$

Let the binary decimal of $\bar{F}(x)$ be expressed as

$$\bar{F}(x) = 0.a_1 a_2 \cdots a_{l(x)} \cdots, \ \forall\, a_i \in \mathbb{F}_2.$$

We use $[A, B]$ to represent a closed interval on the real axis, so

$$\bar{F}(x) \in [0.a_1a_2\cdots a_{l(x)}, 0.a_1a_2\cdots a_{l(x)} + \frac{1}{2^{l(x)}}].$$

If $y \in X$, $x \neq y$, and $f(x)$ is the prefix of $f(y)$, then we have

$$\bar{F}(y) \in [0.a_1a_2\cdots a_{l(x)}, 0.a_1a_2\cdots a_{l(x)} + \frac{1}{2^{l(x)}}].$$

But

$$\bar{F}(y) - \bar{F}(x) \geq \frac{1}{2}p(y) \geq \frac{1}{2}p(x) > \frac{1}{2^{l(x)}},$$

This is contrary to the fact that $\bar{F}(x)$ and $\bar{F}(y)$ are in the same interval. Therefore, we have f as real-time code, that is, Shannon–Fano code is real-time code. Considering its average code length L,

$$L = \sum_{x\in X} p(x)l(x) = \sum_{x\in X} p(x)\left(\left\lceil \log\frac{1}{p(x)}\right\rceil + 1\right) < \sum_{x\in X} p(x)\left(\log\frac{1}{p(x)} + 2\right) = H(X) + 2.$$

We complete the proof of the Lemma.

Let $n \geq 1$, X^n is the power space of the information space, $x = x_1\cdots x_n \in X^n$ is called a message column of length n. In order to improve the coding efficiency, it is often necessary to compress the power space X^n, which is called arithmetic coding. Shannon–Fano code can also be used as arithmetic coding. Its basic method is to find a fast algorithm for calculating joint probability distribution $p(x_1x_2\cdots x_n)$ and cumulative distribution function $F(x)$, and then use Shannon–Fano method to encode $x = x_1\cdots x_n$. We will not introduce the specific details here.

3.8 Channel Coding Theorem

Let X be the input alphabet and Y the output alphabet, and let ξ and η be two random variables with values on X and Y. The probability functions $p(x)$ and $p(y)$ of X and Y and the conditional probability function $p(y|x)$ are

$$p(x) = P\{\xi = x\},\ p(y) = P\{\eta = y\},\ p(y|x) = P\{\eta = y|\xi = x\}\text{respectively.}$$

From the full probability formula,

$$\begin{cases} p(y|x) \geq 0, & \forall\, x \in X, y \in Y. \\ \sum\limits_{y\in Y} p(y|x) = 1, & \forall\, x \in X. \end{cases} \tag{3.78}$$

If X and Y are finite sets, the conditional probability matrix $T = (p(y|x))_{|X|\times|Y|}$ is called the transition probability matrix from X to Y, i.e.,

$$T = \begin{pmatrix} p(y_1|x_1) & p(y_2|x_1) & \dots & p(y_N|x_1) \\ p(y_1|x_2) & p(y_2|x_2) & \dots & p(y_N|x_2) \\ \dots & & & \\ p(y_1|x_M) & p(y_2|x_M) & \dots & p(y_N|x_M) \end{pmatrix}, \tag{3.79}$$

where $|X| = M$, $|Y| = N$. By (3.78), each row of the transition probability matrix T is added to 1.

Definition 3.18 (i) A discrete channel is composed of a finite information space X as the input alphabet, a finite information space Y as the output alphabet, and a transition probability matrix T from X to Y, denote that this discrete channel is $\{X, T, Y\}$. If $X = Y = \mathbb{F}_q$ is q -element finite field, then $\{X, T, Y\}$ is a discrete q-ary channel. In particular, if $q = 2$, then $\{X, T, Y\}$ is called discrete binary channel.

(ii) If $\{X, T, Y\}$ is a discrete q-ary channel and $T = I_q$ is the q-order identity matrix, $\{X, I_q, Y\}$ is called a noise free channel.

(iii) If $\{X, T, Y\}$ is a discrete q-ary channel and $T = T'$ is a q-order symmetric matrix, $\{X, T, Y\}$ is called a symmetric channel.

In discrete channel $\{X, T, Y\}$, codeword spaces X^n and Y^n with length n are defined as

$$X^n = \{x = x_1 \cdots x_n | x_i \in X\}, Y^n = \{y = y_1 \cdots y_n | y_i \in Y\}, n \geq 1.$$

The probabilities of joint events $x = x_1 \cdots x_n$ and $y = y_1 \cdots y_n$ are defined as

$$p(x) = p(x_1 \cdots x_n) = \prod_{i=1}^{n} p(x_i), \quad p(y) = p(y_1 \cdots y_n) = \prod_{i=1}^{n} p(y_i), \tag{3.80}$$

then X and Y become a memoryless source, X^n and Y^n are power spaces, respectively.

Definition 3.19 Discrete channel $\{X, T, Y\}$ is called a memoryless channel if for any positive integer $n \geq 1$, $x = x_1 \cdots x_n \in X^n$, $y = y_1 \cdots y_n \in Y^n$, we have

$$\begin{cases} p(y|x) = \prod_{i=1}^{n} p(y_i|x_i), \\ p(x_i y_i) = p(x_1 y_1), \forall\, i \geq 1. \end{cases} . \tag{3.81}$$

From the joint event probability $p(x_i y_i) = p(x_1 y_1)$ in equation (3.81), then there is

$$p(y_i|x_i) = \frac{p(x_1)}{p(x_i)} p(y_1|x_1). \tag{3.82}$$

The above formula shows that in a memoryless channel, the conditional probability $p(y_i|x_i)$ does not depend on y_i.

Definition 3.19 is the statistical characteristic of a memoryless channel. The following lemma gives a mathematical characterization of a memoryless channel.

Lemma 3.19 *A discrete channel $\{X, T, Y\}$ is a memoryless channel if and only if the product information space XY is a memoryless source, and a power space $(XY)^n = X^n Y^n$.*

Proof If XY is a memoryless source (see Definition 3.9), thn for any $n \geq 1$, and $x = x_1 \cdots x_n \in X^n$, $y = y_1 \cdots y_n \in Y^n$, $xy \in X^n Y^n$, there is

$$p(xy) = p(x_1 \cdots x_n y_1 \cdots y_n) = p(x_1 y_1 \cdots x_n y_n) = \prod_{i=1}^{n} p(x_i y_i).$$

Thus

$$p(x)p(y|x) = p(x) \prod_{i=1}^{n} p(y_i|x_i),$$

so we have

$$p(y|x) = \prod_{i=1}^{n} p(y_i|x_i).$$

$p(x_i y_i) = p(x_1 y_1)$ is given by the definition of memoryless source, so $\{X, T, Y\}$ is a memoryless channel. Conversely, if $\{X, T, Y\}$ is a memoryless channel, by (3.81), there are

$$p(xy) = \prod_{i=1}^{n} p(x_i y_i)$$

and $p(x_i y_i) = p(x_1 y_1)$, then for any $a = a_1 a_2 \cdots a_n \in (XY)^n$, where $a_i = x_i y_i$, we have

$$p(a) = p(x_1 \cdots x_n y_1 \cdots y_n) = p(xy) = \prod_{i=1}^{n} p(x_i y_i) = \prod_{i=1}^{n} p(a_i)$$

and $p(a_i) = p(a_1)$, therefore, XY is a memoryless source, that is, a group of independent and identically distributed random vectors $\xi = (\xi_1, \xi_2, \ldots, \xi_n, \ldots)$ take value on XY, and $(XY)^n = X^n Y^n$ is called power space. The Lemma holds.

The following lemma further characterizes the statistical characteristics of a memoryless channel.

Lemma 3.20 *If $\{X, T, Y\}$ is a discrete memoryless channel, the conditional entropy $H(Y^n|X^n)$ and information $I(X^n, Y^n)$ of information space X^n and Y^n satisfy $\forall\, n \geq 1$,*

$$\begin{cases} H(Y^n|X^n) = nH(Y|X). \\ I(X^n, Y^n) = nI(X, Y). \end{cases} \tag{3.83}$$

Proof Because XY is a memoryless source, we have

$$H(X^nY^n) = H((XY)^n) = nH(XY) = nH(X) + nH(Y|X).$$

On the other hand, by the addition formula of entropy, there is

$$H(X^nY^n) = H(X^n) + H(Y^n|X^n) = nH(X) + H(Y^n|X^n).$$

The combination of the above two formulas has

$$H(Y^n|X^n) = nH(Y|X).$$

According to the definition of mutual information,

$$\begin{aligned} I(X^n, Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &= nH(Y) - nH(Y|X) \\ &= n(H(Y) - H(Y|X)) = nI(X, Y). \end{aligned}$$

The Lemma holds.

Let us define the channel capacity of a discrete channel, this concept plays an important role in channel coding. First, we note that the joint probability distribution $p(xy)$ in the product space XY is uniquely determined by the probability distribution $p(x)$ on X and the probability transformation matrix T, that is $p(xy) = p(x)p(y|x)$; therefore, the mutual information $I(X, Y)$ of X and Y is also uniquely determined by $p(x)$ and T. In fact,

$$\begin{aligned} I(X, Y) &= \sum_{x\in X}\sum_{y\in Y} p(xy)\log\frac{p(xy)}{p(x)p(y)} \\ &= \sum_{x\in X} p(x)\sum_{y\in Y} p(y|x)\log\frac{p(y|x)}{\sum\limits_{x\in X} p(x)p(y|x)}. \end{aligned}$$

Definition 3.20 The channel capacity B of a discrete memoryless channel $\{X, T, Y\}$ is defined as

$$B = \max_{p(x)} I(X, Y), \tag{3.84}$$

where formula (3.84) is the maximum of all probability distributions $p(x)$ on X.

Lemma 3.21 *The channel capacity B of a discrete memoryless channel $\{X, T, Y\}$ is estimated as follows:*

$$0 \leq B \leq \min\{\log|X|,\ \log|Y|\}.$$

Proof The amount of mutual information between the two information spaces is $I(X, Y) \geq 0$ (see Lemma 3.5), so there is $B \geq 0$. By Lemma 3.4,

$$I(X, Y) = H(X) - H(X|Y) \leq H(X) \leq \log|X|$$

and

$$I(X, Y) = H(Y) - H(Y|X) \leq H(Y) \leq \log|Y|,$$

so we have

$$0 \leq B \leq \min\{\log|X|,\ \log|Y|\}.$$

The calculation of information capacity is a problem of solving the conditional extremum of constrained convex function. We will not discuss it in detail here but calculate its channel capacity for two simple channels.

Example 3.8 The channel capacity of noiseless channel $\{X, T, Y\}$ is $B = \log|X|$.

Proof Let $\{X, T, Y\}$ be a noise free channel, then $|X| = |Y|$, and the probability transfer matrix T is the identity matrix, so

$$\begin{aligned} I(X, Y) &= \sum_{x\in X}\sum_{y\in Y} p(xy)\log\frac{p(y|x)}{p(y)} \\ &= \sum_{x\in X} p(x)\sum_{y\in Y} p(y|x)\log\frac{p(y|x)}{p(y)}. \end{aligned}$$

Because $p(y|x) = 0$, if $y \neq x$; $p(y|x) = 1$, if $x = y$. So there is

$$I(X, Y) = \sum_{x\in X} p(x)\log\frac{1}{p(x)} = H(X) \leq \log|X|.$$

Thus

$$B = \max_{p(x)} I(X, Y) = \log|X|.$$

Example 3.9 The channel capacity B of binary symmetric channel $\{X, T, Y\}$ is

$$B = 1 - p\log p - (1-p)\log(1-p) = 1 - H(p),$$

where $p < \frac{1}{2}$, $H(p)$ is the binary entropy function.

Proof In binary symmetric channel $\{X, T, Y\}$, $X = Y = \mathbb{F}_2 = \{0, 1\}$, T is a second-order symmetric matrix

$$T = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}, \ p < 1.$$

Let a be the random variable in the input space $\mathbb{F}_2$ and b be the random variable in the output space $\mathbb{F}_2$, all of which obey the two-point distribution, and then the transfer matrix T of the symmetric binary channel can be represented by the following clearer schematic diagram:

$$\begin{cases} P\{b=1|a=0\} = P\{b=0|a=1\} = p \\ P\{b=0|a=0\} = P\{b=1|a=1\} = 1-p. \end{cases}$$

Calculate mutual information $I(X,Y)$, there is

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

however,

$$\begin{aligned} H(X|Y) &= \sum_{x\in\mathbb{F}_2}\sum_{y\in\mathbb{F}_2} p(xy)\log p(x|y) \\ &= -p\log p - (1-p)\log(1-p) = H(p). \end{aligned}$$

Thus

$$B = \max\{I(X,Y)\} = \max\{H(X) - H(p)\} = 1 - H(p).$$

In order to state and prove the channel coding theorem, we introduce the concept of joint typical sequence. By the Definition 3.13 of Sect. 5 this chapter, if X is a memoryless source, for any small $\varepsilon > 0$ and positive integer $n \geq 1$, in the power space X^n, we define the typical sequence $W_\varepsilon^{(n)}$ as

$$W_\varepsilon^{(n)} = \{x = x_1\cdots x_n \in X^n || -\frac{1}{n}\log p(x) - H(X)| < \varepsilon\}.$$

If $\{X, T, Y\}$ is a memoryless channel, by Lemma 3.19, XY is a memoryless source, in the power space $(XY)^n = X^nY^n$, we define the joint canonical sequence $W_\varepsilon^{(n)}$ as (Fig. 3.4)

$$\begin{aligned} W_\varepsilon^{(n)} = \Big\{ xy \in X^nY^n \Big| &| -\frac{1}{n}\log p(x) - H(X)| < \varepsilon, \ | -\frac{1}{n}\log p(y) - H(Y)| < \varepsilon, \\ &| -\frac{1}{n}\log p(xy) - H(XY)| < \varepsilon \Big\}. \end{aligned} \tag{3.85}$$

Lemma 3.22 (Progressive bisection) *In memoryless channel $\{X, T, Y\}$, the joint typical sequence $W_\varepsilon^{(n)}$ satisfies the following asymptotic bisection properties:*

(i) $\lim\limits_{n\to\infty} P\{xy \in W_\varepsilon^{(n)}\} = 1$;

Fig. 3.4 The transfer matrix

(ii) $(1-\varepsilon)\, 2^{n(H(XY)-\varepsilon)} \leq |W_{\varepsilon}^{(n)}| \leq 2^{n(H(XY)+\varepsilon)}$;
(iii) *If* $x \in X^n$, $y \in Y^n$,*and* $p(xy) = p(x)p(y)$, *then*

$$(1-\varepsilon)\, 2^{-n(I(X,Y)+3\varepsilon)} \leq P\{xy \in W_{\varepsilon}^{(n)}\} \leq 2^{-n(I(X,Y)-3\varepsilon)}.$$

Proof By Lemma 3.13, we have

$$-\frac{1}{n}\log p(X^n) \to H(X), \text{ Convergence according to probability when } n \to \infty;$$

$$-\frac{1}{n}\log p(Y^n) \to H(Y), \text{ Convergence according to probability when } n \to \infty;$$

$$-\frac{1}{n}\log p(X^nY^n) \to H(XY), \text{ Convergence according to probability when } n \to \infty.$$

So when ε is given, as long as n is sufficiently large, there is

$$P_1 = P\left\{|-\frac{1}{n}\log p(x) - H(X)| > \varepsilon\right\} < \frac{1}{3}\varepsilon,$$

$$P_2 = P\left\{|-\frac{1}{n}\log p(y) - H(Y)| > \varepsilon\right\} < \frac{1}{3}\varepsilon,$$

$$P_3 = P\left\{|-\frac{1}{n}\log p(xy) - H(XY)| > \varepsilon\right\} < \frac{1}{3}\varepsilon,$$

where $x \in X^n$, $y \in Y^n$. Thus, it can be obtained

$$P\left\{xy \notin W_{\varepsilon}^{(n)}\right\} \leq P_1 + P_2 + P_3 < \varepsilon.$$

Thus

$$P\left\{xy \in W_{\varepsilon}^{(n)}\right\} > 1-\varepsilon,$$

in other words,

$$\lim_{n\to\infty} P\{xy \in W_{\varepsilon}^{(n)}\} = 1.$$

Property (i) holds. To prove (ii), let $x \in X^n$, $y \in Y^n$, and $xy \in W_{\varepsilon}^{(n)}$, then

$$H(XY) - \varepsilon < -\frac{1}{n} \log p(xy) < H(XY) + \varepsilon.$$

Equivalently,

$$2^{-n(H(XY)+\varepsilon)} < p(xy) < 2^{-n(H(XY)-\varepsilon)}.$$

By total probability formula,

$$1 = \sum_{xy \in X^n Y^n} p(xy) \geq \sum_{xy \in W_\varepsilon^{(n)}} p(xy) \geq |W_\varepsilon^{(n)}|\, 2^{-n(H(XY)+\varepsilon)}.$$

So there is

$$|W_\varepsilon^{(n)}| \leq 2^{n(H(XY)+\varepsilon)}.$$

On the other hand, when n is sufficiently large,

$$\begin{aligned} 1 - \varepsilon < P\{xy \in W_\varepsilon^{(n)}\} &= \sum_{xy \in W_\varepsilon^{(n)}} p(xy) \\ &\leq |W_\varepsilon^{(n)}|\, 2^{-n(H(XY)-\varepsilon)}. \end{aligned}$$

So there is

$$(1-\varepsilon)\, 2^{n(H(XY)-\varepsilon)} \leq |W_\varepsilon^{(n)}| \leq 2^{n(H(XY)+\varepsilon)},$$

property (ii) holds. Now let's prove property (iii). If $p(xy) = p(x)p(y)$, then

$$\begin{aligned} P\{xy \in W_\varepsilon^{(n)}\} &= \sum_{xy \in W_\varepsilon^{(n)}} p(x)p(y) \\ &\leq |W_\varepsilon^{(n)}| 2^{-n(H(X)-\varepsilon)} 2^{-n(H(Y)-\varepsilon)} \\ &\leq 2^{n(H(XY)+\varepsilon-H(X)-H(Y)+2\varepsilon)} \\ &= 2^{-n(I(X,Y)-3\varepsilon)}. \end{aligned}$$

Similarity can prove its lower bound, so we have

$$(1-\varepsilon)\, 2^{-n(I(X,Y)+3\varepsilon)} \leq P\{xy \in W_\varepsilon^{(n)}\} \leq 2^{-n(I(X,Y)-3\varepsilon)}.$$

We have completed the proof of Lemma.

The following lemma has important applications in proving the channel coding theorem. In fact, the conclusion of lemma is valid in general probability space.

Lemma 3.23 *In memoryless channel $\{X, T, Y\}$, if codeword $y \in Y^n$ is uniquely determined by $x \in X^n$, $x' \in X^n$, x' and x are independent, y and x' are also independent.*

Proof If y is uniquely determined by x, then $p(x) = p(y) = p(xy)$, or $p(y|x) = 1$. Therefore, the probability of joint event yxx' is

$$p(yxx') = p(xx') = p(x)p(x') = p(y)p(x').$$

on the other hand,

$$p(yxx') = p(yx').$$

Thus

$$p(yx') = p(y)p(x').$$

The Lemma holds.

In order to define the error probability of channel transmission, we first introduce the workflow of channel coding. After source compression coding, a source message input set is generated,

$$W = \{1, 2, \ldots, M\},\ M \geq 1 \text{ is positive integers.}$$

Injection $f : W \to X^n$ is called coding function, f encodes each input message $w \in W$ as $f(w) \in X^n$. Codeword $x = f(w) \in X^n$ receives codeword $y \in Y^n$ after transmission through channel $\{X, T, Y\}$, we write $x \xrightarrow{T} y$, or $y = T(x)$. Mapping $g : Y^n \to W$ is called decoding function. Therefore, the so-called channel coding is a pair of mapping (f, g). Obviously,

$$C = f(W) = \{f(w)|w \in W\} \subset X^n$$

is a code with length n in codeword space X^n, number of codewords is $|C| = |W| = M$. C is the code of f. The code rate R_C is

$$R_C = \frac{1}{n} \log |C| = \frac{1}{n} \log M.$$

For each input message $w \in W$, if $g(T(f(w))) \neq w$, it is said that the channel transmission is wrong, the transmission error probability λ_w is

$$\lambda_w = P\{g(T(f(w))) \neq w\}, \quad w \in W. \tag{3.86}$$

The transmission error probability of codeword $x = f(w) \in C$ is recorded as $P_e(x)$, obviously, $P_e(x) = \lambda_w$, that is, $P_e(x)$ is the conditional probability

$$\begin{aligned} P_e(x) &= P\{g(T(x)) \neq w | x = f(w)\} \\ &= P\{g(T(f(w))) \neq w\} = \lambda_w. \end{aligned} \tag{3.87}$$

We define the transmission error probability of code $C = f(W) \subset X^n$ as $P_e(C)$,

$$P_e(C) = \frac{1}{M}\sum_{x\in C} P_e(x) = \frac{1}{M}\sum_{w=1}^{M}\lambda_w. \tag{3.88}$$

As before, a code C with length n and number of codewords M is recorded as $C = (n, M)$.

Theorem 3.12 (Shannon's channel coding theorem, 1948) *Let $\{X, T, Y\}$ be a memoryless channel and B be the channel capacity, then*

(i) When $R < B$, there is a column of codes $C_n = (n, 2^{[nR]})$, its transmission error probability $P_e(C_n)$ satisfies

$$\lim_{n\to\infty} P_e(C_n) = 0; \tag{3.89}$$

(ii) Conversely, if the transmission error probability of code $C_n = (n, 2^{[nR]})$ satisfies Eq. (3.89), there is an absolute normal number N_0, and we have the code rate R_{C_n} of C_n satisfies

$$R_{C_n} \le B, \text{ when } n \ge N_0.$$

If $C_n = (n, 2^{[nR]})$, by Lemma 2.27 of Chap. 2,

$$R - \frac{1}{n} < R_{C_n} \le R. \tag{3.90}$$

so (i) of Theorem 3.12 indicates that the code rate is sufficiently close to the channel capacity B, the "good code" with sufficiently small transmission error probability exists. (ii) indicates that the bit rate of the so-called good code with sufficiently small transmission error probability does not exceed the channel capacity. Shannon's proof Theorem 3.12 uses random code technology; this idea of using random method to prove deterministic results is widely used in information theory. At present, it has more and more applications in other fields.

Proof (Proof of theorem 3.12) Firstly, the probability function $p(x_i)$ is arbitrarily selected on the input alphabet X, and the joint probability in power space X^n is defined as

$$p(x) = \prod_{i=1}^{n} p(x_i),\ x = x_1\cdots x_n \in X^n, \tag{3.91}$$

In this way, we get a memoryless source X and power space X^n, which constitute the codeword space of channel coding. Then $M = 2^{[nR]}$ codewords are randomly selected in X^n to obtain a random code $C_n = (n, 2^{[nR]})$. In order to illustrate the randomness of codeword selection, we borrow the source message set $W = \{1, 2, \ldots, M\}$, where $M = 2^{[nR]}$. For every message w, $1 \le w \le M$, the randomly generated codeword is marked as $X^{(n)}(w)$. So we get a random code

$$C_n = \{X^{(n)}(1), X^{(n)}(2), \ldots, X^{(n)}(M)\} \subset X^n.$$

The generation probability $P\{C_n\}$ of C_n is

$$P\{C_n\} = \prod_{w=1}^{M} P\{X^{(n)}(w)\} = \prod_{w=1}^{M} \prod_{i=1}^{n} p(x_i(w)),$$

where $X^{(n)}(w) = x_1(w)x_2(w)\cdots x_n(w) \in X^n$.

We take $A_n = \{C_n\}$ as the set of all random codes C_n, which is called the random code set. The average transmission error probability on random code set A_n is defined as

$$\bar{P}_e(A_n) = \sum_{C_n \in A_n} P\{C_n\} P_e(C_n). \tag{3.92}$$

If you want to prove that for any $\varepsilon > 0$, When n is sufficiently large, $\bar{P}_e(A_n) < \varepsilon$, then there is at least one code $C_n \in A_n$ such that $P_e(C_n) < \varepsilon$, which proves the (i). Therefore, we prove it in two steps.

(1) Principles of constructing random codes and encoding and decoding

We select each message in the source message set $W = \{1, 2, \ldots, M\}$ with equal probability, that is $w \in W$, the selection probability of w is

$$p(w) = \frac{1}{M} = 2^{-[nR]},\ w = 1, 2, \ldots, M.$$

In this way, W becomes an equal probability information space. For each input message w, it is randomly coded as $X^{(n)}(w) \in X^n$, where

$$X^{(n)}(w) = x_1(w)x_2(w)\cdots x_n(w) \in X^n.$$

Codeword $X^{(n)}(w)$ is transmitted through memoryless channel $\{X, T, Y\}$ with conditional probability

$$p(y|X^{(n)}(w)) = \prod_{i=1}^{n} p(y_i|x_i(w))$$

received codeword $y = y_1 y_2 \cdots y_n \in Y^n$. The decoding principle of y is: If $X^{(n)}(w)$ is the only input codeword so that $X^{(n)}(w)y$ is joint typical, that is $X^{(n)}(w)y \in W_\varepsilon^{(n)}$, then decode $g(y) = w$; if there is no such codeword $X^{(n)}(w)$, or there are two or more codewords $X^{(n)}(w)$ and y are joint typical, y cannot be decoded correctly.

(2) Estimating the average error probability of random code set A_n

By (3.92) and (3.88),

$$\begin{aligned}\bar{P}_e(A_n) &= \sum_{C_n \in A_n} P\{C_n\} P_e(C_n) \\ &= \sum_{C_n \in A_n} P\{C_n\} \frac{1}{M} \sum_{x \in C_n} P_e(x) \\ &= \frac{1}{M} \sum_{w=1}^{M} \lambda_w \sum_{C_n \in A_n} P\{C_n\} \\ &= \frac{1}{M} \sum_{w=1}^{M} \lambda_w,\end{aligned} \tag{3.93}$$

where λ_w is given by Eq. (3.86). Because w is input with equal probability, in other words, w is encoded with equal probability. Therefore, the transmission error probability λ_w of w does not depend on w, that is

$$\lambda_1 = \lambda_2 = \cdots = \lambda_M.$$

By (3.93), we have $\bar{P}_e(A_n) = \lambda_1$. To estimate λ_1, we define

$$E_i = \{y \in Y^n | X^n(i) y \in W_\varepsilon^{(n)}\},\ i = 1, 2, \ldots, M, \tag{3.94}$$

If $E_1^c = Y^n \backslash E_1$ is the remainder of E_1, because of the decoding principle,

$$\lambda_1 = P\{E_1^c \cup E_2 \cup \cdots \cup E_M\} \leq P\{E_1^c\} + \sum_{i=2}^{M} P\{E_i\}. \tag{3.95}$$

By property (i) of Lemma 3.22,

$$\lim_{n \to \infty} P\{xy \notin W_\varepsilon^{(n)}\} = 0.$$

So there is

$$\lim_{n \to \infty} P\{X^{(n)}(1) y \notin W_\varepsilon^{(n)}\} = 0.$$

Therefore, when n is sufficiently large,

$$P\{E_1^c\} < \varepsilon.$$

Obviously, codeword $X^{(n)}(1)$ and other codewords $X^{(n)}(i)$, $(i = 2, \ldots, M)$ are independent of each other (see 3.91). By Lemma 3.23, $y = T(X^{(n)}(1))$ and $X^{(n)}(i)$ $(i \neq 1)$ also are independent of each other. Then by the property (iii) of Lemma 3.22,

$$P\{E_i\} = P\{X^{(n)}(i) y \in W_\varepsilon^{(n)}\} \leq 2^{-n(I(X,Y)-3\varepsilon)}\ (i \neq 1).$$

To sum up,

$$\begin{aligned}\bar{P}_e(A_n) = \lambda_1 &\leq \varepsilon + \sum_{i=2}^{M} 2^{-n(I(X,Y)-3\varepsilon)} \\ &\leq \varepsilon + 2^{[nR]} 2^{-n(I(X,Y)-3\varepsilon)} \\ &\leq \varepsilon + 2^{-n(I(X,Y)-R-3\varepsilon)}.\end{aligned}$$

If $R < I(X, Y)$, then $I(X, Y) - R - 3\varepsilon > 0$(when ε is sufficiently small), so when n is large enough, we have $\bar{P}_e(A_n) < 2\varepsilon$. Due to the channel capacity $B = \max\{I(X, Y)\}$, we can choose $p(x)$ to make $B = I(X, Y)$. So when $R < B$, we have $\bar{P}_e(A_n) < 2\varepsilon$, this completes the proof of (i).

To prove (ii), let's look at a special case first. If the error probability of $C = (n, 2^{[nR]})$ is $P_e(C) = 0$, then the bit rate of C is $R_C < B + \frac{1}{n}$, so when n is sufficiently large, there is $R_C \leq B$.

In fact, because $P_e(C) = 0$, decoding function $g : Y^n \to W$ only determines W, there is $H(W|Y^n) = 0$. Because W is equal probability information space, so

$$H(W) = \log |W| = [nR].$$

Using the decomposition of mutual information, there are

$$I(W, Y^n) = H(W) - H(W|Y^n) = H(W) = [nR]. \tag{3.96}$$

on the other hand, $W \to X^n \to Y^n$ forms a Markov chain, by data inequality (see Theorem 3.8)

$$I(W, Y^n) \leq I(X^n, Y^n).$$

By Lemma 3.20,

$$I(W, Y^n) \leq I(X^n, Y^n) = nI(X, Y) \leq nB.$$

By (3.96), there is $[nR] \leq nB$. Because $nR - 1 < [nR] \leq nR$, so $nR < nB + 1$, that is $R < B + \frac{1}{n}$, by (3.90), we have

$$R_C \leq R < B + \frac{1}{n},$$

thus

$$R_C \leq B, \text{ when } n \text{ is sufficiently large.}$$

The above formula shows that when the transmission error probability is 0, as long as n is sufficiently large, there is $R_C \leq B$. Secondly, if the transmission error is allowed, that is, the error probability of C_n is $P_e(C_n) < \varepsilon$, where $C_n = (n, 2^{[nR]})$. Then when n is sufficiently large, we still have $R_{C_n} \leq B$.

In order to prove the above conclusion, we note the error probability of random code C_n is

$$P_e(C_n) = \lambda_w, \tag{3.97}$$

where $w \in W$ is any given message. When w is given, we define a random variable ξ_w with a value on $\{0, 1\}$ as

$$\xi_w = \begin{cases} 1, & \text{if } g(T(f(w))) \neq w; \\ 0, & \text{if } g(T(f(w))) = w. \end{cases}$$

Let $E = (\mathbb{F}_2, \xi_w)$ be a binary information space, by (3.97), then we have

$$P_e(C_n) = P\{\xi_w = 1\}.$$

By Theorem 3.3,

$$\begin{aligned} H(EW|Y^n) &= H(W|Y^n) + H(E|WY^n) \\ &= H(E|Y^n) + H(W|EY^n). \end{aligned} \tag{3.98}$$

Note that E is uniquely determined by Y^n and W, so $H(E|WY^n) = 0$, at the same time, E is a binary information space, $H(E) \leq \log 2 = 1$, there is

$$H(E|Y^n) \leq H(E) \leq 1.$$

On the other hand, the random variable ξ_w is only related to $w \in W$, so

$$H(W|EY^n) = P_e(C_n) \log(|W| - 1) \leq nRP_e(C_n).$$

By (3.98), we have

$$H(W|Y^n) \leq 1 + nRP_e(C_n).$$

Because $f(W) = X^n(W)$ is a function of W, we have the following Fano inequality

$$H(f(W)|Y^n) \leq H(W|Y^n) \leq 1 + nRP_e(C_n).$$

Finally,

$$\begin{aligned} = H(W) &= H(W|Y^n) + I(W, Y^n) \\ &\leq H(W|Y^n) + I(f(W), Y^n) \\ &\leq 1 + nRP_e(C_n) + I(X^n, Y^n) \\ &\leq 1 + nRP_e(C_n) + nB, \end{aligned}$$

because of $nR - 1 < [nR]$, then we have

$$nR < 2 + nRP_e(C_n) + nB.$$

Thus

$$R_{C_n} \leq R < B + \frac{2}{n} + \varepsilon,$$

When n is sufficiently large, we obtain $R_{C_n} \leq B$, which completes the proof of the theorem.

It can be seen from Example 3.9 that the channel capacity $B = 1 - H(p)$ of a binary symmetric channel. Therefore, Theorem 3.12 extends Theorem 2.10 in the previous chapter to a more general memoryless channel; at the same time, it is also proved that the code rate of a good code does not exceed the capacity of the channel.

Exercise 3

1. The joint probability functions of the two information spaces X and Y are as follows:

Y	X	
	0	1
0	$\frac{1}{4}$	$\frac{1}{4}$
1	$\frac{1}{12}$	$\frac{5}{12}$

   Solve $H(X)$, $H(Y)$, $H(XY)$, $H(X|Y)$, $H(Y|X)$, and $I(X, Y)$.
2. Let X_1, X_2, X_3 be three information spaces on $\mathbb{F}_2$, Known $I(X_1, X_2) = 0$, $I(X_1, X_2, X_3) = 1$, prove:

$$H(X_3) = 1, \text{ and } H(X_1X_2X_3) = 2.$$

3. Give an example to illustrate $I(X, Y|Z) \geq I(X, Y)$.
4. Can $I(X, Y|Z) = 0$ be derived from $I(X, Y) = 0$? In turn, can $I(X, Y|Z) = 0$ deduce $I(X, Y) = 0$? Please prove or give examples.
5. Let X, Y, Z be three information spaces, prove:

   (i) $H(XY|Z) \geq H(X|Z)$;
   (ii) $I(XY, Z) \geq I(X, Z)$;
   (iii) $H(XYZ) - H(XY) \leq H(XZ) - H(X)$;
   (iv) $I(X, Z|Y) = I(Z, Y|Z) - I(Z, Y) + I(X, Z)$.

   It also explains under what conditions the equality sign holds.
6. Can $I(X, Y) = 0$ deduce $I(X, Z) = I(X, Z|Y)$?
7. Let the information space be $X = \{0, 1, 2, \ldots\}$ and the value probability $p(n)$ of random variable ξ be

$$p(n) = P\{\xi = n\}, n = 0, 1, \ldots.$$

   Given the mathematical expectation $E\xi = A > 0$ of ξ, find the maximum probability distribution $\{p(n)|n = 0, 1, \ldots\}$ of $H(X)$ and the corresponding maximum information entropy.

8. Let the information space be $X = \{0, 1, 2, \ldots\}$, and take an example of the random variable ξ taken from X, so that $H(X) = \infty$.
9. Let $X_1 = (X, \xi)$, $X_2 = (X, \eta)$ be two information spaces and ξ be a function of η, prove $H(X_1) \leq H(X_2)$, and explain this result.
10. Let $X_1 = (X, \xi)$, $X_2 = (X, \eta)$ be two information spaces and $\eta = f(\xi)$, prove

    (i) $H(X_1) \geq H(X_2)$, give the conditions under which the equal sign holds.
    (ii) $H(X_1|X_2) \geq H(X_2|X_1)$, give the conditions under which the equal sign holds.

References

Bassoli, R., Marques, H., & Rodriguez, J. (2013). Network coding theory, a survey. *IEEE Commun. Surveys Tutor., 15*(4), 1950–1978.

Berger, T. (1971). *Rate distortion theory: a mathematical basis for data compression*. Prentice-Hall.

Blahut, R. E. P. (1965). *Ergodic theory and informtion*. Wiley.

Chung, K. L. (1961). A note on the ergodic theorem of information theory. *Addison. Math Statist., 32*, 612–614.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley.

Csiszár, I., & Körner, J. (1981). *Informaton theory: Coding theorems for discrete memoryless systems*. Academic Press.

EI Gamal, A., & Kim, Y. H. (2011). *Network information theory*. Cambrige University Press

Fragouli, C., Le Boudec, J. Y., & Widmer, J. (2006). Network coding: An instant primer. *ACMSIGCOMM Computer Communication Review, 36*, 63–68.

Gallager, R. G. (1968). *Information theory and reliable communication*. Wiley.

Gray, R. M. (1990). *Entropy and information theory*. Springer.

Guiasu, S. (1977). *Inormation theory with applications*. McGraw-Hill.

Ho, T., & Lun, D. Network coding: An introduction. *Computer Journal*.

Hu, X. H., & Ye, Z. X. (2006). Generalized quantum entropy. *Journal of Mathematical Physics, 47*(2), 1–7.

Ihara, S. (1993). *Information theory for continuous systems*. World Scientific.

Kakihara, Y. (1999). *Abstract methods in information theory*. World Scientific.

McMillan, B. (1953). The basic theorems of information theory. *Annals of Mathematical Statistics, 24*(2), 196–219.

Moy, S. C. (1961). A note on generalizations of Shannon-McMilllan theorem. *Pacific Journal of Mathematics, 11*, 705–714.

Nielsen, M. A., & Chuang, I. L. (2000). *Quantum computation and quantum information*. Cambridge University Press.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Labs Technical Journal, 27*(4), 379–423, 623–656.

Shannon, C. E. (1959). Coding theorem for a discrete source with a fidelity criterion. *IRE National Convention Record, 4*, 142–163.

Shannon, C. E. (1958). Channels with side information at the transmitter. *IBM Journal of Research and Development, 2*(4), 189–193.

Shannon, C. E. (1961). Two-way communication channels. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1*, 611–644.

Thomasian, A. J. (1960). An elementary proof of the AEP of information theory. *Annals of Mathematical Statistics, 31*(2), 452–456.

Wolfowitz, J. (1978). *Coding theorems of information theory* (3rd ed.). Springer-Verlag.

Ye, Z. X., & Berger, T. (1998). *Information measures for discrete random fields*. Science Press.
Yeung, R. W. (2002). *A first course in information theory*. Kluwer Academic.
Qiu, P. (2003). *Information theory and coding*. Higher Education Press. (in Chinese).
Qiu, P., Zhang, C., Yang, S., et al. (2012). *Multi user information theory*. Science Press. (in Chinese).
Ye, Z. (2003). *Fundamentals of information theory*. Higher Education Press. (in Chinese).
Zhang, Z., & Lin, X. (1993). *Information theory and optimal coding*. Shanghai Science and Technology Press. (in Chinese).