信源编码

- 对于信源来说有两个重要问题：
- 1、信源输出信息量的定量度量问题；
- 2、如何有效地表示信源输出问题。

信源概率空间为

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & \cdots & s_q \\ p(s_1) & p(s_2) & \cdots & p(s_q) \end{bmatrix}$$

把信源编码为适合二元信道传输的二元码.二元信道是数字通信中常用的一种信道,它的码符号集为$\{0,1\}$.

解

表 5.1　二元码

信源符号 s_i	$p(s_i)$	码 1	码 2
s_1	$p(s_1)$	00	0
s_2	$p(s_2)$	01	01
s_3	$p(s_3)$	10	001
s_4	$p(s_4)$	11	111

将信源通过一个二元信道传输,就必须把信源符号变换成由 0、1 符号组成的码符号序列.对每个信源符号,可以编成不同的二元码符号序列,就可得到不同的码.

- 信源编码的主要任务是什么？

由于信源符号之间存在分布不均匀和相关性，使得信源存在冗余度，信源编码的主要任务就是减少冗余，提高编码效率。

具体说，就是针对信源输出符号序列的统计特性，寻找一定的方法把信源输出符号序列变换为最短的码字序列。

香农1949年发表“噪声下的通信” ----

为信源编码和信道编码奠定理论基础；

编码定理说明：

（1）必存在一种编码方法，使代码的平均长度可任意接近但不能低于符号熵

（2）达到这目标的途径，就是使概率与码长匹配。

- 信源编码的基本途径有两个:
- 一是使序列中的各个符号尽可能地互相独立，即解除相关性；
- 二是使编码中各个符号出现的概率尽可能地相等，即概率均匀化。

第五章 无失真信源编码

1. 编码的相关概念及分类
2. 定长码及定长码编码方法 编码效率
3. 变长码及变长码编码方法 判别准则
4. 变长编码：香农编码、费诺编码、霍夫曼编码
5. 实用的无失真信源编码方法：
游程编码、算术编码、LZW编码

第一节 信源编码的相关概念

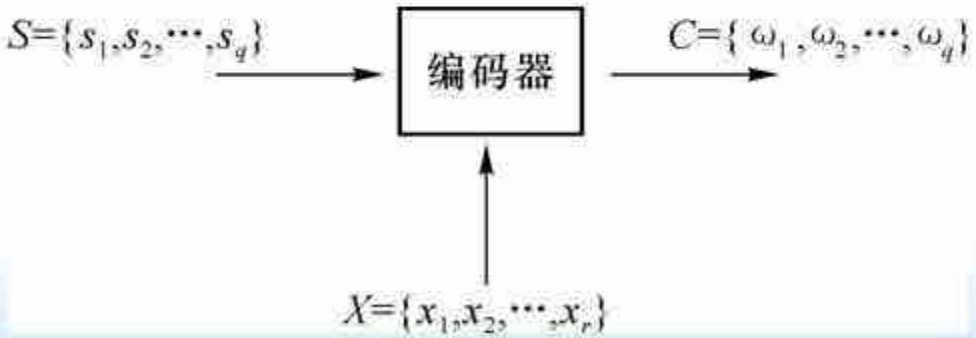
1. 信源编码器及其数学描述
2. 不同分类码的分类
3. 码树图

1、信源编码器

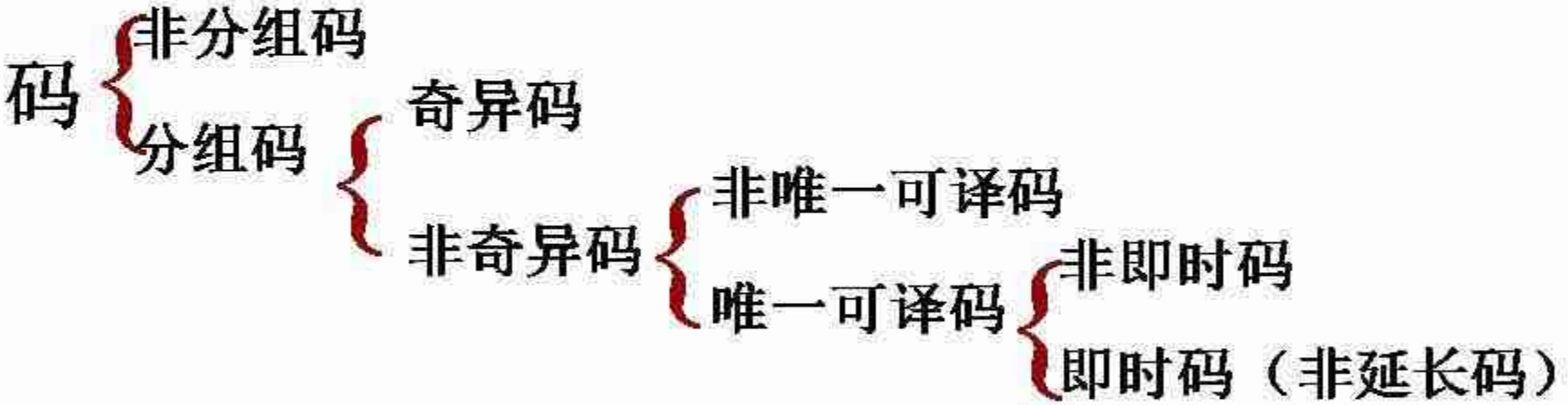
信源输出的符号序列，需要变换成适合信道传输的符号序列，一般称为**码序列**，对信源输出的原始符号按照一定的数学规则进行的这种变换称为**编码**，完成编码功能的器件，称为**编码器**。接收端有一个**译码器**完成相反的功能。

图5-1 信源编码器

- 码就是把信源符号序列变换到码符号序列的一种映射。若要实现无失真编码，那么这种映射必须是一一对应的、可逆的。一般来说，人们总是希望把信源所有的信息毫无保留地传递到接收端，即实现无失真传递，所以首先要对信源实现无失真编码。



2、码的分类



（1）、分组码和非分组码

- 设: 信源消息为符号序列$\mathbf{X_i}$，$X_i=(X_1X_2\cdots X_l\cdots X_L)$，序列中的每个符号取自于符号集A，$X_l\in\{a_1,a_2,\cdots,a_i,\cdots,a_n\}$。而每个符号序列$\mathbf{X_i}$依照固定的码表映射成一个码字$\mathbf{Y_i}$，这样的码称为分组码，有时也叫块码。

  与分组码对应的是非分组码，又称为树码。树码编码器输出的码符号通常与编码器的所有信源符号都有关。

  只有分组码才有对应的码表，而非分组码中则不存在码表。

（2）. 奇异码与非奇异码

- **定义5.2** 若一种分组码中的所有码字都不相同，则称此分组码为非奇异码，否则称为奇异码。

表5.3中，码1是奇异码，码2是非奇异码. 非奇异码是分组码能够正确译码的必要条件，而不是充分条件. 例如传送分组码2时，如果接收端接收到00时，并不能确定发送端的消息是 s_1s_1 还是 s_3.

表5.3 奇异码和非奇异码

信源符号 s_i	码1	码2
s_1	0	0
s_2	11	10
s_3	00	00
s_4	11	01

（3）. 唯一可译码与非唯一可译码

- **定义5.3** 任意有限长的码元序列，如果只能唯一地分割成一个个码字，便称为**唯一可译码**。

- 唯一可译码的物理含义是指不仅要求不同的码字表示不同的信源符号，而且还要求对由信源符号构成的符号序列进行编码时，在接收端仍能正确译码而不发生混淆。唯一可译码首先是非奇异码，且任意有限长的码字序列不会雷同。

（4）. 即时码与非即时码

- **定义5.4** 无需考虑后续的码符号就可以从码符号序列中译出码字，这样的唯一可译码称为**即时码**。

表 5.4　即时码和唯一可译码

信源符号 s_i	码 1	码 2
s_1	1	1
s_2	10	01
s_3	100	001
s_4	1000	0001

同是唯一可译码，其译码方法仍有不同. 如表 5.4 中列出的两组唯一可译码，其译码方法不同. 当传送码 1 时，信道输出端接收到一个码字后不能立即译码，还需等到下一个码字接收到时才能判断是否可以译码. 若传送码 2，则无此限制，接收到一个完整码字后立即可以译码，我们称后一种码为**逗点码**，它是一种即时码，是唯一可译码的一种.

3、码树图

- 即时码可以用树图来构造. 图5.2是一个二元即时码的树图.

树是没有回路的图,所以它也是由节点和弧构成的.树中最顶部的节点称为**根节点**,没有子节点的节点称为**叶子节点**.在构造即时码的树图中,每个节点最多有 r 个子节点,在从此节点到其若干个子节点的弧上分别标柱着 $x_1, x_2, \cdots, x_n$, 这里的 $n \leqslant r$, r 为码符号的个数.将从根节点到叶子节点的各段弧上的码符号顺次连接,就可得到相应的码字.

所有根节点的子节点称为**一阶节点**,所有一阶节点的子节点称为**二阶节点**,依此类推.**n 阶节点**最多有 r^n 个.节点的阶次又称为节点的**深度**.

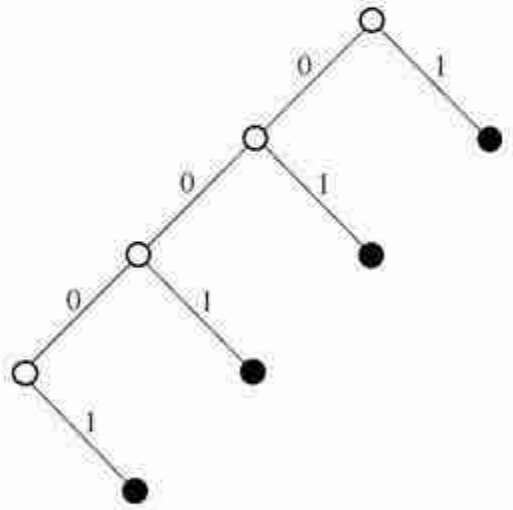
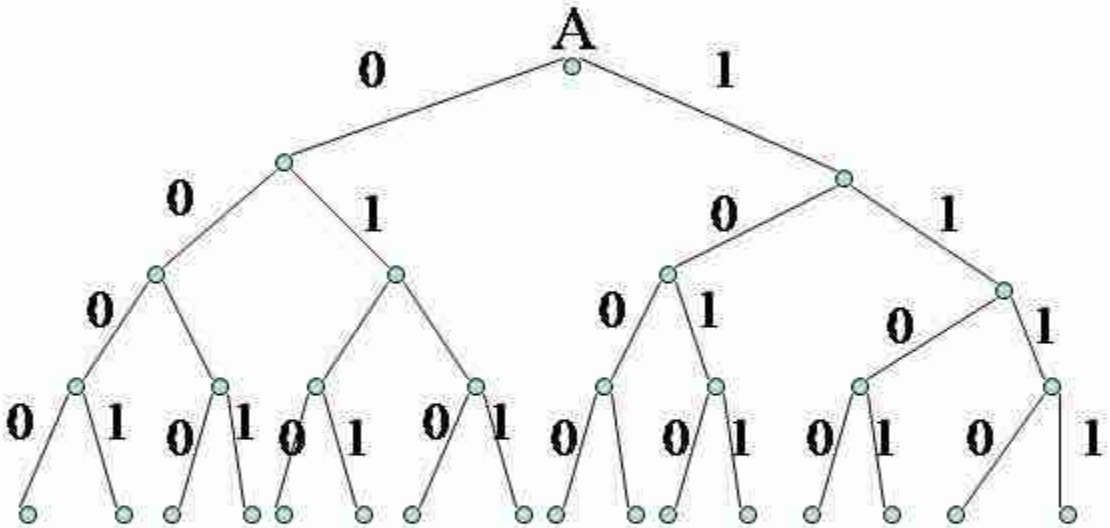
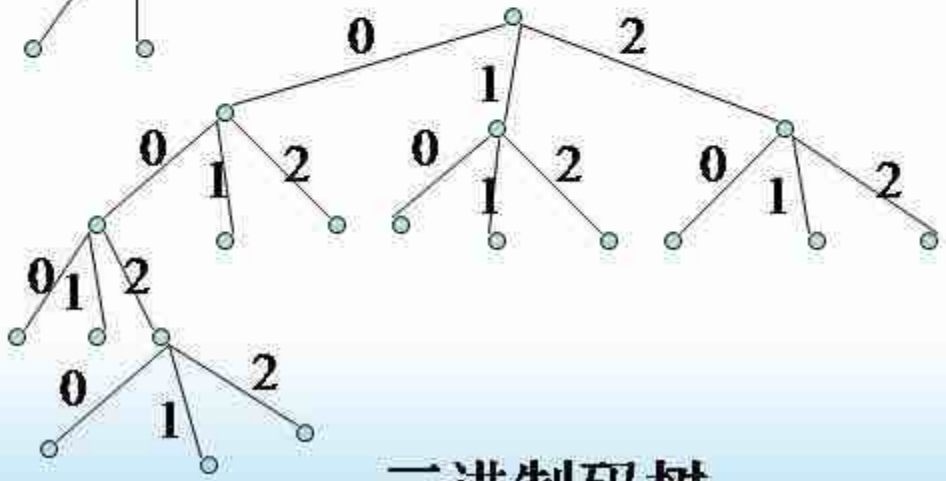


图5.2 二元即时码的树图

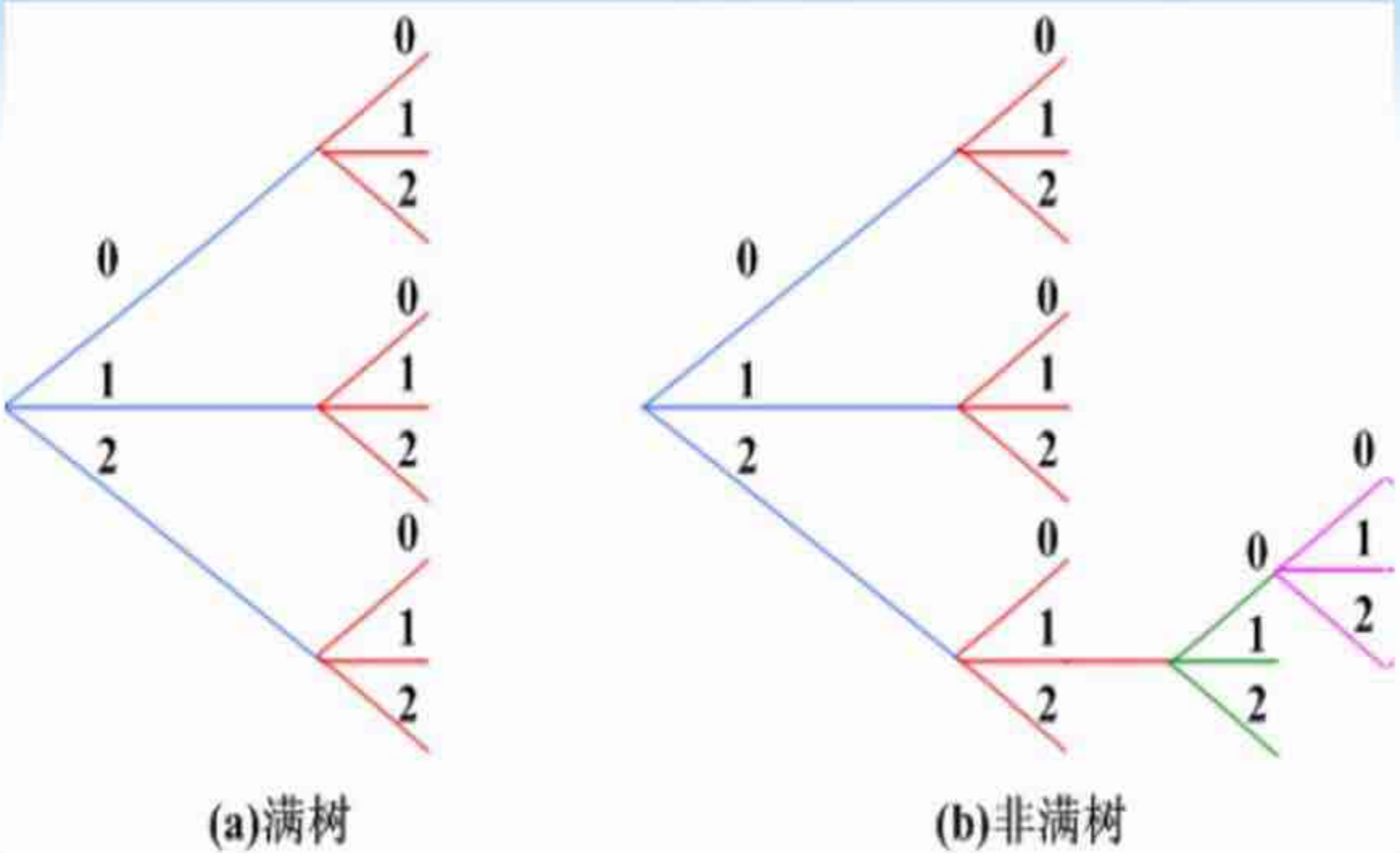
码树图



二进制码树



三进制码树


0
1
2
0
1
2
0
1
2
0
1
2
0
1
2
0
1
2
0
1
2
0
1
2
(a)满树
(b)非满树

即时码的树图还可以用来译码，当接收到一串码符号序列后，首先从树的根节点出发，根据接收到的第一个码符号来选择应走的第一条路径，若沿着所选支路走到中间节点，那就再根据接收到的第二个码符号来选择应走的第二条路径，若又走到中间节点，就再依次继续下去，直到叶子节点为止．走到叶子节点，就可根据所走的枝路立即判断出所接收的码字，同时使系统重新返回根节点，再做下一个接收码字的判断．这样就可以将接收到的一串码符号序列译成对应的信源符号序列．

总结：码的分类

- 码
  - 非分组码
  - 分组码
    - 奇异码
    - 非奇异码
      - 非唯一可译码
      - 唯一可译码
        - 非即时码
        - 即时码（非延长码）

◆奇异码和非奇异码：若信源符号和码字是一一对应的，则该码为非奇异码。反之为奇异码。

◆唯一可译码：任意有限长的码元序列，只能被唯一地分割成一个个的码字，便称为唯一可译码。

◆非即时码和即时码：如果接收端收到一个完整的码字后，不能立即译码，还需等下一个码开始接收后才能判断是否可以译码，这样的码叫做非即时码。

信源符号xi	符号出现概率	码1	码2	码3	码4
x1	1/2	0	0	1	1
x2	1/4	11	10	10	01
x3	1/8	00	00	100	001
x4	1/8	11	01	1000	0001

奇异码与非奇异码　码1是奇异码2是非奇异码

唯一可译码与非唯一可译码

码4是即时码3是非即时码　即时码与非即时码

无失真的编码---唯一可译码

一般来说，若要实现无失真的编码，所编的码必须是唯一可译码，否则，就会因译码带来错误与失真.

表 5.5　定长码

信源符号 s_i	码 1	码 2
s_1	00	00
s_2	01	11
s_3	10	10
s_4	11	11

表 5.5 中码 2 是奇异码. 当接收到码符号 11 后既可译成 s_2 也可译成 s_4，所以不能唯一地译码. 而码 1 是等长非奇异码，它是一个唯一可译码.

对于定长码来说，若定长码是非奇异码，则它的任意有限长 N 次扩展码一定也是非奇异码，因此定长非奇异码一定是唯一可译码.

第二节 、定长码及定长编码定理

1. 唯一可译定长码存在的条件
2. 定长信源编码定理
3. 编码效率

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ 0.40 & 0.18 & 0.10 & 0.10 & 0.07 & 0.06 & 0.05 & 0.04 \end{bmatrix}$$

- 如何编码：给定的信源，如何确定定长编码的码长度？
- 编码效率如何：如何计算编码效率，有效性指标是否满足？
- 如果是N次扩展信源，如何确定定长编码的码长度？其编码效率如何？
- 对于定长编码，如果要提高编码效率，有什么方法？

1、唯一可译定长码存在的条件

- 若对一个有q个信源符号的信源S进行定长编码，那么信源S存在唯一可译定长码的条件是
  - $$q \leq r^l \tag{5.1}$$
  - 其中，r是码符号集中的码元数，l是定长码的码长。

对于定长唯一可译码，平均每个原始信源符号至少需要 $\log_r q$ 个码符号来表示.

例如英文电报有 32 个符号(26 个英文字母加 6 个标点符号)，即 $q=32$，若利用二元码，则 $r=2$，若对信源 S 的每个符号 s_i，$i=1,2,\cdots,32$ 进行二元编码，则 $l \geqslant \log_2 q = \log_2 32 = 5$，也就是说，每个英文电报符号至少要用 5 位二元码符号进行编码才能得到唯一可译码.

扩展信源唯一可译定长码存在的条件

- 如果对信源S的N次扩展信源S^N进行定长编码，若要编得的定长码是唯一可译码，则必须满足
- $$q^N \le r^l \tag{5.2}$$
  - 其中，q是信源S的符号个数，q^N是信源S的N次扩展信源S^N的符号个数，r是码符号集X的码符号数。

对式(5.2)两边取对数得 $N\log_2 q \leqslant l\log_2 r$，则

$$\frac{l}{N} \geqslant \frac{\log_2 q}{\log_2 r}$$

其中，$\frac{l}{N}$表示S^N中平均每个原始信源符号所需要的码符号个数.

2、定长编码定理

- 定长编码的信息传输效率是很低的。提高信息传输效率的方法有：
  - 方法1 考虑符号之间的依赖关系，对信源S的扩展信源进行编码。
  - 方法2对于概率等于0或非常小的符号序列不予编码。这样可能会造成一定的误差，但是当N足够大时，这种误差概率可以任意小，即可做到几乎无失真编码。
  -
- **定理5.2** 离散无记忆信源的熵为$H(S)$，若对信源长为N的序列进行定长编码，码符号集X中有r个码符号，码长为l，则对于任意$\varepsilon > 0$ ，只要满足

$$\frac{l}{N} \geq \frac{H(S)+\varepsilon}{\log r}$$

- 则当N足够大时，可实现几乎无失真编码，即译码错误概率 δ 任意小；
- 反之，如果

$$\frac{l}{N} \leq \frac{H(S)-2\varepsilon}{\log r}$$

- 则不可能实现几乎无失真编码，即当N足够大时，译码错误概率为1。

定理5.2的条件又可写成 $l\log_2 r > NH(S)$，这个式子表明只要 l 长码符号序列所能携带的最大信息量大于 N 长信源序列所携带的信息量，就可以实现无失真编码，当然条件是 N 足够大.

- **问题是：当达到了这样的N之后，满足了不失真的条件，但编码效率如何？**

3、编码效率

定义 5.6 设熵为 $H(S)$的离散无记忆信源，若对信源的长为 N 的符号序列进行定长编码，码符号集中码符号个数为 r，设码字长为 l，定义 $R'=\frac{l}{N}\log_2 r$ 比特/信源符号为**编码速率**，它表示平均每个信源符号编码后能载荷的最大信息量.

这时，定理 5.2 的条件可表述为 $R'\geqslant H(S)+\varepsilon$，即编码速率大于信源的熵才能实现几乎无失真编码. 为了衡量编码效果，引进编码效率.

定义 5.7 定义 $\eta=\frac{H(S)}{R'}=\frac{H(S)}{\frac{l}{N}\log_2 r}$ 为**编码效率**.

- 由定理5.2可得最佳编码效率为 $\eta = \dfrac{H(S)}{H(S)+\varepsilon}, \varepsilon > 0$，所以 $\varepsilon = \dfrac{1-\eta}{\eta} H(S)$
-
- 在已知方差和信源熵的条件下，信源符号序列长度N与最佳编码效率η和允许错误概率 P_E的关系为：

$$N \geq \frac{D[I(s_i)]}{\varepsilon^2 \delta} = \frac{D[I(s_i)]}{H^2(S)} \frac{\eta^2}{(1-\eta)^2 \delta}$$

- 当允许错误概率越小，编码效率又要高，那么信源符号序列长度N必须越长．在实际情况下，要实现几乎无失真的定长编码，N需要的长度将会大到难以实现。

【例 5.2】

设有离散无记忆信源

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ 0.40 & 0.18 & 0.10 & 0.10 & 0.07 & 0.06 & 0.05 & 0.04 \end{bmatrix}$$

如果对信源符号采用定长二元编码，要求编码效率 $\eta = 90\%$，允许错误概率 $P_E \leqslant 10^{-6}$，求所需信源序列长度 N.

解

信息熵 $H(S) = E[-\log_2 p(s_i)] = -\sum_{i=1}^{8} p(s_i)\log_2 p(s_i) = 2.55$ 比特/信源符号

自信息的方差

$$D[I(s_i)] = \sum_{i=1}^{8} p(s_i)[-\log_2 p(s_i)]^2 - H^2(S) = 7.82$$

$$\varepsilon = \frac{1-\eta}{\eta} H(S) = 0.28$$

所以

$$N \geqslant \frac{D[I(s_i)]}{\varepsilon^2 \delta} = \frac{7.82}{0.28^2 \times 10^{-6}} \approx 9.8 \times 10^7 \approx 10^8$$

即信源序列长度 N 需长达 10^8 以上才能实现上述给定的要求.

【例 5.3】
设离散无记忆信源 $\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 \\ 3/4 & 1/4 \end{bmatrix}$ 要求 $\eta = 0.96, \delta \leqslant 10^{-5}$，求 N.

解

信源熵 $H(S) = \frac{1}{4}\log_2 4 + \frac{3}{4}\log_2 \frac{4}{3} = 0.811$ 比特/信源符号

自信息的方差

$$D[I(s_i)] = \sum_{i=1}^{2} p(s_i)[-\log_2 p(s_i)]^2 - H^2(S)$$

$$= \frac{3}{4} \times \left(\log_2 \frac{3}{4}\right)^2 + \frac{1}{4} \times \left(\log_2 \frac{1}{4}\right)^2 - (0.811)^2$$

$$= 0.4715$$

因为

$$\varepsilon = \frac{1-\eta}{\eta} H(S)$$

所以

$$N \geqslant \frac{D[I(s_i)]}{\varepsilon^2 \delta} = \frac{D[I(s_i)]}{\delta} \frac{\eta^2}{(1-\eta)^2 H^2(S)}$$

$$= \frac{0.4715}{10^5} \times \frac{(0.96)^2}{(0.04)^2 \times (0.811)^2}$$

$$\approx 4.13 \times 10^7$$

即信源序列长度长达 4×10^7 以上，才能实现给定的要求，这在实际中是很难实现的.

一般来说，当 N 有限时，高传输效率的定长码往往要引入一定的失真和错误，但是变长码则可以在 N 不大时实现无失真编码.

第三节 、变长码及变长编码定理

1. 克劳夫特不等式、麦克米伦不等式
2. 平均码长及信息传输率
3. 香农第一编码定理
   （无失真变长编码定理）
4. 最佳编码

变长码要成为唯一可译码不仅本身应是非奇异的，而且它的有限长 N 次扩展码也应是非奇异的. 例如表 5.6 中码 2 是非奇异码，但是当信宿收到“00”时，它不能判断是 s_1s_1 还是 s_3，所以不是唯一可译码.

表 5.6　变长码

信源符号 s_i	概率分布	码 1	码 2	码 3	码 4
s_1	1/2	0	0	1	1
s_2	1/4	11	10	10	01
s_3	1/8	00	00	100	001
s_4	1/8	11	01	1000	0001

1、 Kraft不等式和McMillan不等式

- 什么条件才可以构成即时码和唯一可译码

克劳夫特不等式、麦克米伦不等式

定理 5.3 设信源符号集为 $S=\{s_1,s_2,\cdots,s_q\}$，码符号集为 $X=\{x_1,x_2,\cdots,x_r\}$，对信源进行编码，得到的码为 $C=\{w_1,w_2,\cdots,w_q\}$，码长分别为 $l_1, l_2,\cdots,l_q$. 即时码存在的充要条件是

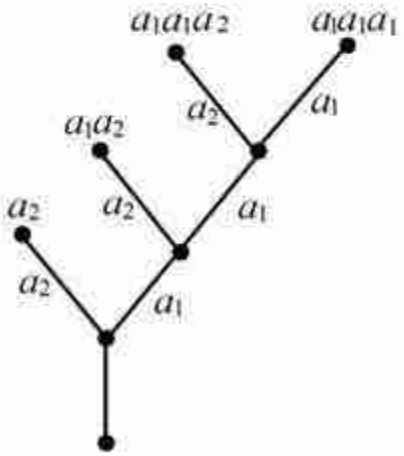
$$\sum_{i=1}^{q} r^{-l_i} \leqslant 1 \tag{5.11}$$

这称为 Kraft 不等式.

- 例3.1.1设二进制码树中X∈｛x1，x2，x3，x4｝，K1=1，K2=2，K3=2，K4=3，应用上述判断定理，可得

$$\sum_{i=1}^{4} 2^{-Ki} = 2^{-1} + 2^{-2} + 2^{-2} + 2^{-3} = \frac{9}{8} > 1$$

因此，不存在满足这种Ki的唯一可译码，
用树码进行检查如图



树码

• Kraft不等式和McMillan不等式的意义

由 Kraft 不等式可知，给定 r 和 q，只要允许码字长度可以足够长，则码长总可以满足 Kraft 不等式而得到即时码，Kraft 不等式指出了即时码的码长必须满足的条件. 后来，McMillan 证明对于唯一可译码的码长也必须满足此不等式. 在码长的选择上唯一可译码并不比即时码有更宽松的条件. 对于唯一可译码，该不等式又称为 **McMillan 不等式**.

定理 5.4 指出了唯一可译码中 r、q、l_i 之间的关系，如果满足这个不等式的条件，则一定能够构成至少一种唯一可译码，否则，无法构成唯一可译码. 它给出了唯一可译变长码的存在性.

表 5.6　变长码

信源符号 s_i	概率分布	码 1	码 2	码 3	码 4
s_1	1/2	0	0	1	1
s_2	1/4	11	10	10	01
s_3	1/8	00	00	100	001
s_4	1/8	11	01	1000	0001

例如在表 5.6 中，码 1、码 2 码长为 $l_1=1, l_2=l_3=l_4=2$，$\sum_{i=1}^{4} 2^{-l_i}=\frac{5}{4}\geqslant 1$，所以不能构成唯一可译码. 而码 3、码 4 的码长为 $l_1=1, l_2=2, l_3=3, l_4=4$，$\sum_{i=1}^{4} 2^{-l_i}=\frac{15}{16}\leqslant 1$，可以构成唯一可译码. 当然满足此条件的也不一定就是唯一可译码，例如 $C=\{1,01,011,0001\}$，但至少可以找到一种唯一可译码. 同样满足此码长条件的码也可能不是即时码，但至少可以找到一种即时码.

另外，从定理 5.3 和定理 5.4 可以得到一个重要的结论，即任何一个唯一可译码均可用一个相同码长的即时码来代替，因为即时码很容易用树图法构造，因此要构造唯一可译码，只要构造即时码就可以了.

2、平均码长及信息传输率

由 5.3.2 节讨论可知，对于已知信源 S 可用码符号集合 X 进行变长编码，而且对同一信源用同一码符号集编成的即时码或唯一可译码可有很多种，究竟哪一种最好呢？从高效传输信息的角度希望选择由短的码符号组成的码字，就是用码长作为选择标准，为此引入码的平均长度.

定义 5.8 设有信源

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & \cdots & s_q \\ p(s_1) & p(s_2) & \cdots & p(s_q) \end{bmatrix}$$

编码后的码字分别为 $w_1, w_2, \cdots, w_q$，各码字相应的码长分别为 $l_1, l_2, \cdots, l_q$. 因为是唯一可译码，信源符号 s_i 和码字 w_i 一一对应，则定义此码的**平均码长**为

$$\overline{L} = \sum_{i=1}^{q} p(s_i) l_i \text{ 码符号/信源符号} \tag{5.20}$$

其中，$\overline{L}$ 表示每个信源符号平均需用的码元数.

当信源给定时，信源熵 $H(S)$ 就确定了，而编码后每个信源符号平均用 $\overline{L}$ 个码元来变换，故平均每个码元载荷的信息量即编码后信源的信息传输率.

定义 5.9 编码后信源的**信息传输率**为

$$R = H(X) = \frac{H(S)}{\overline{L}} \text{ 比特/码符号} \tag{5.21}$$

如果传输一个码符号平均需要 t 秒时间，则编码后信源每秒钟提供的信息量为

$$R_t = \frac{H(S)}{\overline{L}t} \tag{5.22}$$

可以看出，$\overline{L}$ 越小，则 R_t 越大，信息传输率就越高，因此我们感兴趣的码是使平均码长 $\overline{L}$ 最短的码.

定义 5.10 对于给定的信源和码符号集，若有一个唯一可译码，其平均码长 $\overline{L}$ 小于所有其他唯一可译码，则称这种码为**紧致码**或**最佳码**.

无失真信源编码的核心问题就是寻找紧致码.

下面的定理给出了紧致码的平均码长可能达到的理论极限.

平均码长的极限值

定理 5.6 若一个离散无记忆信源 S，熵为 $H(S)$，用拥有 r 个码符号的码符号集 $X=\{x_1,x_2,\cdots,x_r\}$ 对 S 进行无失真编码，则总可以找到一种唯一可译码，其平均码长满足

$$\frac{H(S)}{\log_2 r}\leqslant \overline{L}<1+\frac{H(S)}{\log_2 r} \tag{5.23}$$

不要求掌握证明过程　　JENSEN 不等式

定理 5.6 说明，码字的平均长度 $\overline{L}$ 不能小于极限值 $\frac{H(S)}{\log_2 r}$，否则唯一可译码不存在，同时定理又给出了平均码长的上界. 这不是说大于这个上界就不能构成唯一可译码，而是说即使平均码长小于这个上界也一定存在唯一可译码.

另外还可以看到平均码长的极限值与无失真定长信源编码定理中的极限值是一致的.

3、香农第一编码定理

与无失真定长信源编码定理一样，无失真变长信源编码定理(香农第一定理)也是一个极限定理，是一个存在性定理. 极限和定长编码的极限是一样的.

定理 5.7 设离散无记忆信源 S 的信源熵 $H(S)$，它的 N 次扩展信源 $S^N=\{s_1,s_2,\cdots,s_{q^N}\}$，其熵 $H(S^N)$. 并用码符号 $X=\{x_1,x_2,\cdots,x_r\}$ 对信源 S^N 进行编码，总可以找到一种唯一可译码，使信源 S 中每个信源符号所需的平均码长满足

$$\frac{H(S)}{\log_2 r}\leqslant\frac{\overline{L}_N}{N}<\frac{H(S)}{\log_2 r}+\frac{1}{N} \tag{5.34}$$

或者写成

$$H_r(S)\leqslant\frac{\overline{L}_N}{N}<H_r(S)+\frac{1}{N} \tag{5.35}$$

由于离散无记忆信源的 N 次扩展信源 S^N 的熵 $H_r(S^N)$ 是信源 S 的熵 $H(S)$ 的 N 倍，即

$$H_r(S^N)=NH_r(S) \tag{5.37}$$

代入式(5.36)得

$$NH_r(S)\leqslant\overline{L}_N\leqslant NH_r(S)+1 \tag{5.38}$$

当 $N\to\infty$ 时，有

$$\lim_{N\to\infty}\frac{\overline{L}_N}{N}=H_r(S) \tag{5.40}$$

定理 5.7 的结论推广到平稳遍历的有记忆信源(如马尔可夫信源)便有

$$\lim_{N\to\infty}\frac{\overline{L}_N}{N}=\frac{H_\infty}{\log_2 r} \tag{5.41}$$

式中，H_∞ 为有记忆信源的极限熵.

定理 5.7 是香农信息论的主要定理之一. 定理指出，要做到无失真信源编码，每个信源符号平均所需最少的 r 元码元数就是信源的熵值(以 r 进制单位为信息量单位). 若编码的平均码长小于信源的熵值，则唯一可译码不存在，在译码或反变换时必然要带来失真或差错，同时定理还指出，通过对扩展信源进行变长编码，当 $N\to\infty$ 时，平均码长 $\overline{L}$(这时它等于 $\frac{\overline{L}_N}{N}$)可达到这个极限值. 可见，信源的信息熵 $H(S)$ 是无失真信源编码码长的极限值，也可以认为信源的信息熵($H(S)$ 或 H_∞)是表示每个信源符号平均所需最少的二元码符号数.

无失真信源编码定理通常又称为**无噪信道编码定理**，此定理可以表述为：总能对信源的输出进行适当的编码，使得在无噪无损信道上能无差错地以最大信息传输率 C 传输信息，但要使信道的信息传输率 R 大于 C 而无差错地传输则是不可能的.

为了衡量各种编码是否已达到极限情况，我们定义变长码的编码效率.

定义 5.12 设对信源 S 进行无失真编码所得到的平均码长为 $\overline{L}$，则 $\overline{L} \geqslant H_r(S)$，定义

$$\eta = \frac{H_r(S)}{\overline{L}} \tag{5.44}$$

为**编码效率**，$\eta \leqslant 1$.

对同一信源来说，码的平均码长 $\overline{L}$ 越短，越接近极限 $H_r(S)$，信道的信息传输率越高，越接近无噪无损信道的信道容量，这时 η 也越接近于 1，所以用码的编码效率 η 来衡量各种编码的优劣.

另外，为了衡量各种编码与最佳码的差距，引入码的剩余度的概念.

定义 5.13 定义

$$\gamma = 1 - \eta = 1 - \frac{H_r(S)}{\overline{L}} \tag{5.45}$$

为码的**剩余度**.

在二元无噪无损信道中 $r=2$，$\eta = \frac{H(S)}{\overline{L}}$，所以在二元无噪无损信道中信息传输率 $R = \frac{H(S)}{\overline{L}} = \eta$.

注意它们数值相同，单位不同. η 是个无单位的比值，而 R 的单位是比特/码符号. 因此在二元信道中可直接用码的效率来衡量编码后信道的信息传输率是否提高了. 当 $\eta = 1$ 时，即 $R = 1$，达到二元无噪无损信道的信道容量，编码效率最高，码剩余度为零.

与定长码一样，通过对扩展信源进行编码，可以提高编码后信道的信息传输率.

【例 5.5】

有一离散无记忆信源

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 \\ 3/4 & 1/4 \end{bmatrix}$$

求其信息传输率及二次、三次、四次扩展信源的信息传输率.

解

信源熵

$$H(S)=\frac{1}{4}\log_2 4+\frac{3}{4}\log_2\frac{4}{3}=0.811$$

用二元码符号{0,1}来构造一个即时码：$s_1\to 0, s_2\to 1$.

这时，$\overline{L}=1$，编码效率 $\eta=\frac{H(S)}{\overline{L}}=0.811$，信源的信息传输率 $R=0.811$.

表 5.9　变长编码的编码效率

s_i	$p(s_i)$	即时码
s_1s_1	9/16	0
s_1s_2	3/16	10
s_2s_1	3/16	110
s_2s_2	1/16	111

如果对信源 S 的二次扩展信源 S^2 进行编码，得到它的一种即时码如表 5.9 所示.

这时码的平均长度

$$\overline{L_2}=\frac{9}{16}\times 1+\frac{3}{16}\times 2+\frac{3}{16}\times 3+\frac{1}{16}\times 3=\frac{27}{16}\text{二元码符号/二个信源符号}$$

信源符号的平均码长 $\overline{L}=\frac{\overline{L_2}}{2}=\frac{27}{32}$二元码符号/信源符号

编码效率 $\eta_2=\frac{0.811\times 32}{27}=0.961$，$R_2=0.961$ 比特/二元码符号

可见编码复杂了些，但信息传输效率有了提高.

用同样方法进一步对信源 S 的三次和四次扩展信源进行编码，并求出其编码效率为 $\eta_3=0.985$，$\eta_4=0.991$. 信道的信息传输率分别为 $R_3=0.985$，$R_4=0.991$.

将此例与例 5.3 相比较，对于同一信源，要求编码效率达到 96% 时，变长码只需对二次扩展信源($N=2$)进行编码，而等长码则要求 $N \geqslant 4.13\times10^{7}$. 因此用变长码编码时，$N$ 不需很大就可以达到相当高的编码效率，而且可实现绝对无失真编码，随着扩展信源次数 N 的增加，编码的效率越来越接近于 1. 编码后信道的信息传输率 R 也越来越接近于无噪无损二元信道的信道容量 $C=1$ 比特/二元码符号，从而达到信源与信道匹配，使信道得到充分利用.

【例 5.3】

设离散无记忆信源

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 \\ 3/4 & 1/4 \end{bmatrix}$$

要求 $\eta=0.96, \delta \leqslant 10^{-5}$，求 N.

$$N \geqslant \frac{D[I(s_i)]}{\varepsilon^2 \delta} = \frac{D[I(s_i)]}{\delta} \frac{\eta^2}{(1-\eta)^2 H^2(S)} \approx 4.13\times10^{7}$$

4、最佳编码

(1)最佳码定义是什么?

凡是能载荷一定的信息量，且码字的平均长度最短，可分离的变长码的码字集合都可称为最佳码。

(2)最佳编码思想是什么?

将概率大的信息符号编以短的码字，概率小的符号编以长的码字，使得平均码字长度最短。

(3)最佳码的编码主要方法有哪些?

香农（Shannon）、费诺（Fano）、哈夫曼（Huffman）编码等。