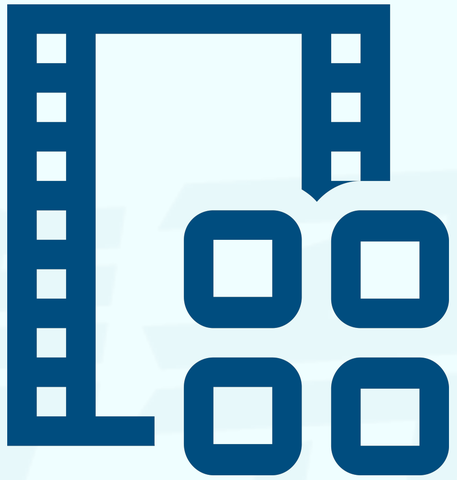
无失真信源编码

信息论与编码原理不挂科 第五讲

无失真信源编码

信息论与编码原理不挂科 第五讲



4大模块

13道题目

信息论与编码原理不挂科 第五讲

无失真信源编码

模块1 信源编码器与基本术语

模块2 分组码与唯一可译性

模块3 定长码与定长编码定理

模块4 变长码

信源编码的作用与目的

- 信源编码的作用：

  1. 使信源适合于信道的传输，用信道能传输的符号代表信源发出的消息。
  2. 在不失真或允许一定失真的条件下，用尽可能少的符号来传递信源消息。

- 信源编码的目的：提高通信的有效性，通常通过压缩信源的冗余度来实现。

  冗余度取决于符号间记忆的相关性与符号概率分布的非均匀性。

  压缩方式：概率匹配——统计编码（*Huffman* 编码，算术编码）

  去除码符号间的相关性，再对各独立分量（标量）进行编码（变换编码）

  利用条件概率进行编码（预测编码）

  利用联合概率进行编码（无记忆信源的扩展编码）

信源编码理论基础

- 信源编码理论基础：
  - 无失真信源编码定理
  - 限失真信源编码定理

信源编码理论基础

- 信源编码理论基础：

  - **无失真信源编码定理** → 统计匹配编码，根据信源概率分布选用与之匹配的编码
  - 限失真信源编码定理

信源编码器与基本术语

小节1 信源编码的基本术语

小节2 几类常见的编码类型

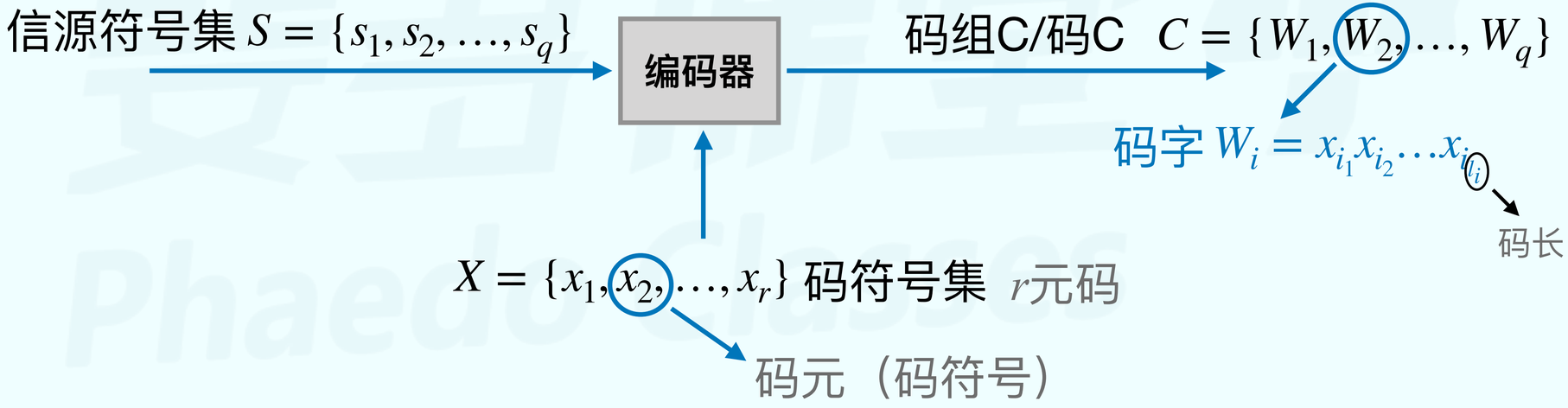
信源编码器与基本术语

小节1 信源编码的基本术语

小节2 几类常见的编码类型

信源编码

将**信源符号序列**按一定的数学规律映射成**码符号序列**的过程。



信源编码的过程是将信源符号集中的消息（符号）s_i（或长为 N 的信源符号序列）映射成由码符号 x_i 组成的长度为 l_i 的一个对应的码符号序列 W_i 。

信源编码器与基本术语

小节1 信源编码的基本术语

小节2 几类常见的编码类型

码的种类

- 定长码：若一组码中所有码字的码长相同，则称为定长码。
- 变长码：若一组码中所有码字的码长各不相同，则称为变长码。
- 奇异码：若一组码中存在相同的码字，则称为奇异码。
- 非奇异码：若一组码中所有码字都不相同，则称为非奇异码。
- N次扩展码：信源符号集 $S=\{s_1, s_2, \ldots, s_q\}$ $\longleftrightarrow$ 码组 $C=\{W_1, W_2, \ldots, W_q\}$

$\downarrow$ N次扩展（无记忆）

$S^N=\{\alpha_1, \alpha_2, \ldots, \alpha_{q^N}\}$ $\longleftrightarrow$ N次扩展码 $C^N=\{V_1, V_2, \ldots, V_{q^N}\}$

$\alpha_i = s_{i_1} s_{i_2} \ldots s_{i_N}$ $\qquad$ $V_j = W_{j_1} W_{j_2} \ldots W_{j_N}$ （其中 $j=1,2,\ldots,q^N$）

例题5–1 信源概率空间 $\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 \\ p(s_1) & p(s_2) & p(s_3) & p(s_4) \end{bmatrix}$，指出其中的定长码、变长码、奇异码、非奇异码，

并对码2进行无记忆二次扩展。

信源符号	符号概率	码 1	码 2	码 3
s_1	$p(s_1)$	00	0	0
s_2	$p(s_2)$	01	01	11
s_3	$p(s_3)$	10	001	00
s_4	$p(s_4)$	11	111	11

解析5–1 定长码：码 1

变长码：码 2 、码 3

奇异码：码 3

非奇异码：码 1 、码 2

例题5–1 信源概率空间 $\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 \\ p(s_1) & p(s_2) & p(s_3) & p(s_4) \end{bmatrix}$，指出其中的定长码、变长码、奇异码、非奇异码，

并对码2进行无记忆二次扩展。

解析5–1

信源符号	码字	码 2
s_1	W_1	0
s_2	W_2	01
s_3	W_3	001
s_4	W_4	111

将码 2 进行二次扩展 →

二次扩展信源符号 $\alpha_j(j=1,2,\ldots,16)$	二次扩展码码字 $W_j(j=1,2,\ldots,16)$
$\alpha_1 = s_1s_1$	$V_1 = W_1W_1 = 00$
$\alpha_2 = s_1s_2$	$V_2 = W_1W_2 = 001$
$\alpha_3 = s_1s_3$	$V_3 = W_1W_3 = 0001$
……	……
$\alpha_{16} = s_4s_4$	$V_{16} = W_4W_4 = 111111$

分组码与唯一可译性

小节1 唯一可译性

小节2 唯一可译性的判别准则

分组码与唯一可译性

小节1 唯一可译性

小节2 唯一可译性的判别准则

唯一可译性

分组码：将信源符号集中的每个信源符号 s_i 映射为一个固定的码字 W_i 。

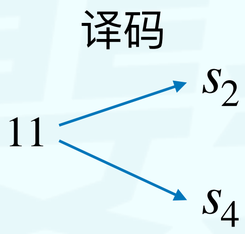
唯一可译码：任意一串有限长的码序列符号只能被唯一地译为对应的信源符号序列。

唯一可译码充要条件：编码任意次扩展均为非奇异码 $\begin{cases} \text{码字与信源符号一一对应} \\ \text{码字序列与信源符号序列一一对应} \end{cases}$

唯一可译码的判断

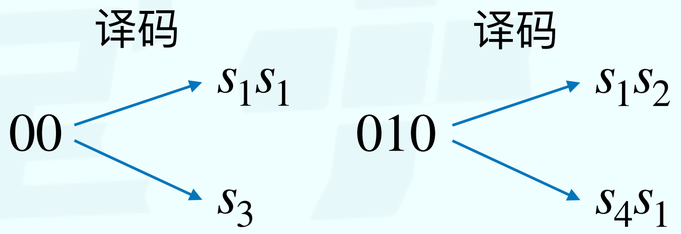
- 奇异码一定不是唯一可译码

s_1	0
s_2	11
s_3	00
s_4	11



- 非奇异码不一定是唯一可译码

s_1	0
s_2	10
s_3	00
s_4	01



- 等长非奇异码一定是唯一可译码
- 唯一可译码又可分为即时码和非即时码

即时码：某个唯一可译码在接收到一个完整的码字时，无需参考后续的码符号就能立即译码，则为即时码。

例题5-2 判断下列两组码字的即时性。

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

解析5-2 给定码字 11010010001

若用码1译码 11010010001 → 1,10,…? / 1,1,…? 需要后续码字方可译码

若用码2译码 11010010001 → 1,1,01,001,0001 无需看后续码字即可译码

则码1是非即时码，码2是即时码。

★ 即时码的充要条件：码组中任一码字都不是其他码字的前缀。

另：等长非奇异码一定是即时码。

分组码与唯一可译性

小节1 唯一可译性

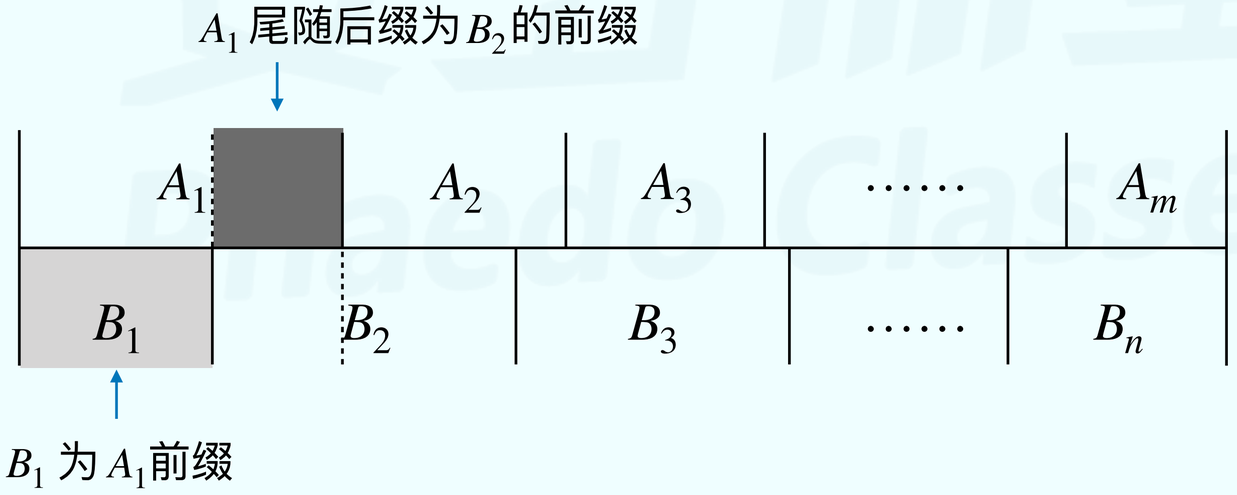
小节2 唯一可译性的判别准则

唯一可译码的判别准则

码符号序列译码二义性的存在条件 { 存在前缀码同时存在后缀码

码符号序列的尾部以不同码字结束

示意图：



唯一可译码的判断方法：将码中所有可能的尾随后缀组成一个集合 F，当且仅当 F 中没有包含任一码字时，该码字为唯一可译码。

唯一可译码的判别准则

判断步骤：1.初始化：$S_0 = C$（将已知码置于集合 S_0 中）

2.构造集合 S_1：考察 S_0 中所有码字，若一个码字是另一个码字的前缀，则将后缀置于 S_1 中。

3.构造集合 $S_n(n > 1)$：将 S_0 与 S_{n-1} 比较

（1）如果 S_0 中有码字是 S_{n-1} 元素中的前缀，则将后缀置于 S_n 中。

（2）同样，若 S_{n-1} 中有元素是 S_0 中码字的前缀，则也将 S_0 中码字的后缀置于 S_n 中。

4.检验 S_n：

（1）若 S_n 是空集，则码 C 是唯一可译码，结束。

（2）若 S_n 中某个元素与 S_0 中的某个元素相同，则码 C 不是唯一可译码，结束。

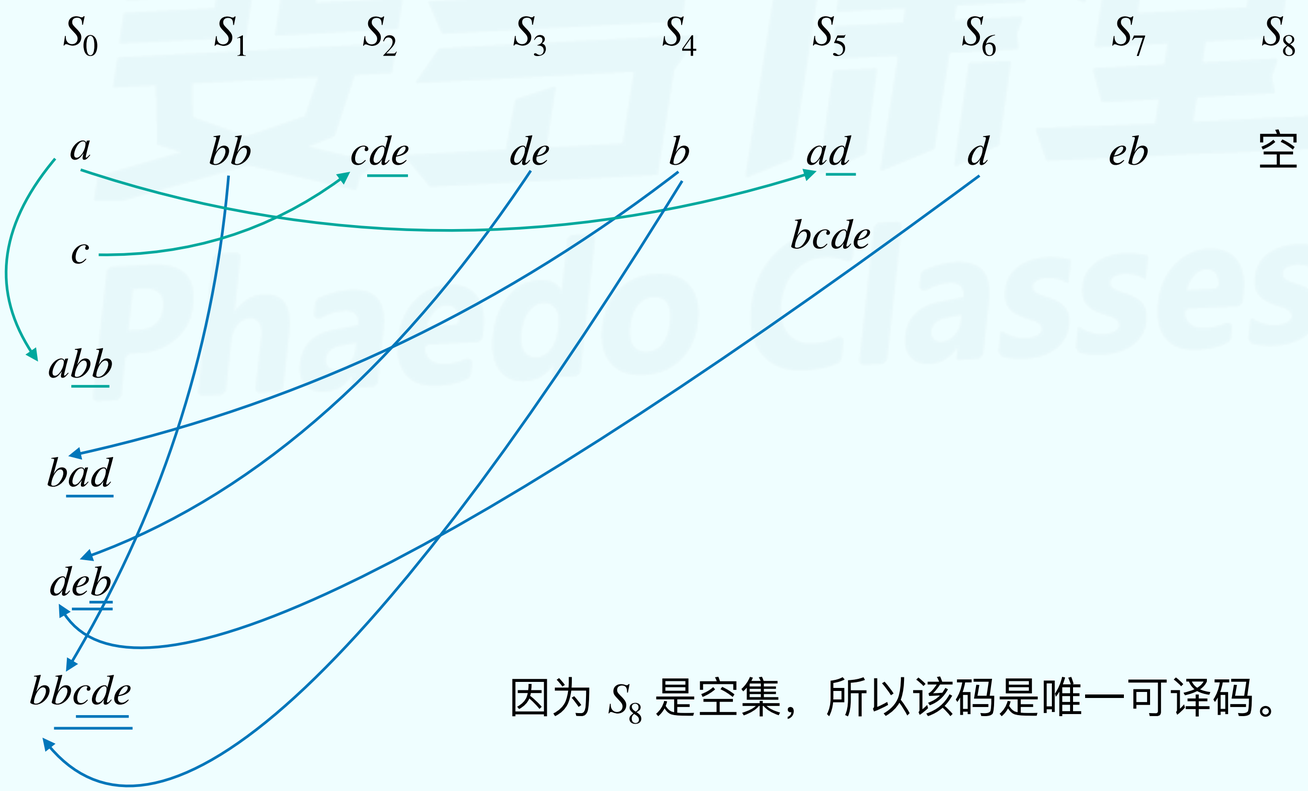
（3）若上述两个条件均不满足，则返回步骤 3 。

例题5–3 判断下列码是否是唯一可译码。

(1) $S_0 = \{a, c, abb, bad, deb, bbcde\}$

(2) $S_0 = \{a, c, ad, abb, bad, deb, bbcde\}$

解析5–3 (1) 列表



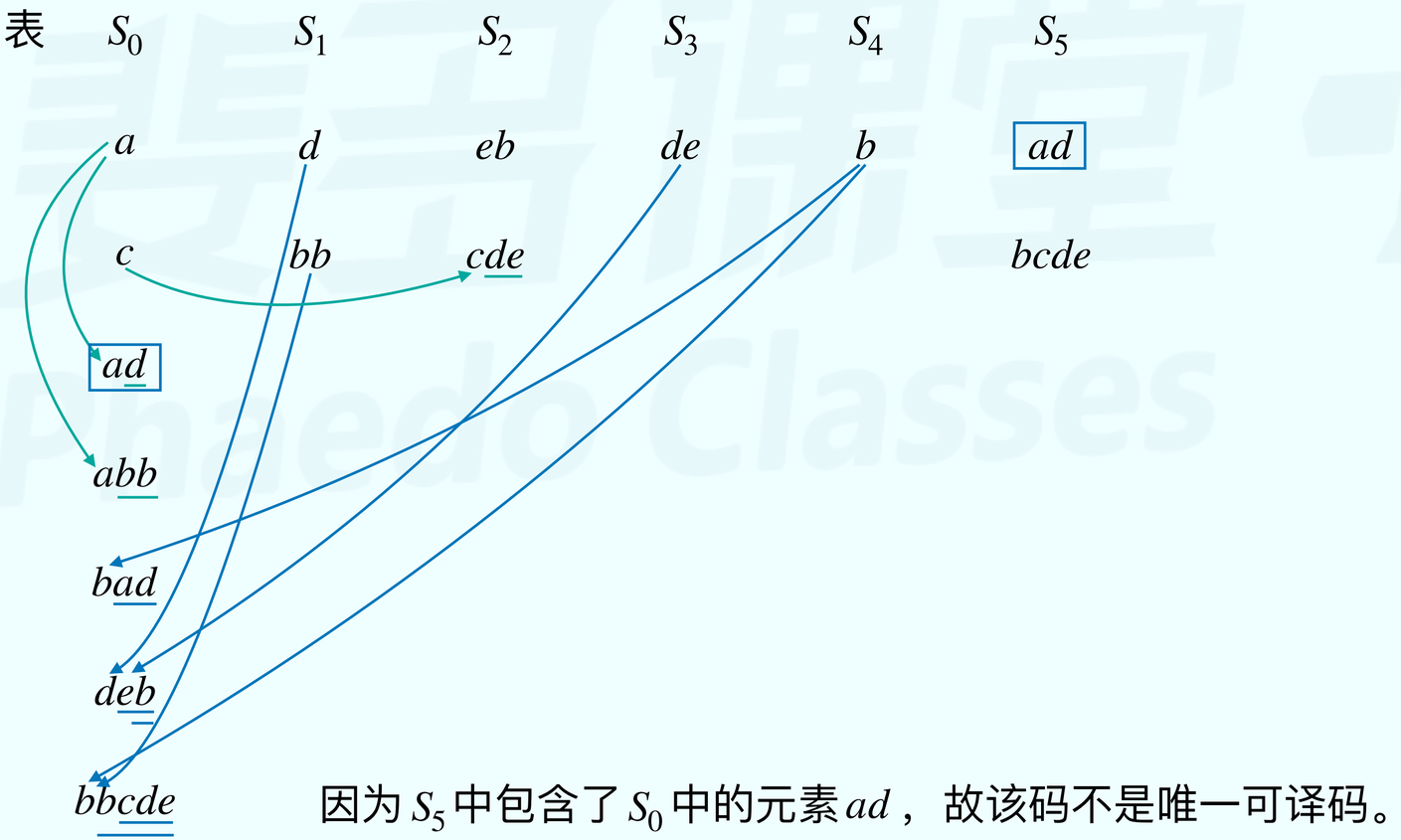
因为 S_8 是空集，所以该码是唯一可译码。

例题5–3 判断下列码是否是唯一可译码。

(1) $S_0 = \{a, c, abb, bad, deb, bbcde\}$

(2) $S_0 = \{a, c, ad, abb, bad, deb, bbcde\}$

解析5–3 (2) 列表



因为 S_5 中包含了 S_0 中的元素 ad，故该码不是唯一可译码。

「小结」各类码之间的相互关系

- 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
        - 即时码
        - 非即时码

定长码与定长编码定理

小节1 定长码

小节2 定长信源编码定理

定长码与定长编码定理

小节1 定长码

小节2 定长信源编码定理

定长码

定长码是所有码字码长相同的码组。

我们在前面讲到定长的非奇异码一定是唯一可译码，故我们来讨论定长码长与非奇异性的关系。

简单信源 S 进行定长编码时：

- 设信源符号集中共有 q 个符号（消息）$S=\{s_1, s_2, \ldots, s_q\}$；
- 码符号集中共有 r 种码元 $X=\{x_1, x_2, \ldots, x_r\}$；
- 编码后定长码码长为 l，则码字总数为 r^l，

若满足非奇异性，则需 $r^l \geq q$，即码字总数 $\geq$ 消息数。

若对 N 次扩展信源进行定长编码，满足非奇异性，则需 $r^l \geq q^N$。

例题5–4 英文字母表中，每一字母用定长编码转换成二进制表示，码字的最短长度应为多少？若对英语字母信源的 2 次、4 次扩展信源进行二元编码，则码字的最短长度又是多少？

解析5–4 ① 对英文字母信源进行编码时，信源符号数 $q=26$ 、码符号数 $r=2$ 。

$$r^l \geq q \longrightarrow l \geq \frac{\log q}{\log r}=\frac{\log 26}{\log 2}=4.7$$

所以最短长度 $l_{min}=5$ 二元符号/信源符号。

② 对英文字母信源的2次扩展信源进行编码时，信源符号数 $q^N=26^2$ 、码符号数 $r=2$ 。

$$r^l \geq q^N \longrightarrow l \geq \frac{\log q^N}{\log r}=\frac{\log 26^2}{\log 2}=9.4$$

所以最短长度 $l_{min}=10$ 二元符号/ 2 个信源符号，每个信源符号对应的平均码长为 $\frac{l_{min}}{2}=5$ 二元符号/信源符号。

③ 对英文字母信源的4次扩展信源进行编码时，信源符号数 $q^N=26^4$ 、码符号数 $r=2$ 。

$$r^l \geq q^N \longrightarrow l \geq \frac{\log q^N}{\log r}=\frac{\log 26^4}{\log 2}=18.8$$

所以最短长度 $l_{min}=19$ 二元符号/ 4 个信源符号，每个信源符号对应的平均码长为 $\frac{l_{min}}{4}=4.75$ 二元符号/信源符号。

定长码与定长编码定理

小节1 定长码

小节2 定长信源编码定理

定长信源编码定理

设离散平稳无记忆信源的熵为 $H(S)$，若对 N 次扩展信源 S^N 进行长度为 l 定长编码，则对 $\forall \varepsilon > 0$，

- 只要满足 $\frac{l}{N} \geq \frac{H(s)+\varepsilon}{\log r}$，则当扩展次数 N 足够大时，可几乎无失真编码，即译码错误概率无限小。
- 反之，若 $\frac{l}{N} \leq \frac{H(s)-2\varepsilon}{\log r}$，则不可能实现无失真编码，$N$ 足够大时，译码错误概率为 1。

定长信源编码的编码信息率与编码效率

- **编码信息率 R 的计算：** $R=\dfrac{H(S)}{l/N}=\dfrac{bit/\text{信源符号}}{r\ \text{元码符号/信源符号}}=bit/r$ 元码符号，指编码后平均每个码符号载荷的实际信息量。

- **编码效率 η 的计算：** $\eta=\dfrac{R}{\log r}=\dfrac{H(S)}{\log r}\cdot\dfrac{N}{l}$ 。若 $\dfrac{l}{N}\geq\dfrac{H(s)+\varepsilon}{\log r}$，则 $\eta\leq\dfrac{H(S)}{H(S)+\varepsilon}$，称为最佳编码效率。

定长信源编码定理的说明

- 当选定定长码的码长满足 $\frac{l}{N} \geq \frac{H(s)+\varepsilon}{\log r}$ 时，N 足够大时，可实现几乎无失真编码。
- 在这种编码方式下，只对扩展信源中的典型序列进行编码，非典型序列被舍弃， 虽然总概率极小，但还是有可能出现，此时只有当N趋于无穷大时，错误概率才趋近于0。
- 定长信源编码定理同样适用于离散平稳有记忆信源，将信源熵 $H(S)$ 换为 H_{∞} 极限熵即可。
- 如果对信源的 N 次扩展进行编码，给定编码效率和错误概率时，容许错误概率越小，编码效率越高时，要求信源序列长度，即扩展次数 N 越长，要实现几乎无失真编码，必须以 N 大到难以实现为代价。

变长码

小节1 Kraft不等式与McMillan不等式

小节2 变长码的平均码长与平均码长界限定理

小节3 变长无失真信源编码定理

小节4 常用变长编码方法

变长码

小节1 Kraft不等式与McMillan不等式

小节2 变长码的平均码长与平均码长界限定理

小节3 变长无失真信源编码定理

小节4 常用变长编码方法

Kraft不等式与McMillan不等式

- *Kraft* 不等式

  设信源符号集为 $S=\{s_1, s_2, \ldots, s_q\}$，码符号集为 $X=\{x_1, x_2, \ldots, x_r\}$，码字为 $C=\{W_1, W_2, \ldots, W_q\}$，

  码字的码长分别为 $l_1, l_2, \ldots, l_q$，则即时码存在的充要条件是 $\sum_{i=1}^{q} r^{-l_i} \leq 1$，该不等式称为 *Kraft* 不等式。

  该不等式给出了即时码的码长必须满足的条件。

- *McMillan* 不等式

  在上述条件下，唯一可译码存在的充要条件是 $\sum_{i=1}^{q} r^{-l_i} \leq 1$。

Kraft不等式与McMillan不等式的说明

- *Kraft* 不等式与 *McMillan* 不等式在形式上完全相同。

  即时码属于唯一可译码。但在码长的选择上，唯一可译码并不比即时码有更宽的条件。

  若存在一个码长为 $l_1, l_2, \ldots, l_q$ 的唯一可译码，则一定存在相同码长的即时码。

- 上述不等式给出的是存在性定理。

  即：当满足 *Kraft*（或 *McMillan* 不等式），必然可构造出满足其码长要求的即时码（或唯一可译码），否则不可，为存在性的验证。两不等式可作为判断一种码**不是**即时码（或唯一可译码）的依据，而不能作为判断判断一种码**是**即时码（唯一可译码）的依据。

例题5-5 设信源 $\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\ p(s_1) & p(s_2) & p(s_3) & p(s_4) & p(s_5) & p(s_6) \end{bmatrix}$，且满足 $\sum_i p(s_i) = 1$。将此信源编码为 r 元唯一可译变长码，其对应的码长 $(l_1, l_2, l_3, l_4, l_5, l_6) = (1,1,2,3,2,3)$，求 r 值的下限。

解析5-5 若要构造出唯一可译变长码，则相关码长必须满足 *McMillan* 不等式 $\sum_{i=1}^{q} r^{-l_i} \le 1$。

代入 $(l_1, l_2, l_3, l_4, l_5, l_6) = (1,1,2,3,2,3)$ 可得：$r^{-1} + r^{-1} + r^{-2} + r^{-3} + r^{-2} + r^{-3} \le 1$，即 $r^{-1} + r^{-2} + r^{-3} \le \frac{1}{2}$。

当 $r = 2$ 时，$r^{-1} + r^{-2} + r^{-3} = \frac{7}{8} > \frac{1}{2}$，不等式不成立。

当 $r = 3$ 时，$r^{-1} + r^{-2} + r^{-3} = \frac{13}{27} < \frac{1}{2}$，不等式成立。

故 r 值的下限为 3。

变长码

小节1 Kraft不等式与McMillan不等式

小节2 变长码的平均码长与平均码长界限定理

小节3 变长无失真信源编码定理

小节4 常用变长编码方法

变长码的平均码长

给定信源 $\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & \cdots & s_q \\ p(s_1) & p(s_2) & \cdots & p(s_q) \end{bmatrix}$，编码后的码字 $W_1, W_2, \ldots, W_q$ 的长度为 $l_1, l_2, \ldots, l_q$，

定义平均码长为 $\bar{L} = \sum_{i=1}^{q} p(s_i) l_i$，即码字长度对信源符号概率的统计平均。单位为码符号/信源符号。

将平均码长最小的唯一可译码称为紧致码或最佳码。码率 $R = \frac{H(s)}{\bar{L}}$ 比特/码符号

对 $\bar{L} = \sum_{i=1}^{q} p(s_i) l_i$ 的说明：

- 结构上，若 q 、r 、l_i 符合 $Kraft$ 不等式，则必定存在唯一可译码（即时码）。
- $\bar{L}$ 与码的结构和信源 S 的统计特性有关，不同 l_i 与 $p(s_i)$ 的搭配，$\bar{L}$ 不同。
- 信源编码的目的是提高通信的有效性，即希望找到紧致码或最佳码。

平均码长界限定理

定义：给定熵为$H(S)$的离散无记忆信源，用r个码符号对其进行编码，则一定能找到一种无失真信源编码构成唯一可译码，使其平均码长满足 $\frac{H(S)}{\log r} \leq \bar{L} \leq \frac{H(S)}{\log r} + 1$。

证明过程略。

但通过证明我们可以得到一个重要的结论：在变长编码过程中，大概率消息/符号用较短的码字表示，小概率消息/符号用较长的码字表示，通过这样的概率匹配可以切实降低平均码长。

变长码

变长无失真信源编码定理（香农第一定理）

内容：设离散无记忆信源的熵为 $H(S)$ ，它的 N 次扩展信源为 S^N，对扩展信源 S^N 进行编码，一定可以找到一种编码方法构成唯一可译码，使平均码长满足 $\frac{H(S)}{\log r} \leq \frac{\bar{L}_N}{N} \leq \frac{H(S)}{\log r} + \frac{1}{N}$。

说明：

- 平均码长界限定理只考虑单个信源符号，而香农第一定理考虑的是消息序列，将 $H(S^N) = NH(S)$ 代入平均码长界限定理即可证得香农第一定理。
- $N \to \infty$ 时，$\lim\limits_{N\to\infty} \frac{\bar{L}_N}{N} = \frac{H_\infty}{\log r}$ 。即一般离散信源需考虑符号间依赖性进一步降低平均码长。
- 无失真信源编码平均码长下界为信源熵 $H_r(S)$，码率为 $R = \frac{H(S)}{\bar{L}}$ ，其中 $\bar{L} = \frac{\bar{L}_N}{N}$ ；效率为 $\eta = \frac{R}{\log r} = \frac{H(S)/\bar{L}}{\log r}$ 。

变长无失真信源编码定理（香农第一定理）

- 无失真信源编码平均码长下界为信源熵 $H_r(S)$，码率为 $R=\dfrac{H(S)}{\bar{L}}$，其中 $\bar{L}=\dfrac{\bar{L}_N}{N}$；

  效率为 $\eta=\dfrac{R}{\log r}=\dfrac{H(S)/\bar{L}}{\log r}$。

$$\eta=\frac{R}{\log r}=\frac{H(S)/\bar{L}}{\log r}=\frac{H(S)}{\bar{L}\cdot\log r}=\frac{H_r(S)}{\bar{L}}$$

编码后每个码符号实际载荷的信息量

编码前每个信源符号实际载荷的信息量

编码信源符号所需的平均码长下界

编码信源符号的实际平均码长

编码后每个码符号最大能载荷的信息量

编码后每个信源符号能够载荷的最大信息量

香农第一定理给出了本课程的第一个理论极限：平均码长的压缩下限，它和信源的熵有关。

例题5-6 设某无记忆二元信源，概率 $P_1 = P(1) = 0.1$，$P_2 = P(0) = 0.9$ 采用下述编码方案，首先根据 0 的长度编为 9 个中间码字，再将九个码字换成二元变长码如下表：

信源符号	中间码	二元码字
1	s_0	1000
01	s_1	1001
001	s_2	1010
0001	s_3	1011
00001	s_4	1100
000001	s_5	1101
0000001	s_6	1110
00000001	s_7	1111
00000000	s_8	0

(1) 试问最后的二元变长编码是否唯一可译?

(2) 试求中间码对应的信源序列的平均长度 $\bar{L}_1$；

(3) 试求中间码对应的二元变长码字的平均码长 $\bar{L}_2$；

(4) 计算比值 $\bar{L}_2/\bar{L}_1$ 并解释其意义，计算这种编码方式的编码效率。

例题5-6 设某无记忆二元信源，概率 $P_1 = P(1) = 0.1$，$P_2 = P(0) = 0.9$ 采用下述编码方案，首先根据 0 的长度编为 9 个中间码字，再将九个码字换成二元变长码如下表：

信源符号	中间码	二元码字
1	s_0	1000
01	s_1	1001
001	s_2	1010
0001	s_3	1011
00001	s_4	1100
000001	s_5	1101
0000001	s_6	1110
00000001	s_7	1111
00000000	s_8	0

（1）试问最后的二元变长编码是否唯一可译?

解析5-6 （1）二元码字中任意码字均不是其他码字的前缀，因此其为即时码，从而其为唯一可译码。

例题5-6 设某无记忆二元信源，概率 $P_1 = P(1) = 0.1$，$P_2 = P(0) = 0.9$ 采用下述编码方案，首先根据 0 的长度编为 9 个中间码字，再将九个码字换成二元变长码如下表：

信源符号	中间码	二元码字
1	s_0	1000
01	s_1	1001
001	s_2	1010
0001	s_3	1011
00001	s_4	1100
000001	s_5	1101
0000001	s_6	1110
00000001	s_7	1111
00000000	s_8	0

(2) 试求中间码对应的信源序列的平均长度 $\bar{L}_1$；

解析5-6 (2) 信源无记忆，即序列中各符号间相互独立。

故概率 $p(1) = p(1) = 0.1$　$p(01) = p(0)p(1) = 0.09$　$p(001) = p^2(0)p(1) = 0.081$　$p(0001) = p^3(0)p(1) = 0.0729$

$p(00001) = p^4(0)p(1) = 0.06561$　$p(000001) = p^5(0)p(1) = 0.059049$　$p(0000001) = p^6(0)p(1) = 0.0531441$

$p(0000001) = p^7(0)p(1) = 0.04782969$　$p(0000000) = p^8(0) = 0.43046721$

故信源序列平均长度 $\bar{L}_1 = \sum_{i=1}^{8} p(s_i)l_i = 5.7bit/$信源符号序列。

例题5-6 设某无记忆二元信源，概率 $P_1 = P(1) = 0.1$，$P_2 = P(0) = 0.9$ 采用下述编码方案，首先根据 0 的长度编为 9 个中间码字，再将九个码字换成二元变长码如下表：

信源符号	中间码	二元码字
1	s_0	1000
01	s_1	1001
001	s_2	1010
0001	s_3	1011
00001	s_4	1100
000001	s_5	1101
0000001	s_6	1110
00000001	s_7	1111
00000000	s_8	0

（3）试求中间码对应的二元变长码字的平均码长 $\bar{L}_2$；

解析5-6 （3）平均码长 $\bar{L}_2 = \sum_{i=1}^{8} p(s_i) l_i' \approx 2.71$ 码符号/信源符号序列

例题5-6 设某无记忆二元信源，概率 $P_1 = P(1) = 0.1$，$P_2 = P(0) = 0.9$ 采用下述编码方案，首先根据 0 的长度编为 9 个中间码字，再将九个码字换成二元变长码如下表：

信源符号	中间码	二元码字
1	s_0	1000
01	s_1	1001
001	s_2	1010
0001	s_3	1011
00001	s_4	1100
000001	s_5	1101
0000001	s_6	1110
00000001	s_7	1111
00000000	s_8	0

(4) 计算比值 $\bar{L}_2/\bar{L}_1$ 并解释其意义，计算这种编码方式的编码效率。

解析5-6 (4) $\dfrac{\bar{L}_2}{\bar{L}_1} \approx 0.476$ 码符号 / 二进制符号，为平均每个二进制信源符号所需的二元码符号数，即平均码长。

$$\eta = \frac{R}{\log r} = \frac{H(S)}{\frac{\bar{L}_2}{\bar{L}_1}} = \frac{-0.9\log 0.9 - 0.1\log 0.1}{0.476} \approx 0.986$$

变长码

费诺码（Fano）

- 编码步骤：

1）将信源符号按概率从大到小的顺序排列，令 $p(s_1) \geq p(s_2) \geq \ldots \geq p(s_q)$；

2）将依次排列的信源符号按概率分成两组，使每组概率和尽可能接近或相等；

3）给每一组分配一位码元 $''0''$ 或 $''1''$（分配方式不同导致结果不唯一）；

4）将每一组再按同样的方式划分，重复步骤 2 、3 ，直至概率不再可分，结束。

例题5–7 设单符号离散信源如下，要求对信源进行二进制 *Fano* 编码。

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ 0.20 & 0.19 & 0.18 & 0.17 & 0.15 & 0.10 & 0.01 \end{bmatrix}$$

解析5–7

信源符号	符号概率	第一次分组	第二次分组	第三次分组	第四次分组	码字	码长
s_1	0.20	0	0			00	2
s_2	0.19		1	0		010	3
s_3	0.18			1		011	3
s_4	0.17	1	0			10	2
s_5	0.15		1	0		110	3
s_6	0.10			1	0	1110	4
s_7	0.01				1	1111	4

平均码长为 $\bar{L} = \sum_{i=1}^{7} p(s_i) l_i = 2.74$ 码符号/信源符号

信源熵 $H(S) = -\sum_{i=1}^{7} p(s_i) \log p(s_i) = 2.61$ *bit*/ 信源符号

码率 $R = \dfrac{H(S)}{\bar{L}} = \dfrac{2.61}{2.74} = 0.953$ *bit*/ 码符号

效率 $\eta = \dfrac{R}{\log r} = \dfrac{0.953}{\log 2} = 95.3\,\%$

香农码（Shannon）

- 编码步骤：

1）将信源符号按概率从大到小的顺序排列，令 $p(s_1) \geq p(s_2) \geq \cdots \geq p(s_q)$；

2）按下式计算第 i 个符号对应的码字码长：$-\log p(s_i) \leq l_i < -\log p(s_i) + 1$（自信息量上取整即为码长）；

3）计算第 i 个符号的累加概率 $P_i = \sum_{k=1}^{i-1} p(s_k)$（第 i 个符号的累加概率为前 $i-1$ 个符号概率之和）；

4）将累加概率 P_i 变换为二进制小数，取小数点后 l_i 位作为第 i 个符号的码字。

（或者利用公式 $P_i \times r^{l_i}$，其中 r 为码元数，l_i 为对应码长，将结果的整数部分化为二进制数）

香农编码结果唯一。

例题5-8 设单符号离散信源如下，对此信源进行二进制香农编码。

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ 0.20 & 0.19 & 0.18 & 0.17 & 0.15 & 0.10 & 0.01 \end{bmatrix}$$

解析5-8

信源符号	符号概率	累加概率	$-\log p(s_i)$	码长	$P_i \times 2^{l_i}$	码字
s_1	0.20	0	2.34	3	0	0
s_2	0.19	0.20	2.41	3	1.6	001
s_3	0.18	0.39	2.48	3	3.12	011
s_4	0.17	0.57	2.56	3	4.56	100
s_5	0.15	0.74	2.74	3	5.92	101
s_6	0.10	0.89	3.34	4	14.24	1110
s_7	0.01	0.99	6.66	7	126.72	1111110

平均码长为 $\bar{L} = \sum_{i=1}^{7} p(s_i) l_i = 3.14$ 码符号/信源符号

信源熵 $H(S) = -\sum_{i=1}^{7} p(s_i) \log p(s_i) = 2.61$ *bit*/ 信源符号

码率 $R = \dfrac{H(S)}{\bar{L}} = \dfrac{2.61}{3.14} = 0.831$ *bit*/ 码符号

效率 $\eta = \dfrac{R}{\log r} = \dfrac{0.831}{\log 2} = 83.1\,\%$

霍夫曼码（Huffman）「霍夫曼码是紧致码」

- 编码步骤：

1）将信源符号按概率从大到小的顺序排列，令 $p(s_1) \geq p(s_2) \geq \cdots \geq p(s_q)$；

2）给两个概率最小的信源符号 S_{n-1} 和 S_n 各分配一个码元 "0" 和 "1"，并将这两个信源符号合并成一个新符号，并用这两个最小的概率之和作为新符号的概率，结果得到一个包含 $(n-1)$ 个的信源符号的新信源，称为信源的第一次缩减信源，用 S_1 表示（0和1的分配有多种方法不影响解）；

3）将缩减信源 S_1 的符号仍按概率从大到小排列，重复步骤 2 得到只含 $(n-2)$ 个符号的缩减信源 S_2（如果合并后的概率与其他概率相等，将合并概率当大概率排列，此时码长方差比较小）；

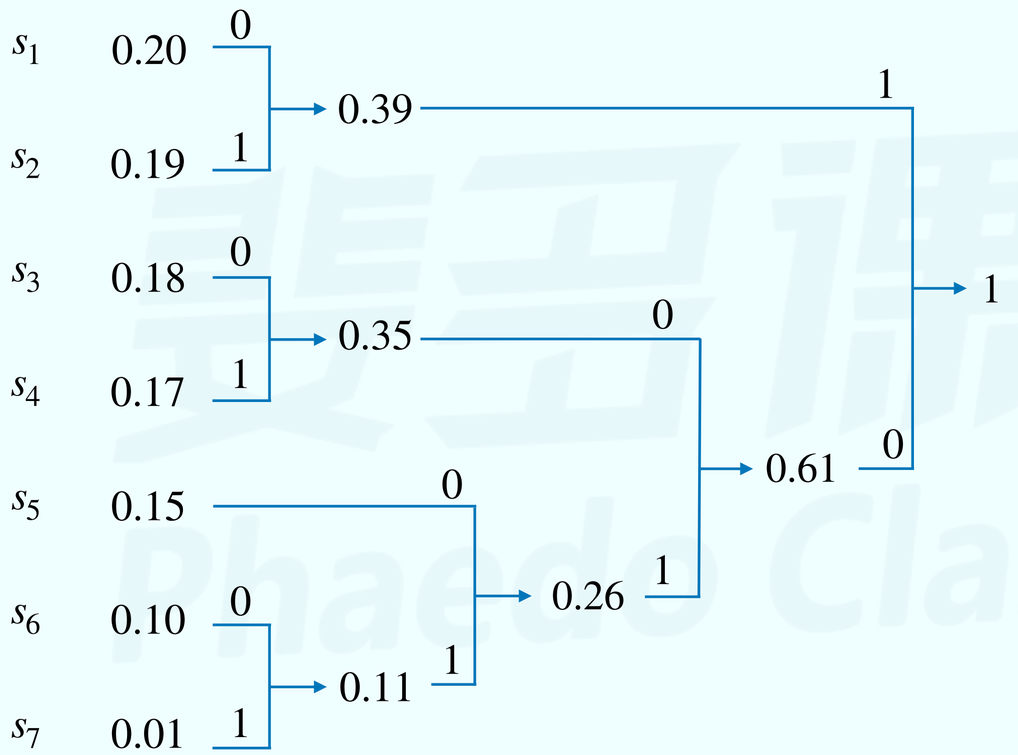
4）重复上述步骤，直至缩减信源只剩两个符号为止，此时所剩的两个符号的概率之和必为 1；

5）从最后一级缩减信源开始，依编码路径向前返回，就得到各信源符号所对应的码字。

例题5-9 设单符号离散信源如下，对此信源进行二进制霍夫曼编码。

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ 0.20 & 0.19 & 0.18 & 0.17 & 0.15 & 0.10 & 0.01 \end{bmatrix}$$

解析5-9



初始化：0.20,0.19,0.18,0.17,0.15,0.10,0.01

第一次缩减序列：0.20,0.19,0.18,0.17,0.15,0.11

第二次缩减序列：0.26,0.20,0.19,0.18,0.17

第三次缩减序列：0.35,0.26,0.20,0.19

第四次缩减序列：0.39,0.35,0.26

第五次缩减序列：0.61,0.39

第六次缩减序列：1

故对应的编码为

信源符号	s_1	s_2	s_3	s_4	s_5	s_6	s_7
码符号	10	11	000	001	010	0110	0111
码长	2	2	3	3	3	4	4

例题5-9 设单符号离散信源如下，对此信源进行二进制霍夫曼编码。

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 \\ 0.20 & 0.19 & 0.18 & 0.17 & 0.15 & 0.10 & 0.01 \end{bmatrix}$$

解析5-9

信源符号	s_1	s_2	s_3	s_4	s_5	s_6	s_7
码符号	10	11	000	001	010	0110	0111
码长	2	2	3	3	3	4	4

平均码长为 $\bar{L} = \sum_{i=1}^{7} p(s_i) l_i = 2.72$ 码符号/信源符号

信源熵 $H(S) = -\sum_{i=1}^{7} p(s_i) \log p(s_i) = 2.61\ bit/$ 信源符号

码率 $R = \dfrac{H(S)}{\bar{L}} = \dfrac{2.61}{2.72} = 0.96\ bit/$码符号

效率 $\eta = \dfrac{R}{\log r} = \dfrac{0.96}{\log 2} = 96.0\,\%$

霍夫曼编码注意事项

霍夫曼编码结果不唯一。

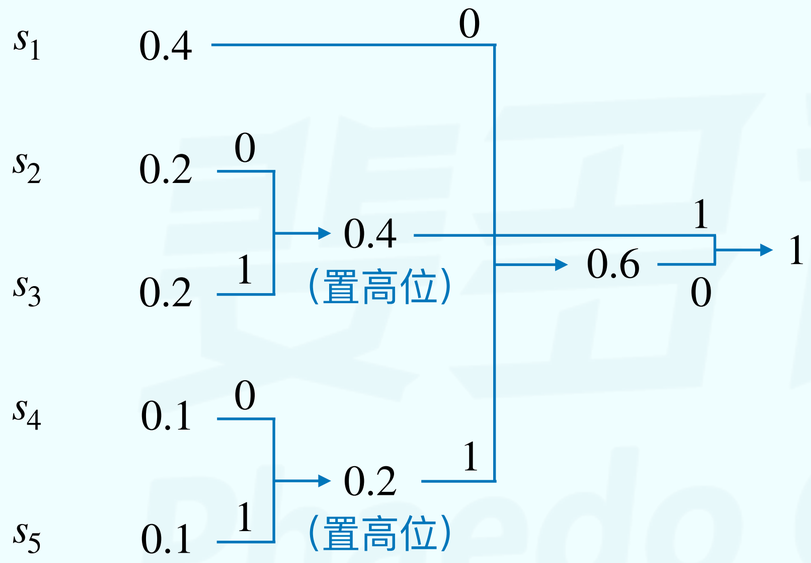
- 每次对缩减信源两个概率最小的符号分配 $''0''$ 和 $''1''$ 是任意的，因此可得到不同码字，但码长 l_i 不变，故平均码长不变，结果间无本质区别。（本课程一般默认高概率为 0，低概率为 1）

- 缩减信源时，若合并后的概率与其他概率相等，这几个概率的次序排列不同，结果就不同，编码正确。但码字不同，码长也不相同。在解题过程中，默认选择将合并后的概率排列高位，这样码长方差较小，易于实现。

例题5–10 设单符号离散信源如下，对此信源进行二进制霍夫曼编码。

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{bmatrix}$$

解析5–10



初始化：0.4,0.2,0.2,0.1,0.1

第一次缩减序列：0.4,0.2,0.2,0.2

第二次缩减序列：0.4,0.4,0.2

第三次缩减序列：0.6,0.4

第六次缩减序列：1

故对应的编码为

信源符号	s_1	s_2	s_3	s_4	s_5
码符号	00	10	11	010	011

r 进制霍夫曼编码

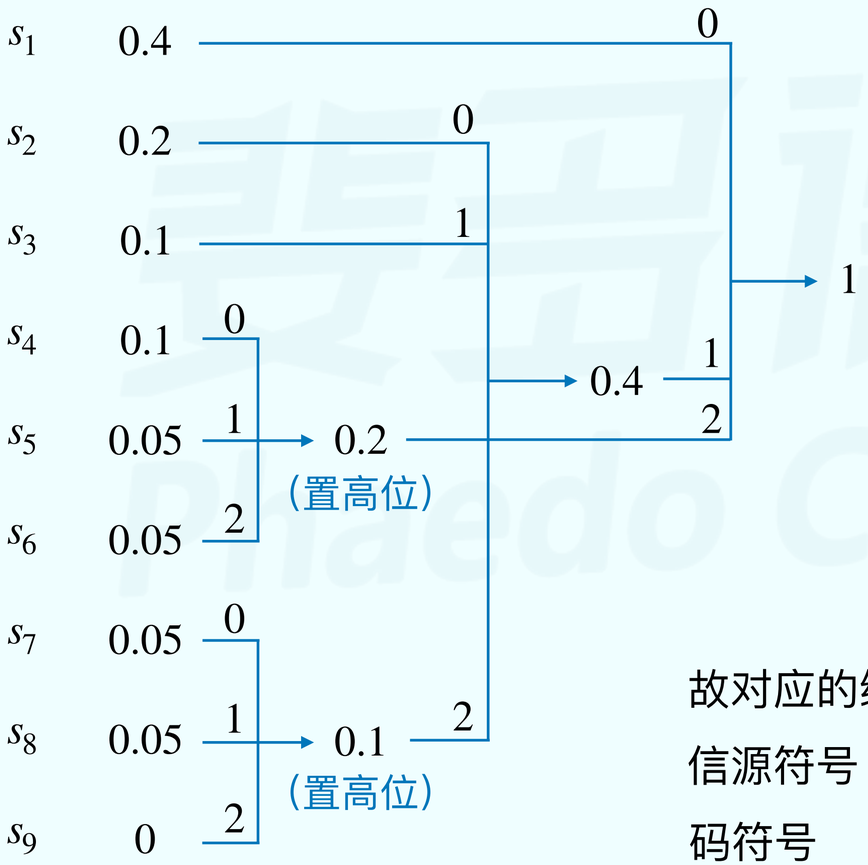
- 构造 r 进制霍夫曼码时，每一次缩减都将概率最小的 r 个符号合并。
- 为充分利用短码使平均码长最短，必须使最后一次缩减信源有 r 个符号，

  信源符号数 $q = r + k \cdot (r-1)$ 可进行完全缩减。

  不符合 $q = r + k \cdot (r-1)$ 时，补充一些概率为 0 的符号，使符号总数满足 $r + k \cdot (r-1)$ 。

例题5–11 信源空间为 $\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ 0.4 & 0.2 & 0.1 & 0.1 & 0.05 & 0.05 & 0.05 & 0.05 \end{bmatrix}$，码符号为 $X = \{0,1,2\}$，试构造一种三元紧致码。

解析5–11 构造三元紧致码，一共有8个消息符号，$r + k \cdot (r-1) = 3 + 2k$ 可以等于9但不能等于8，需补一个概率为0的符号；



初始化：0.4,0.2,0.1,0.1,0.05,0.05,0.05,0.05,0

第一次缩减序列：0.4,0.2,0.1,0.1,0.1,0.05,0.05

第二次缩减序列：0.4,0.2,0.2,0.1,0.1

第三次缩减序列：0.4,0.4,0.2

第四次缩减序列：1

故对应的编码为

信源符号	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9
码符号	0	10	11	20	21	22	120	121	122

例题5-12 现有一副离散量化后的图像，图像的灰度量化分成8级，见下表。表中数字为相应像素上的灰度级。

1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	3	3	3
3	3	3	3	3	3	3	4	4	4
4	4	4	4	4	4	4	5	5	5
5	5	5	5	6	6	6	6	6	6
7	7	7	7	7	8	8	8	8	8

另有一无噪无损二元信道，单位时间（秒）内传输100个二元符号。

（1）现将图像通过给定的信道传输，不考虑图像的任何统计特性，并采用二元等长码，需要多久才能传送完这幅图像?

（2）若考虑图像的统计特性（不考虑图像像素之间的依赖性）求此图像的信源熵$H(S)$，并对每个灰度等级进行最佳二元编码。问平均每个像素需用多少个二元码符号来表示？此时需要多少时间传输完此图像?

（3）从理论上简要说明这幅图像还可以压缩，而且平均每个像素所需的二元码符号数可小于$H(S)$。

例题5-12 现有一副离散量化后的图像，图像的灰度量化分成8级，见下表。表中数字为相应像素上的灰度级。

1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	3	3	3
3	3	3	3	3	3	3	4	4	4
4	4	4	4	4	4	4	5	5	5
5	5	5	5	6	6	6	6	6	6
7	7	7	7	7	8	8	8	8	8

另有一无噪无损二元信道，单位时间（秒）内传输100个二元符号。

（1）现将图像通过给定的信道传输，不考虑图像的任何统计特性，并采用二元等长码，需要多久才能传送完这幅图像?

解析5-12 （1）灰度级为8级，可用3位二进制数表示，100点共有300位二进制数。

若不考虑统计特性，需用$\frac{300\ bit}{100\ bit/s}=3\ s$时间完成传送。

例题5-12 现有一副离散量化后的图像，图像的灰度量化分成8级，见下表。表中数字为相应像素上的灰度级。

1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
2	2	2	2	2	2	2	3	3	3
3	3	3	3	3	3	3	4	4	4
4	4	4	4	4	4	4	5	5	5
5	5	5	5	6	6	6	6	6	6
7	7	7	7	7	8	8	8	8	8

另有一无噪无损二元信道，单位时间（秒）内传输100个二元符号。

（2）若考虑图像的统计特性（不考虑图像像素之间的依赖性）求此图像的信源熵$H(S)$，并对每个灰度等级进行最佳二元编码。问平均每个像素需用多少个二元码符号来表示？此时需要多少时间传输完此图像？

解析5-12 （2）由表格可得各灰度等级的概率分布为

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 0.4 & 0.17 & 0.10 & 0.10 & 0.07 & 0.06 & 0.05 & 0.05 \end{bmatrix}$$

例题5–12 现有一副离散量化后的图像，图像的灰度量化分成8级，见下表。表中数字为相应像素上的灰度级。

表格（略）

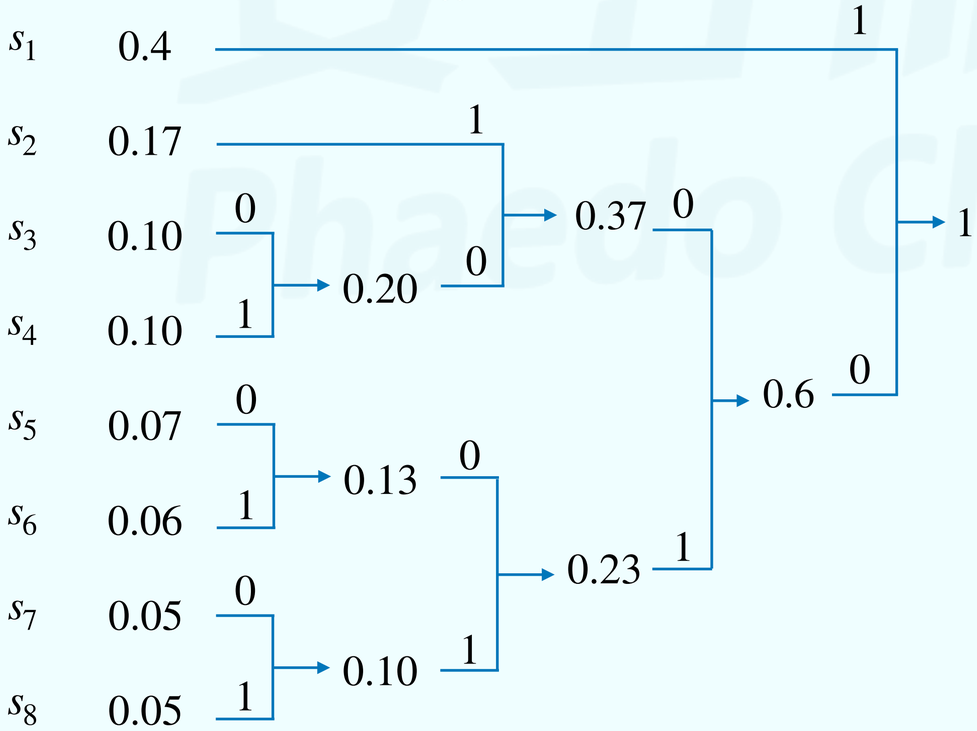
另有一无噪无损二元信道，单位时间（秒）内传输100个二元符号。

(2) 若考虑图像的统计特性（不考虑图像像素之间的依赖性）求此图像的信源熵 $H(S)$，并对每个灰度等级进行最佳二元编码。问平均每个像素需用多少个二元码符号来表示？此时需要多少时间传输完此图像？

解析5–12 (2)

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 0.4 & 0.17 & 0.10 & 0.10 & 0.07 & 0.06 & 0.05 & 0.05 \end{bmatrix}$$

对其进行最佳编码（*Huffman* 编码）



编码结果为

信源符号	1	2	3	4	5	6	7	8
码符号	1	001	0000	0001	0100	0101	0110	0111

平均码长为 $\bar{L} = \sum_{i=1}^{8} p(i)l_i$

$$= 0.4\times1 + 0.17\times3 + 0.10\times4 + 0.10\times4 + 0.07\times4 + 0.06\times4 + 0.05\times4 + 0.05\times4$$

$$= 2.63 \text{码符号/信源符号}$$

100位共需用263个码符号，共需 $\frac{263}{100} = 2.63\ s$ 传输时间。

例题5–12 现有一副离散量化后的图像，图像的灰度量化分成 8 级，见下表。表中数字为相应像素上的灰度级。

表格（略）

另有一无噪无损二元信道，单位时间（秒）内传输 100 元符号。

（3）从理论上简要说明这幅图像还可以压缩，而且平均每个像素所需的二元码符号数可小于 $H(S)$ 。

解析5–12 （3）若考虑图像之间的依赖关系，信源熵需考察有记忆信源的极限熵 H_∞ ，则 $H_\infty < H(S)$ ，根据香农第一定理，平均码长的压缩极限为信源熵 $H(S)$ ，若进一步压缩，则可使平均码长逼近 H_∞ ，且小于 $H(S)$ 。

例题5-13 设有一个信源发出符号A和B，它们是相互独立的发出，并已知$P(A)=0.25$，$P(B)=0.75$。

（1）计算该信源的熵；

（2）若用二进制代码组传输消息，0代表A，1代表B，试求码元符号的概率$P(0)$和$P(1)$；

（3）若该信源发出二重扩展消息，采用费诺编码，求其平均传输速率以及码元符号的概率$P(0)$和$P(1)$；

解析5-13 （1）该信源的熵为 $H(S)=H(\frac{3}{4},\frac{1}{4})=-\frac{3}{4}\log\frac{3}{4}-\frac{1}{4}\log\frac{1}{4}\approx 0.811 bit/symbol$

（2）明显，单符号二进制代码的码元概率为$P(0)=0.25$，$P(1)=0.75$

（3）信源发出二重扩展消息，则码字一共有4种情况：00，01，10，11

根据信源符号是相互独立的发出，我们可以确定四个码字的概率如右表

s_i	$p(s_i)$
$AA:00$	$\frac{1}{16}$
$AB:01$	$\frac{3}{16}$
$BA:10$	$\frac{3}{16}$
$BB:11$	$\frac{9}{16}$

进行费诺编码（过程略，请读者自行按照刚才的编码方法练习）

可求得二重扩展消息的编码结果为

s_i	$p(s_i)$	W_i
$AA:00$	$\frac{1}{16}$	111
$AB:01$	$\frac{3}{16}$	110
$BA:10$	$\frac{3}{16}$	10
$BB:11$	$\frac{9}{16}$	0

例题5-13 设有一个信源发出符号A和B，它们是相互独立的发出，并已知$P(A)=0.25$，$P(B)=0.75$。

（1）计算该信源的熵；

（2）若用二进制代码组传输消息，0代表A，1代表B，试求码元符号的概率$P(0)$和$P(1)$；

（3）若该信源发出二重扩展消息，采用费诺编码，求其平均传输速率以及码元符号的概率$P(0)$和$P(1)$；

解析5-13 (3) 进行费诺编码（过程略，请读者自行按照刚才的编码方法练习）

可求得二重扩展消息的编码结果为

s_i	$p(s_i)$	W_i
$AA:00$	$\frac{1}{16}$	111
$AB:01$	$\frac{3}{16}$	110
$BA:10$	$\frac{3}{16}$	10
$BB:11$	$\frac{9}{16}$	0

因此单符号平均码长为 $\bar{l}=\frac{1}{2}\sum_{i=1}^{4}p(s_i)l_i=\frac{1}{2}(\frac{1}{16}\times 3+\frac{3}{16}\times 3+\frac{3}{16}\times 2+\frac{9}{16}\times 1)=0.844$ 码符号/信源符号

故平均传输速率为 $R=\frac{H(s)}{\frac{L_N}{N}}=\frac{H(s)}{\bar{l}}=\frac{0.811}{0.844}=0.961bit/$码符号

例题5-13 设有一个信源发出符号A和B，它们是相互独立的发出，并已知$P(A)=0.25$，$P(B)=0.75$。

（1）计算该信源的熵；

（2）若用二进制代码组传输消息，0代表A，1代表B，试求码元符号的概率$P(0)$和$P(1)$；

（3）若该信源发出二重扩展消息，采用费诺编码，求其平均传输速率以及码元符号的概率$P(0)$和$P(1)$；

解析5-13 (3) 进行费诺编码（过程略，请读者自行按照刚才的编码方法练习）

可求得二重扩展消息的编码结果为

s_i	$p(s_i)$	W_i
$AA:00$	$\frac{1}{16}$	111
$AB:01$	$\frac{3}{16}$	110
$BA:10$	$\frac{3}{16}$	10
$BB:11$	$\frac{9}{16}$	0

求解码元符号的概率$P(0)$和$P(1)$：

$$P(0)=\frac{\sum_{i=1}^{4}[p(s_1)\times 0+p(s_2)\times 1+p(s_3)\times 1+p(s_4)\times 1}{2\bar{l}}=\frac{15}{27}$$

$$P(1)=\frac{\sum_{i=1}^{4}[p(s_1)\times 3+p(s_2)\times 2+p(s_3)\times 1+p(s_4)\times 0}{2\bar{l}}=\frac{12}{27}$$

例题5-13 设有一个信源发出符号A和B，它们是相互独立的发出，并已知$P(A)=0.25$，$P(B)=0.75$。

（1）计算该信源的熵；

（2）若用二进制代码组传输消息，0代表A，1代表B，试求码元符号的概率$P(0)$和$P(1)$；

（3）若该信源发出二重扩展消息，采用费诺编码，求其平均传输速率以及码元符号的概率$P(0)$和$P(1)$；

解析5-13 **补充：求解扩展信源编码二元码元符号的概率$P(0)$和$P(1)$的方法**

$$P(0)=\frac{\sum_{i=1}^{s^N}\text{扩展信源消息概率}\times\text{对应码字中0的个数}}{\text{扩展消息平均码长}}$$

$$P(1)=\frac{\sum_{i=1}^{s^N}\text{扩展信源消息概率}\times\text{对应码字中1的个数}}{\text{扩展消息平均码长}}$$

三种编码方式小结

名称	结果唯一性	平均码长	编码效率	适用范围
费诺码	不唯一	居中	居中	分组概率相等或接近的码
香农码	唯一	长	低	思想可扩展至算术编码
霍夫曼码	不唯一	短	高	无特殊要求，且综合性能最优 且为紧致码，最佳码