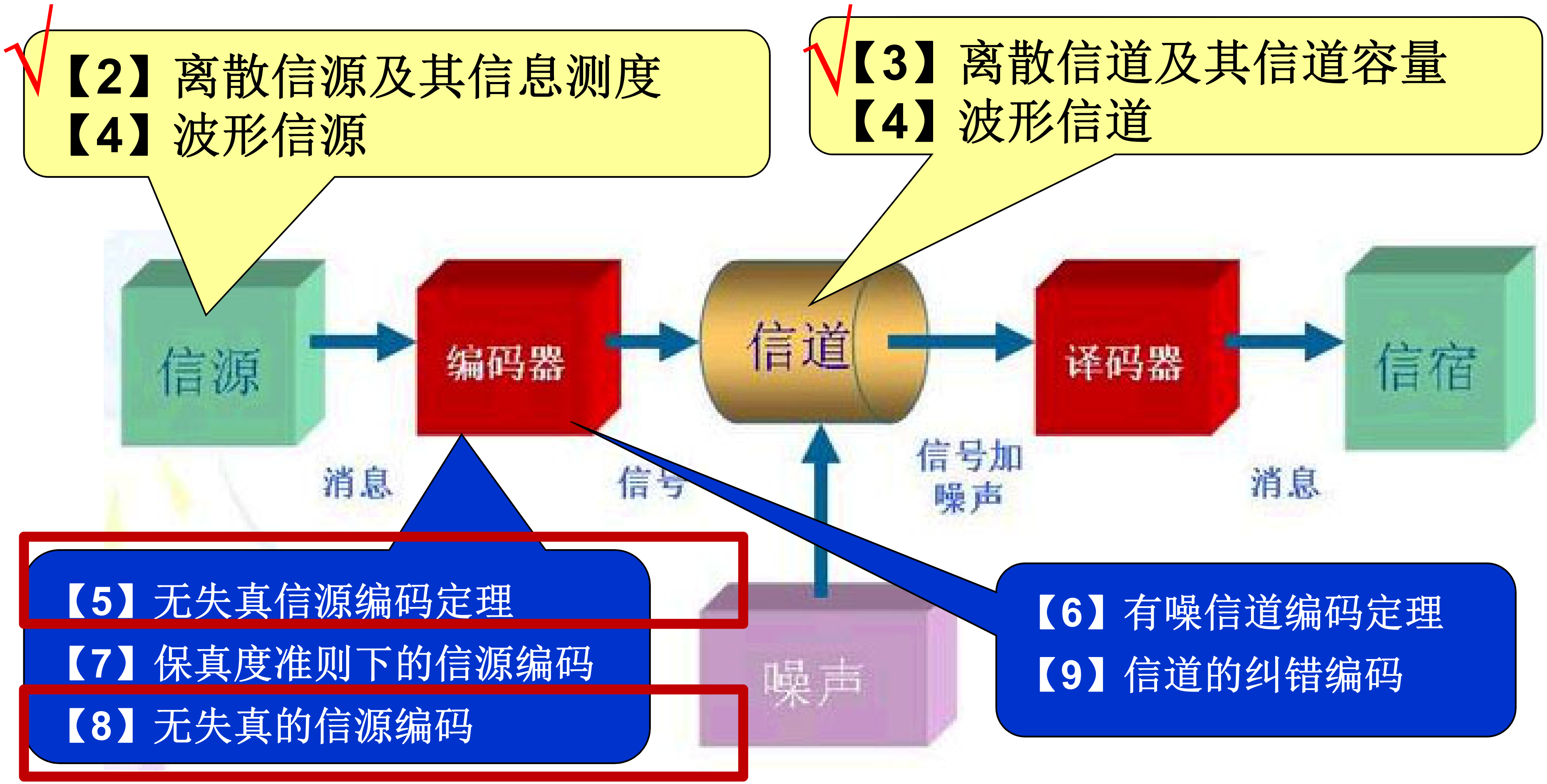
第五+八章 无失真信源编码

2022/5/9

信息论的旅程

- 从本章开始，分析信源特性，通过对信源符号进行编码以提高信息传输的有效性。



第五、八章的研究内容

- 如何实现无损信道传输中的信源信道匹配？
  - 将原信源编码转换为新的信源 - 概率空间 ★
- 无失真信源编码的主要研究内容
  - 第五章：无失真信源编码定理
    - 信源编码的目的 – 有效地表示！
    - 无失真信源编码基本理论
  - 第八章：无失真信源编码
    - 常用算法

主要内容

- 编码器
  - 编码器概论
  - 信源编码器与基本术语
- 分组码
  - 唯一可译性、即时码
- 定长码和定长编码定理
- 变长码
  - 变长无失真信源编码定理 – 香农第一定理
  - 几种变长编码方法（第八章）

主要内容

- 信源编码器
  - 编码器概论
  - 基本术语
  - *N*次扩展码
- 分组码
- 定长码和定长编码定理（等长＝定长）
- 变长码

1.1、概述 – 编码器概论

- 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来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。

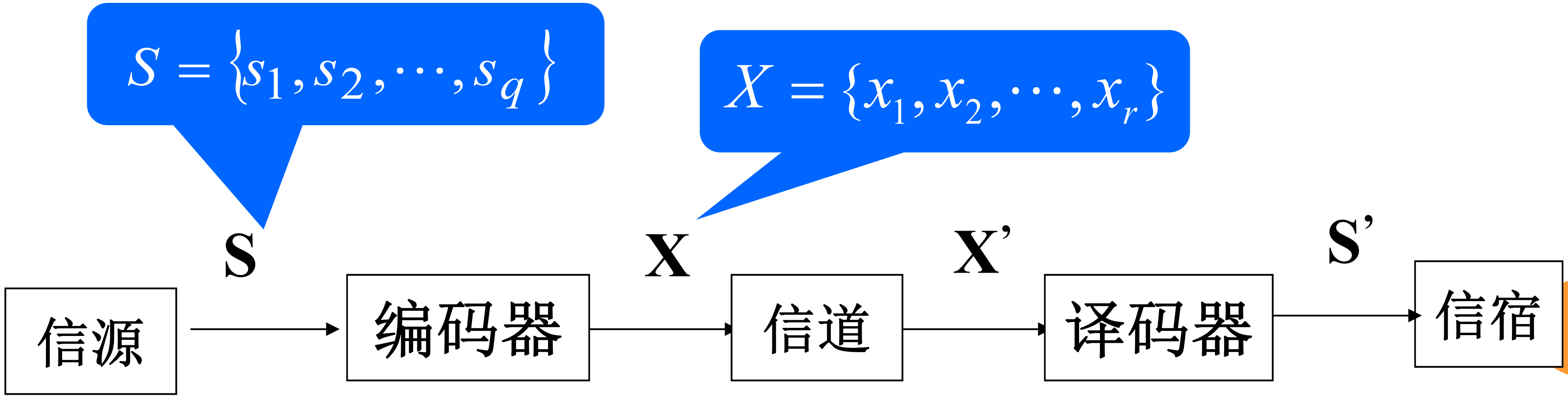
- 信源编码目的 – 提高通信有效性
  - 通常通过压缩信源的冗余度来实现。
  - 采用的一般方法是压缩每个信源符号的平均比特数。

1.1、概述 － 编码器概论（续）

- 信源编码理论是信息论的一个重要分支，其理论基础：无失真信源编码定理；限失真信源编码定理。
- 本章主要介绍无失真信源编码，它实质上是一种统计匹配编码，根据信源的不同概率分布而选用与之相匹配的码。
- 信源的统计剩余度主要决定于以下两个因素
  - 无记忆信源中，符号概率分布的非均匀性。
  - 有记忆信源中，符号间的相关性及符号概率分布的非均匀性。

1.1、概述 － 信源编码器模型

- 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。



信源编码器模型

1.1、概述 － 信源编码器模型（续）

$$S=\{s_1,s_2,\cdots,s_q\} \longrightarrow \boxed{\text{编码器}} \longrightarrow C=\{W_1,W_2,\cdots,W_q\}$$

$$X:\{x_1,x_2,\ldots,x_r\}$$

$$W_i=\{x_{i_1}x_{i_2}\cdots x_{i_{l_i}}\}$$

- 将信源符号集中的符号 s_i（或者长为*N*的信源符号序列）映射成由码符号 x_i 组成的长度为　　的一一对应的码符号序列 W_i 。

1.2、概述 － 基本术语★★

信源符号集

$$S=\{s_1,s_2,\cdots,s_q\}$$

编码器

代码组C / 码C

$$C=\{W_1,W_2,\cdots,W_q\}$$

码符号集

$$X:\{x_1,x_2,\ldots,x_r\}$$

码元 / 码符号

r元码

码字

$$W_i=\{x_{i_1}x_{i_2}\cdots x_{i_{l_i}}\}$$

码长

平均码长

$$\boxed{\overline{L}=\sum_i p(s_i)l_i}$$

定长码、变长码；奇异码、非奇异码

1.2、概述 － 基本术语 － 例题5.1

例题：设有二元信道的信源编码器，其概率空间如右：

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 \\ p(s_1) & p(s_2) & p(s_3) & p(s_4) \end{bmatrix}$$

信源符号	出现概率	码字	码1	码2	码3
S_1	$p(S_1)$	W_1	00	0	0
S_2	$p(S_2)$	W_2	01	01	11
S_3	$p(S_3)$	W_3	10	001	00
S_4	$p(S_4)$	W_4	11	111	11

定长码： 码1　　　　变长码：码2、码3

非奇异码： 码1、码2　　　奇异码：码3

1.3、概述 – *N*次扩展码

- 实际接收：*N*次无记忆扩展信源 --〉*N*次扩展码

$$S=\{s_1,s_2,\cdots,s_q\} \longleftrightarrow C=\{W_1,W_2,\cdots,W_q\}$$

$$s_i \longleftrightarrow W_i$$

$$S^N=\{\alpha_1,\alpha_2,\cdots,\alpha_{q^N}\} \longleftrightarrow C^N=\{\mathbf{W}_1,\mathbf{W}_2,\cdots,\mathbf{W}_{q^N}\}$$

$$\alpha_j=s_{j_1}s_{j_2}\cdots s_{j_N} \longleftrightarrow \mathbf{W}_j=W_{j_1}W_{j_2}\cdots W_{j_N}$$

$$j=1,2,\cdots,q^N$$

$$j_1,j_2,\cdots,j_N=1,2,\cdots,q$$

1.3、概述 – N次扩展码

- 例题5.1 - 续

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

二次扩展信源符号 $\alpha_j (j=1,2,\ldots,16)$	二次扩展码码字 $\mathbf{W}_j \left(j=1,2,\cdots,16\right)$
$\alpha_1 = s_1 s_1$	$\mathbf{W}_1 = W_1 W_1 = 00$
$\alpha_2 = s_1 s_2$	$\mathbf{W}_2 = W_1 W_2 = 001$
$\alpha_3 = s_1 s_3$	$\mathbf{W}_3 = W_1 W_3 = 0001$
$\alpha_{16} = s_4 s_4$	$\mathbf{W}_{16} = W_4 W_4 = 111111$

基本术语复习

信源符号集

$$S=\{s_1,s_2,\cdots,s_q\}$$

编码器

代码组C / 码C

$$C=\{W_1,W_2,\cdots,W_q\}$$

码符号集

$$X:\{x_1,x_2,\ldots,x_r\}$$

码元 / 码符号

r元码

码字

$$W_i=\{x_{i_1}x_{i_2}\cdots x_{i_{l_i}}\}$$

码长

平均码长

$$\boxed{\overline{L}=\sum_i p(s_i)l_i}$$

定长码、变长码；奇异码、非奇异码

主要内容

- 信源编码器
- 分组码
- 定长码和定
- 变长码

定义

唯一可译性

即时码的判别与构造

2.1、分组码

- 分组码：将信源符号集中的每个信源符号映射成一个固定的码字 。
- 特点：集间符号为一对一或多对一。

例题5.2：信源S有四种不同的符号S_1，S_2，S_3，S_4，其概率分别为$p(S_1)$、$p(S_2)$、$p(S_3)$、$p(S_4)$；码符号集为X：{0、1}，可得到如下五种码。

信源符号	码1	码2	码3	码4	码5
S_1	0	0	00	1	1
S_2	11	10	01	10	01
S_3	00	00	10	100	001
S_4	11	01	11	1000	0001

2.2、分组码 － 唯一可译性

- **唯一可译码**：任意一串有限长的码符号序列只能被<u>唯一地译为对应的信源符号序列</u>，则此码为唯一可译码。<u>没有二义性</u>

- 唯一可译码的**充要条件**：编码的任意次扩展均为非奇异码。
  - 码字与信源符号一一对应
  - 不同的信源符号序列对应不同的码字序列

2.2、分组码 – 唯一可译性 – 分析

译码

- 奇异码一定不是唯一可译码

11 → s_2

11 → s_4

- 非奇异码不一定是唯一可译码

0 1 0 0 0 0 1 0

0 | 1 0 | 0 0 | 0 1 | 0 $\xrightarrow{\text{译码}}$ $s_1 s_2 s_3 s_4 s_1$

0 1 | 0 0 | 0 0 | 1 0 $\xrightarrow{\text{译码}}$ $s_4 s_3 s_3 s_2$

- 等长非奇异码一定是唯一可译码

0 0 0 1 1 0 1 1 → $s_1 s_2 s_3 s_4$

符号	码1	码2	码3	码4	码5
S_1	0	0	00	1	1
S_2	11	10	01	10	01
S_3	00	00	10	100	001
S_4	11	01	11	1000	0001

变长非奇异码 ??

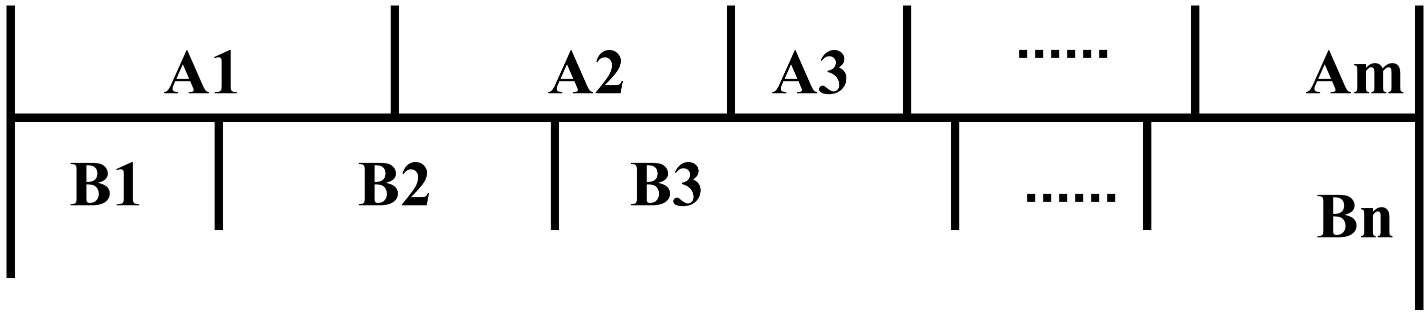
2.2、变长码 – 唯一可译码判别准则

*N*次扩展码的奇异性检查很难！

0 1 0 0 0 0 1 0

0 | 1 0 | 0 0 | 0 1 | 0 $\xrightarrow{\text{译码}}$ $s_1 s_2 s_3 s_4 s_1$

0 1 | 0 0 | 0 0 | 1 0 $\rightarrow$ $s_4 s_3 s_3 s_2$



- 码符号序列译码二义性的存在条件
  - 存在前缀码
  - 排除前缀的剩余码可能和其他码再次构成前缀码
  - 码符号序列的尾部一定是一个码字

2.2、变长码 － 唯一可译码判别准则（续）

步骤： 1、初始化：$S_0 = C$ ★★

2、构造 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，则将后缀作为S_1 的元素。

3、构造 $S_n(n>1)$：将 S_0与 S_{n-1} 比较。

（1）如果 S_0中有码字是S_{n-1} 中元素的前缀，则将相应的后缀放入 S_n 中；

（2）同样 S_{n-1}中若有元素是 S_0 中码字的前缀，也将相应的后缀放入 S_n 中。

4、检查 S_n ：

（1）若S_n 是空集，则码 C 是唯一可译码，结束；

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

（3）如果上述两个条件都不满足，则返回步骤3。

2.2、变长码 – 唯一可译码判别准则 – 例题5.3

S_0	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8
$\boxed{a}$	bb	cde	de	b	ad	d	eb	空
c					$bcde$			
$\boxed{abb}$								
bad								
deb								
$bbcde$								

S_8为空集，该码是唯一可译码

2.2、变长码 – 唯一可译码判别准则 – 例题5.4

S_0	S_1	S_2	S_3	S_4	S_5
a	d	eb	de	b	$\boxed{ad}$
c	bb	cde			$bcde$
$\boxed{ad}$					
abb					
bad					
deb					
$bbcde$					

结论：S_5中包含S_0中元素，故该变长码不是唯一可译码

2.3、分组码 － 即时码

- 码4和码5，均为唯一可译码。但是：
  - 码4：不能即时译码 1 1 0 1 0 0 1 0 0 0
  - 码5：一个码字完全出现后，即可译码

1	0 1	0 0 1	0 0 0 1

- 即时码★：唯一可译码在接收到一个完整的码字时，<u>无需参考</u>后续的码符号就能立即译码。

符号	码1	码2	码3	码4	码5
S_1	0	0	00	1	1
S_2	11	10	01	10	01
S_3	00	00	10	100	001
S_4	11	01	11	1000	0001

2.3、分组码 – 即时码的判断

- 唯一可译码成为即时码的充要条件：★★
  设 W_i 是C中任一码字，则要求其它码字 W_i，都不是码字 W_i 的前缀。

  证明

充分性：若不存在前缀，则收到完整码字自然可以实现即时译码

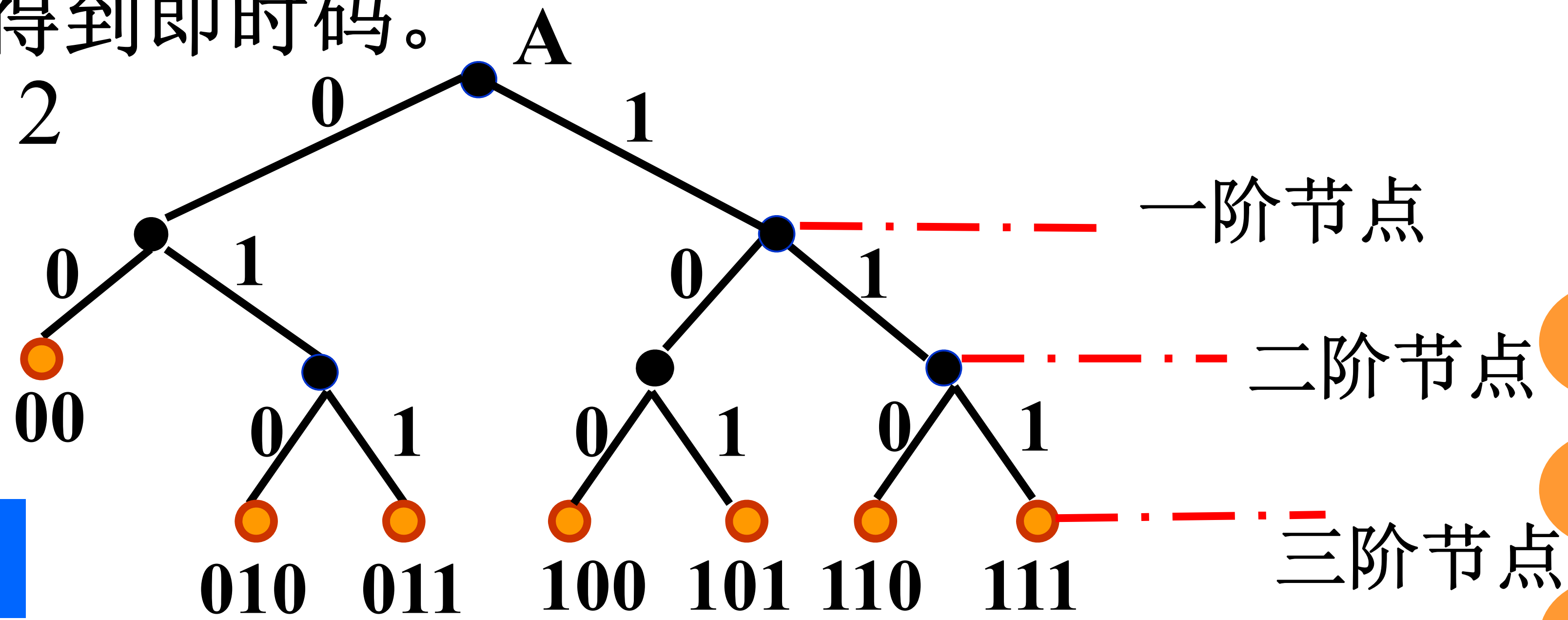
必要性：反证法->*满足即时码，但存在前缀码字*

若有前缀，则译码必须参考后续码字，不可能实现即时译码，假设错误，原命题成立！

2.3、即时码的构造方法 － 对于前缀的要求

- 树图法构造即时码（r 进制树）
  - 树中每个中间节点都伸出1至 r 个树枝，将所有的码字都安排在终端节点上就可以得到即时码。

$r = 2$



整树

编码过程满足对于前缀的要求

2.3、即时码的构造方法 – 即时码的判断

- 利用构造方法，来判断给定码字是否为即时码：{0，10，110，111}

2.3、各类码之间的相互关系 – 小结

- 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

符号	码1	码2	码3	码4	码5
S_1	0	0	00	1	1
S_2	11	10	01	10	01
S_3	00	00	10	100	001
S_4	11	01	11	1000	0001

主要内容

- 信源编码器
- 分组码
- 定长码及编码定理
- 变长码

1.唯一可译定长码的存在条件

2.定长编码定理

3.1、定长码 － 唯一可译定长码存在的条件

- 定长码：非奇异码一定是唯一可译码
  - 非奇异码：信源符号与码字一一对应

- 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 $S=\{s_1,s_2,\cdots,s_q\}$
  - 码符号集中共有*r*种码元 $X:\{x_1,x_2,\ldots,x_r\}$
  - 定长码码长为
  - 若满足非奇异性，则 $\boxed{q \le r^l}$

码元数目

该条件是必要条件，而不是充分条件。

3.1、唯一可译定长码存在的条件 – 例题5.5

- 英文字母表中，每一字母用定长编码转换成二进制表示，码字的最短长度是多少？

解：

信源符号数 $q=26$

码符号数 $r=2$

$$r^l \ge q \quad \Rightarrow \quad l \ge \frac{\log q}{\log r} = \frac{\log 26}{\log 2} = 4.7$$

$$\therefore \quad l_{\min} = 5$$

3.1、扩展信源－唯一可译定长码存在的条件

- 如果对*N*次扩展信源 s^N 进行定长编码，要满足非奇异性，须满足以下条件：

$$q^N \le r^l$$

定长非奇异码的存在条件

其中：$S^N = \{\alpha_1, \alpha_2, \cdots, \alpha_{q^N}\}$ $S = \{s_1, s_2, \cdots, s_q\}$

$$\alpha_j = s_{j_1} s_{j_2} \cdots s_{j_N}$$

$$s_{j_k} \in \{s_1, s_2, \ldots, s_q\}$$

3.2、定长码 – 定长信源编码定理 – 前言

- 理论意义远大于实际意义！★★

1. 为什么要考虑扩展信源的定长编码？
   - 扩展信源的实际应用
   - 不同符号序列的出现概率差异较大！
2. 如何有效地进一步压缩编码的平均码长？
   - 减少参与编码的信源符号数目 $q^N \le r^l$
3. 是否会带来译码错误？错误率是多少？

   什么是错误率

   - 求解具体参数
   1. 参与编码的信源符号集合的特点
   2. 未参与编码的信源符号集合的特点
   3. 对应的误差和编码效率是多少？

3.2、定长码 – 定长信源编码定理（定理5.3）

- 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 定长编码，则对于任意$\varepsilon>0$，只要满足 $\boxed{\dfrac{l}{N} \geq \dfrac{H(S)+\varepsilon}{\log r}}$

则当N足够大时，可实现几乎无失真编码，即译码错误概率 P_E为任意小；

反之，如果 $\boxed{\dfrac{l}{N} \leq \dfrac{H(S)-2\varepsilon}{\log r}}$

则不可能实现无失真编码，当N足够大时，译码错误概率 P_E为1。

返回

3.2、定长信源编码定理 － 证明思路

信源S的N次扩展信源的定长编码存在唯一可译码的必要条件

$$\boxed{q^N \le r^l \Rightarrow N\log q \le l\log r \Rightarrow \frac{l}{N} \ge \frac{\log q}{\log r}}$$

简单信源S

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

N次扩展信源

数目剧增

$$\begin{bmatrix} S^N \\ P \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_{q^N} \\ p(\alpha_1) & p(\alpha_2) & \cdots & p(\alpha_{q^N}) \end{bmatrix}$$

$\frac{l}{N}\log r$ $H(S)$ $\boxed{\frac{l}{N} \ge \frac{H(S)+\varepsilon}{\log r}}$ $\boxed{\frac{l}{N} \le \frac{H(S)-2\varepsilon}{\log r}}$

想法：将信源序列分为两部分（经常出现的、不经常出现的），只对经常出现的序列进行编码。

关于“序列中平均每个符号的自信息”与“信源熵”之间的关系？

设有一离散无记忆信源，其概率空间为

$$\begin{pmatrix} X \\ P \end{pmatrix} = \begin{pmatrix} x_1 = 0 & x_2 = 1 & x_3 = 2 & x_4 = 3 \\ 3/8 & 1/4 & 1/4 & 1/8 \end{pmatrix}$$

该信源发出的信息符号序列为（202 120 130 213 001 203 210 110 321 010 021 032 011 223 210），求：

（1） 此信息的自信息量是多少？

（2） 在此信息中平均每个符号携带的信息量是多少？

$$I(\alpha) = I(0) * \text{‘0’的个数} + I(1) * N(1) + I(2) * N(2) + I(3) * N(3)$$

此消息平均每个符号的信息量：

$$I(\alpha)/\text{符号总数} = [I(0) * N(0) + I(1) * N(1) + I(2) * N(2) + I(3) * N(3)] / \text{符号数}$$

$$\xrightarrow{\text{当N趋于无穷时}} I(0) * p(0) + I(1) * p(1) + I(2) * p(2) + I(3) * p(3) = H(S)$$

3.2、定长信源编码定理 －引理

- 渐进等分割性：

  若$S_1S_2..S_N$随机序列中$S_i(i=1,2,\ldots,N)$相互统计独立

  并且服从同一概率分布$P(S)$，又$\alpha_i=(s_{i1}s_{i2}\ldots s_{iN})\in S_1S_2..S_N$，则

  $$-\frac{1}{N}\log P(\alpha_i)=-\frac{1}{N}\log P(s_{i1}s_{i2}\ldots s_{iN})\text{以概率收敛于}H(S)$$

  $$(i=1,2,\ldots,q^N \quad i_1,i_2,\ldots i_N=1,2,\ldots,q)$$

- 何谓以概率收敛?
  - 随着N的增大，每个符号所含信息量趋于/等于$H(S)$的概率越来越大

    即对于任意小的$\varepsilon>0$，

    $$\lim_{N\to\infty}P\left\{\left|\frac{I(\alpha_i)}{N}-H(S)\right|<\varepsilon\right\}=1$$

3.2、定长信源编码定理 －引理（续）

- 因此可以将扩展信源中的信源序列分为两个互补的子集

$$G_{\varepsilon}=\left\{\alpha_{j}:\left|\frac{I(\alpha_{j})}{N}-H(S)\right|<\varepsilon\right\}$$

$$\overline{G_{\varepsilon}}=\left\{\alpha_{j}:\left|\frac{I(\alpha_{j})}{N}-H(S)\right|\geq\varepsilon\right\}$$

$$\boxed{p(G_{\varepsilon})+p\left(\overline{G_{\varepsilon}}\right)=1}$$

ε 典型序列集是那些(序列中符号)平均“自信息”以任意小地接近信息熵的N长序列的集合

3.2、定长信源编码定理 －引理（续）

（对已知信源，求解其序列的均值/方差）

$$I(s_i) = -\log p(s_i)$$

$$D[I(s_i)] = E[I(s_i) - H(S)]^2 = E[I^2(s_i)] - H^2(S)$$

$$= \sum_{i=1}^{q} p(s_i)[\log p(s_i)]^2 - H^2(S) < \infty$$

$$I(\alpha_j) = -\log p(\alpha_j) = -\log\left(\prod_{k=1}^{N} p(s_{j_k})\right) = \sum_{k=1}^{N} I(s_{j_k})$$

$$E[I(\alpha_j)] = H(S^N) = NH(S)$$

$$D[I(\alpha_j)] = ND[I[s_i]] = N\{E[I^2(s_i)] - [H(S)]^2\}$$

$$= N\left\{\sum_{i=1}^{q} p(s_i)[\log p(s_i)]^2 - \left[-\sum_{i=1}^{q} p(s_i)\log p(s_i)\right]^2\right\}$$

q为有限值时，对应方差有限

3.2、定长信源编码定理 －引理（续）

- ***Chebyshev*** **(契比雪夫)不等式**

$$\boxed{P\{|X-\mu|\ge\varepsilon\}\le\frac{\sigma^2}{\varepsilon^2}}$$

$$p\left\{\ \left|I(\alpha_j)-NH(S)\right|\ \ge N\varepsilon\right\}\le\frac{D[I(\alpha_j)]}{(N\varepsilon)^2}$$

$$\Downarrow$$

$$p\left\{\left|\frac{I(\alpha_j)}{N}-H(S)\right|\ge\varepsilon\right\}\le\frac{D[I(s_i)]}{N\varepsilon^2}$$

$$\boxed{\delta(N,\varepsilon)=\frac{D[I(s_i)]}{N\varepsilon^2}}$$

$$\lim_{N\to\infty}\delta(N,\varepsilon)=\lim_{N\to\infty}\frac{D[I(s_i)]}{N\varepsilon^2}=0$$

自信息 $I(\alpha_j)$的均值 $\dfrac{I(\alpha_j)}{N}$ 以概率收敛于信源熵

3.2、定长信源编码定理 －引理（续）

ε典型序列集和非ε典型序列集具有如下特性:

对于任意小的数$\varepsilon \geq 0$，$\delta \geq 0$，当N足够大时，则

（1） $P(G_\varepsilon) > 1-\delta$

$P(\overline{G}_\varepsilon) \leq \delta$

分析：

根据渐进等分割性定理：

$$\lim_{N\to\infty} P\left\{\left|\frac{I(\alpha_i)}{N} - H(S)\right| < \varepsilon\right\} = 1,$$

可利用契比雪夫不等式来分析

- 说明 ★

  - **在N次扩展信源中，信源序列可分为两大类：**

1. **ε典型序列：经常出现的信源序列，当N趋近于无穷时，这类序列出现的概率趋于1**
2. **非ε典型序列：当N趋近于无穷时，出现的概率趋于0**

3.2、定长信源编码定理 －引理（续）

ε 典型序列集和非**ε**典型序列集具有的第二个性质：

对于任意小的数$\varepsilon \geq 0$，$\delta \geq 0$，当N足够大时，则

（2）若$\alpha_i \in G_\varepsilon$,则：

$$2^{-N[H(S)+\varepsilon]} < P(\alpha_i) < 2^{-N[H(S)-\varepsilon]}$$

分析：

性质（2）可由ε典型序列的定义得出。若$\boldsymbol{\alpha_i \in G_\varepsilon}$,必满足：

$$\left|\frac{\boldsymbol{I(\alpha_i)}}{\boldsymbol{N}} - \boldsymbol{H(S)}\right| < \boldsymbol{\varepsilon}$$,即：

$$\boldsymbol{\varepsilon > \frac{I(\alpha_i)}{N} - H(S) > -\varepsilon}$$

则得：

$$2^{-N[\boldsymbol{H(S)}+\varepsilon]} < \boldsymbol{P(\alpha_i)} < 2^{-N[\boldsymbol{H(S)}-\varepsilon]}$$

- 说明：所有ε典型序列出现的概率近似相等，可粗略的认为典型序列出现的概率都等于$\boldsymbol{2^{-NH(S)}}$

3.2、定长信源编码定理 －引理（续）

ε 典型序列集和非ε典型序列集具有的第三个性质：

对于任意小的数$\varepsilon \geq 0$，$\delta \geq 0$，当N足够大时，则

(3)设M_G表示ε典型序列集中的序列数目，则

$$(1-\delta)2^{N[H(S)-\varepsilon]} \leq M_G \leq 2^{N[H(S)+\varepsilon]}$$

分析

$$1 = \sum_{i} P(\alpha_i) \geq \sum_{\alpha_i \in G_\varepsilon} P(\alpha_i) \geq \sum_{\alpha_i \in G_\varepsilon} 2^{-N[H(S)+\varepsilon]} = M_G 2^{-N[H(S)+\varepsilon]}$$

根据性质1和性质2：

$$1-\delta < P(G_\varepsilon) \leq M_G 2^{-N[H(S)-\varepsilon]}$$

占总集合比例：$\dfrac{M_G}{q^N} = \dfrac{2^{N[H(S)+\varepsilon]}}{q^N} = 2^{-N[\log q - H(S) - \varepsilon]}$

- 说明：ε典型序列集虽然是高概率集，但数目常常比非典型序列数要少很多。

3.2、定长信源编码定理 －证明

- 整个信源序列可分为 $\boxed{p(G_\varepsilon)+p(\overline{G_\varepsilon})=1}$

$$G_\varepsilon=\left\{\alpha_j:\left|\frac{I(\alpha_j)}{N}-H(S)\right|<\varepsilon\right\}\quad \overline{G_\varepsilon}=\left\{\alpha_j:\left|\frac{I(\alpha_j)}{N}-H(S)\right|\geq\varepsilon\right\}$$

对经常出现的信源序列进行编码，定义为 M_G 个

$$\boxed{r^l\geq M_G}\qquad \boxed{(1-\delta)2^{N[H(S)-\varepsilon]}\leq M_G\leq 2^{N[H(S)+\varepsilon]}}$$

$$r^l\geq 2^{N[H(S)+\varepsilon]}\Rightarrow l\log r\geq N[H(S)+\varepsilon]\Rightarrow \frac{l}{N}\geq\frac{H(S)+\varepsilon}{\log r}$$

3.2、定长信源编码定理 －证明（续）

- 当选定定长码的码字长度满足下式时，常出现集合中的所有序列都可以确定唯一码字来一一对应。

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

$$\boxed{l \log r > NH(S)}$$

- 此时，非典型序列被舍弃，因此造成译码错误。此时的错误概率

$$p\left\{\left|\frac{I(\alpha_j)}{N} - H(S)\right| \geq \varepsilon\right\} \leq \frac{D[I(s_j)]}{N\varepsilon^2}$$

$$p_E = p(\overline{G_\varepsilon}) \leq \frac{D[I(s_i)]}{N\varepsilon^2} = \delta(N,\varepsilon)$$

$\overline{G_\varepsilon}$集合中的元素出现时

当N趋于无穷时，译码错误概率趋于**0**。

3.2、定长信源编码定理 －证明（续）

反之，如果l满足下式： $\boxed{(1-\delta)2^{N[H(S)-\varepsilon]} \le M_G \le 2^{N[H(S)+\varepsilon]}}$

$$\frac{l}{N} \le \frac{H(S)-2\varepsilon}{\log r}, \quad 即 r^l \le 2^{N[H(S)-2\varepsilon]}$$

根据M_G的下界可知，此时选取的码字总数小于典型序列数.

将可以给予不同码字对应的码字序列的概率和记作：

$$P[G_\varepsilon 中 r^l 个 \alpha_i] \le r^l \cdot \max_{\alpha_i \in G_\varepsilon} P(\alpha_i) \qquad \boxed{2^{-N[H(S)+\varepsilon]} < P(\alpha_i) < 2^{-N[H(S)-\varepsilon]}}$$

$$\le 2^{N[H(S)-2\varepsilon]} \cdot 2^{-N[H(S)-\varepsilon]} = 2^{-N\varepsilon}$$

正确译码概率即为:$1-P_E \le 2^{-N\varepsilon}$

所以$P_E \ge 1-2^{-N\varepsilon}$

由此可见，当$N \to \infty$时，舍弃很多典型序列，

因此译码错误概率$P_E \to 1$

3.2、定长信源编码定理 – 证明总结

- 基本：唯一可译码存在条件 $q^N \le r^l$
- 对数目庞大的信源序列集合进行分类
  - ε 典型序列集的性质（渐近等分割性）
    - 这类序列出现的概率趋于1
    - 每个典型序列接近等概分布 $2^{-NH(S)}$
    - 序列数目占信源序列的比值很小
- 只考虑对经常出现的 ε 典型序列进行定长编码

3.2、定长信源编码定理（续）

- 平稳有记忆信源
  - 极限熵存在

例如：英文电报信源极限熵为 1.4比特/符号

每个原始信源符号所需要的码长理论极限

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

$$\Rightarrow \quad > 1.4\text{二元符号}/\text{信源符号}$$

- **编码信息率*R***
  - 编码后平均每个信源符号所能载荷的最大信息量

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

- **编码效率η**

$$\eta = \frac{H(S)}{R} = \frac{NH(S)}{l\log r}$$

最佳编码效率

$$\boxed{\eta = \frac{H(S)}{H(S)+\varepsilon}}$$

3.2、定长信源编码定理（续）

$$\boxed{p\left\{\left|\frac{I(\alpha_j)}{N}-H(S)\right|\geq\varepsilon\right\}\leq\frac{D[I(s_i)]}{N\varepsilon^2}\leq\delta}$$

$$\boxed{\delta(N,\varepsilon)=\frac{D[I(s_i)]}{N\varepsilon^2}}$$

- 序列码长*N*的分析
  - 当小于指定错误概率

$$N\geq\frac{D[I(s_i)]}{\varepsilon^2\delta}$$

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

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

对于一个给定信源：

序列长度、最佳编码效率、允许错误概率之间关系

3.2、定长信源编码定理 － 例题5.6

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ 0.4 & 0.18 & 0.10 & 0.10 & 0.07 & 0.06 & 0.05 & 0.04 \end{bmatrix}$$

要求：定长二元编码，编码效率 $\eta = 90\%$，且允许错误概率 $\delta \le 10^{-6}$，则需要至少多少个信源符号一起编码？

第一步 $H(S) = E[-\log p(s_i)] = -\sum_{i=1}^{8} p(s_i)\log p(s_i) = 2.55 bit/sym$

$$D[I(s_i)] = \sum_{i=1}^{8} p(s_i)[\log p(s_i)]^2 - [H(S)]^2 = \sum_{i=1}^{8} p(s_i)[\log p(s_i)]^2 - [2.55]^2 = 7.82$$

第二步

$$N \ge \frac{D[I(s_i)]}{\varepsilon^2 \delta} \qquad \eta = H(S)/[H(S) + \varepsilon] \Rightarrow \varepsilon = 0.28$$

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

3.2、定长信源编码定理 －例题5.6分析

- 问题的讨论：
  - 如果直接对信源（8个信源符号）进行二进制定长编码，则每个信源符号需要3比特表示。即 l/N=3 （N=1）。对N次扩展信源的ε典型序列进行编码，降为多少？($H(S)+\varepsilon$=2.55+0.28=2.83)
  - 若考虑只对 ε 典型序列进行编码，对于给定编码效率和错误概率，若要求容许错误概率越小、编码效率越高，则信源序列长度N必须越长。
  - 从例题中可以看到，要实现几乎无失真的定长编码，N的长度将会大到难以实现。因此为提高编码有效性需要付出很大的代价。

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

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

3.2、定长信源编码定理 – 例题5.7

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 \\ 3/4 & 1/4 \end{bmatrix}$$ 要求：等长二元编码，编码效率 $\eta = 96\%$，允许错误概率 $\delta \le 10^{-5}$

第一步 $H(S) = H(1/4, 3/4) = 0.811 bit/sym$

$$D[I(s_i)] = \sum_{i=1}^{2} p(s_i)[\log p(s_i)]^2 - [H(S)]^2 = \sum_{i=1}^{2} p(s_i)[\log p(s_i)]^2 - 0.811^2 = 0.4715$$

第二步 $$N \ge \frac{D[I(s_i)]}{\varepsilon^2 \delta} = \frac{D[I(s_i)]}{H^2(S)} \frac{\eta^2}{(1-\eta)^2 \delta}$$

$$N \ge \frac{0.4715}{(0.811)^2} \frac{(0.96)^2}{0.04^2 * 10^{-5}} = 4.13 * 10^7$$

$$\varepsilon = H(S)/\eta - H(S) = 0.811/0.96 - 0.811 = 0.034$$

主要内容

- 信源编码器
- 分组码
- 定长码
- 变长码

前言

平均码长界定定理

变长无失真信源编码定理

变长编码方法

4.1、变长码 – 概论

符号	码1	码2	码3	码4	码5
S_1	0	0	00	1	1
S_2	11	10	01	10	01
S_3	00	00	10	100	001
S_4	11	01	11	1000	0001

- 分组码
  - 非奇异码、奇异码
  - 唯一可译码、非唯一可译码
  - 即时码（逗点码/非延长码）、非即时码

- 信源编码的三种主要方法
  - 匹配编码 – 根据概率分布，代码长度不同
  - 变换编码 – 先信号变换，再编码
  - 识别编码 – 对有标准形状的符号进行编码

4.1、变长码 – Kraft不等式 – 定理5.4

信源符号数和码字长度何条件下可构成即时码？

若：信源符号集为 $S=\{s_1,s_2,\cdots,s_q\}$

码符号集 $X=\{x_1,x_2,\ldots,x_r\}$

设码字为 $W=\{W_1,W_2,\cdots,W_q\}$

其码长分别为 $l_1,l_2,\ldots,l_q$,

则即时码存在的充分必要条件是

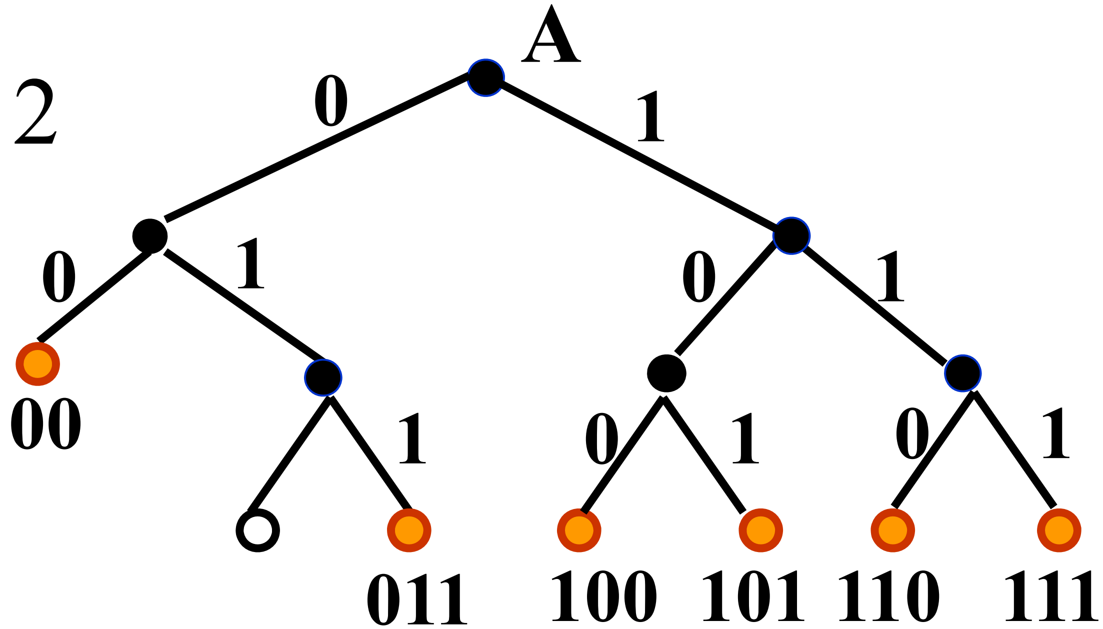
$$\sum_{i=1}^{q} r^{-l_i} \le 1$$

上式被称为克拉夫特（***Kraft***）不等式。

4.1、Kraft不等式 － 证明思路

- 观察树图结构，可得到如下不等式

$r=2$



$$\sum_{i=1}^{q} r^{L-l_i} \le r^L \quad \Rightarrow \quad \boxed{\sum_{i=1}^{q} r^{-l_i} \le 1}$$

4.1、Kraft不等式 － 定理证明

分析　树图法可构造即时码；

- 充分性
  - 满足*Kraft*不等式，则存在即时码

$$\sum_{i=1}^{q} r^{-l_i} \le 1 \Rightarrow r^{-l_1} + r^{-l_2} + \cdots + r^{-l_i} + \cdots + r^{-l_q} \le 1$$

设：最大码长l，且长度为i的共有n_i个，则

$l_i = 1 \Rightarrow$ 有n_1项r^{-1}

$l_i = 2 \Rightarrow$ 有n_2项r^{-2} $\Rightarrow \sum_{i=1}^{l} n_i r^{-i} \le 1 \xRightarrow{\boxed{*r^l}} \sum_{i=1}^{l} n_i r^{l-i} \le r^l$

$\vdots$

$l_i = l \Rightarrow$ 有n_l项r^{-l}　$n_l \le r^l - n_1 r^{l-1} - n_2 r^{l-2} - \cdots - n_{l-1} r$

4.1、Kraft不等式 － 定理证明（续）

$$n_l \le r^l - n_1 r^{l-1} - n_2 r^{l-2} - \cdots - n_{l-1} r$$ 两边同乘 r^{-1}

$$0 < n_l r^{-1} \le r^{l-1} - n_1 r^{l-2} - n_2 r^{l-3} - \cdots - n_{l-1} r^0$$

$$n_{l-1} < r^{l-1} - n_1 r^{l-2} - n_2 r^{l-3} - \cdots - n_{l-2} r$$

$$n_{l-2} < r^{l-2} - n_1 r^{l-3} - n_2 r^{l-4} - \cdots - n_{l-3} r$$

$n_3 < r^3 - n_1 r^2 - n_2 r$ 二阶节点可选n_2个点作为终端节点

$n_2 < r^2 - n_1 r$ 则二阶共有节点 $r(r-n_1) = r^2 - n_1 r$

$n_1 < r$ 一阶：选择n_1个点作为终端节点

剩余中间节点继续构造码树，满足所列不等式时，可以构造整个树图。即时码存在，充分性证明完毕。

4.1、Kraft不等式 － 定理证明（续）

- 必要性
  - 存在即时码，则满足*Kraft*不等式。
    - 上述过程反推即可。

- 说明
  - 该不等式给出了即时码的码长必须满足的条件！

4.1、McMillan不等式 － 定理5.5

- 在定理5.4给定的条件下，唯一可译码存在的充分必要条件

★ $$\sum_{i=1}^{q} r^{-l_i} \le 1$$

该不等式与*Kraft*不等式形式完全相同

■ 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 $l_1, l_2, \cdots, l_q$ 的唯一可译码，则一定存在具有相同码长的即时码。

■ 给出了唯一可译码/即时码存在的充要条件。

补充作业

（10分）有下列三种码字长度：

码长	l_1	l_2	l_3	l_4	l_5
C_1	2	1	2	4	1
C_2	2	2	2	3	1
C_3	1	4	6	1	1

1. 设编码符号集为$X=\{0, 1, 2\}$，上表中哪种码长可用来构造出唯一可译码？
2. 对每种可用的码长，构造出一个即时码。

4.2、平均码长界限定理 – 相关概念

- 码平均长度：
  - 信源**S** $\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & \cdots & s_q \\ p(s_1) & p(s_2) & \cdots & p(s_q) \end{bmatrix}$
  - 编码后的码字 $W_1, W_2, \cdots, W_q$
  - 码字对应长度 $l_1, l_2, \cdots, l_q$，则平均码长

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

★

■ 紧致码/最佳码：该唯一可译码对应的码平均长度小于所有其他的唯一可译码。

■ 编码后的信息传输率 $R = H(X) = \dfrac{H(S)}{\overline{L}}$ 比特/码符号

■ 编码效率

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

4.2、变长码 – 平均码长界限定理 – 定理5.7

- 平均码长界限定理
  - 离散无记忆信源的熵为$H(S)$
  - 用r个码元符号进行编码
  - 则总可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：

$$\underbrace{\frac{H(S)}{\log r}}_{\text{平均码长下界}} \leq \bar{L} < \underbrace{\frac{H(S)}{\log r}+1}_{\text{大于上界也可构成唯一可译码}}$$

4.2、平均码长界限定理 － 下界的证明

$$\boxed{\frac{H(S)}{\log r} \le \overline{L}}$$

$$\begin{aligned}
&H(S) - \overline{L} * \log r \\
&= -\sum_{i=1}^{q} p(s_i)\log p(s_i) - \log r \sum_{i=1}^{q} p(s_i) l_i \\
&= -\sum_{i=1}^{q} p(s_i)\log p(s_i) + \sum_{i=1}^{q} p(s_i)\log r^{-l_i} \\
&= \sum_{i=1}^{q} p(s_i)\log \frac{r^{-l_i}}{p(s_i)} \\
&\le \log \sum_{i=1}^{q} p(s_i)\frac{r^{-l_i}}{p(s_i)} \\
&= \log \sum_{i=1}^{q} r^{-l_i} \\
&\le 0
\end{aligned}$$

$$\boxed{\sum_{i=1}^{q} r^{-l_i} \le 1}$$

等号成立条件

$$\frac{r^{-l_i}}{p(s_i)} = 1 \qquad (i = 1,2,\ldots,q)$$

$$\Rightarrow l_i = -\frac{\log p(s_i)}{\log r} = -\log_r p(s_i)$$

$$H(S) - \overline{L}\log r \le 0 \Rightarrow \frac{H(S)}{\log r} \le \overline{L}$$

4.2、平均码长界限定理 – 例题5.8

$$l_i = -\frac{\log p(s_i)}{\log r} = -\log_r p(s_i) \Rightarrow p\left(s_i\right) = r^{-l_i}$$ （l_i是正整数）

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 \\ 1/2 & 1/4 & 1/8 & 1/8 \end{bmatrix}$$ 二元码：X={0， 1}

$\because p(s_i)$均呈现2^{-l_i}的形式：$p(s_1) = 2^{-1}, l_1 = 1$　　$p(s_2) = 2^{-2}, l_2 = 2$

$p(s_3) = 2^{-3}, l_3 = 3$　　$p(s_4) = 2^{-3}, l_4 = 3$

$$\therefore \quad l_1 = 1, l_2 = 2, l_3 = 3, l_4 = 3$$

$$\overline{L} = \frac{1}{2} * 1 + \frac{1}{4} * 2 + \frac{1}{8} * 3 + \frac{1}{8} * 3 = \frac{14}{8} \text{码符号/信源符号}$$

$$H(S) = -\frac{1}{2} * \log\frac{1}{2} - \frac{1}{4} * \log\frac{1}{4} - 2(\frac{1}{8} * \log\frac{1}{8}) = \frac{14}{8} bit/\text{信源符号}$$

$$\text{下界} = \frac{H(S)}{\log r} = \frac{14/8}{\log 2} = \frac{14}{8} \text{码符号/信源符号}$$

4.2、平均码长界限定理 – 上界的证明

- 存在性证明：构造一种唯一可译码，使平均码长小于上界。

$$\boxed{\bar{L} < \frac{H(S)}{\log r} + 1}$$

1 定义信源S 的概率分布 $p(s_i) = r^{-m_i}, m_i \le l_i < m_i + 1$

$$\Longrightarrow -\log_r p(s_i) \le l_i < -\log_r p(s_i) + 1$$

$$\frac{1}{p(s_i)} \le r^{l_i} < \frac{r}{p(s_i)} \Rightarrow p(s_i) \ge r^{-l_i} > \frac{p(s_i)}{r}$$

$$\Rightarrow \sum_{i=1}^{q} p(s_i) \ge \sum_{i=1}^{q} r^{-l_i} > \sum_{i=1}^{q} \frac{p(s_i)}{r} \Rightarrow 1 \ge \sum_{i=1}^{q} r^{-l_i} > \frac{1}{r}$$

唯一可译码
存在！

2 $$\sum_{i=1}^{q} p(s_i) \cdot l_i < -\sum_{i=1}^{q} p(s_i) \cdot \log_r p(s_i) + \sum_{i=1}^{q} p(s_i)$$

$$\Rightarrow \bar{L} < H_r(S) + 1 \Rightarrow \bar{L} < \frac{H(S)}{\log r} + 1$$

具体的编码算法 – Huffman编码

- 主要内容
  - 如何构造二元*Huffman*码？
  - *Huffman*编码所得码是否唯一？同一信源的不同编码之间的区别？
  - *r*元*Huffman*码的构造
  - 紧致码的证明

4.4、Huffman码 – 基本概念（续）

- *Huffman*码为紧致码/最佳码 ★

- 编码效率

$$\boxed{\eta = \frac{H(S)/\overline{L}}{\log r}} = \frac{H_r(S)}{\overline{L}} = \frac{H(S)}{\overline{L} * \log r}$$

用码的效率来衡量各种编码的优劣

4.4 变长码 －二元Huffman码的构造

编码步骤如下：

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，结果得到一个只包含$(n-1)$个信源符号的新信源。称为信源的第一次缩减信源，用S_1表示。

4.4、二元Huffman码的构造（续）

3. 将缩减信源S_1的符号仍按概率从大到小顺序排列，重复步骤2，得到只含$(n-2)$个符号的缩减信源S_2。

4. 重复上述步骤，直至缩减信源只剩两个符号为止，此时所剩两个符号的概率之和必为1。然后从最后一级缩减信源开始，依编码路径返回，就得到各信源符号所对应的码字。

以信源空间的概率分布为基准
用概率匹配方法进行信源编码

4.4、二元Huffman码的构造 － 例题5.9

- 设单符号离散信源如下，要求对信源编二元霍夫曼码。

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

解：构造*Huffman*码

信源符号	s_1	s_2	s_3	s_4	s_5	s_6	s_7
码字	**10**	**11**	**010**	**011**	**001**	**0000**	**0001**

4.4、二元Huffman码的构造 － 例题（续）

平均码长

$$\overline{L}=\sum_{i=1}^{7}p(s_i)l_i=2.72 \text{ 码元符号/信源符号}$$

信源熵

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

编码效率

$$\eta=\frac{H(S)/\overline{L}}{\log r}=\frac{2.61/2.72}{\log 2}=96\%$$

4.4、Huffman编码后的码字不是唯一的

- 造成非唯一的主要原因
  - 每次对缩减信源两个概率最小的符号的“0”或“1”分配是任意的，故得到的码字不同。
    - 不同的码元分配，得到的具体码字不同，但码长不变，平均码长也不变，所以没有本质区别。
  - 缩减信源时，若合并后的概率与其他概率相等，这几个概率的次序可任意排列，编出的码都是正确的，但得到的码字不相同。
    - 不同的排列方法得到的码长不相同。

4.4、Huffman编码非唯一 – 例题5.10

- 单符号离散信源，用两种不同的方法进行二进制霍夫曼编码

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

1）将合并后的新符号排在其它相同概率符号的后面（先参与编码），编码结果如下：

信源符号	s_1	s_2	s_3	s_4	s_5
码字	1	01	001	0000	0001

$$\overline{L} = \sum_{i=1}^{q} p(s_i) l_i = 2.2$$

码元符号/信源符号

4.4、二元Huffman编码 – 例题5.10（续）

2）将合并后的新符号排在其它相同概率符号的前面（后参与编码），编码结果如下：

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

$$\overline{L} = \sum_{i=1}^{q} p(s_i) l_i = 2.2 \quad \text{码元符号/信源符号}$$

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

4.4、Huffman编码 – 例题5.10（续）

- 分析
  - 两种方法平均码长相等。
  - 计算两种码的码长方差：

$$\sigma_1^{\ 2} = \sum_{i=1}^{5} p(s_i)(l_i - \overline{L})^2 = 1.36$$

$$\sigma_2^{\ 2} = \sum_{i=1}^{5} p(s_i)(l_i - \overline{L})^2 = 0.16$$

第二种方法得到码字的长度变化较小，易于实现。

4.4、r元Huffman码的构造过程

- 对r进制编码构成的r叉树：若所有中间节点都是r个分支，则对应的码字数（即，信源个数）必为 $r+k(r-1)$。其中，k为信源缩减次数。

4.4、r元Huffman码的构造过程

- 当将待编码的符号个数为$r+k(r-1)$，则可进行完全缩减。
- 当符号个数不满足条件时，则必须是最后一次缩减时有r个信源符号 —> 充分利用短码。

方法：第一次缩减时，补充一些概率为零的符号，使符号总数等于$r+k(r-1)$。

或者，用于第一次缩减的符号个数为$r-x$，x求解如下所示：

$$r+k(r-1)=q+x \quad \Rightarrow \quad \underline{x=\min_{>0}\left[k(r-1)-(q-r)\right]}$$

4.4、r元Huffman码的构造过程 – 例题5.11

- 例题：对如下单符号离散无记忆信源编三进制霍夫曼码。

$$\begin{bmatrix} S \\ P(S) \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ 0.4 & 0.18 & 0.1 & 0.1 & 0.07 & 0.06 & 0.05 & 0.04 \end{bmatrix}$$

- 这里：***r*=3**，***q*=8**

$$x = \min_{>0}\left[2 * k - 5\right] = 1$$

- 所以第一次取***r*－*x*=2**个符号进行编码。

$$\boxed{x = \min_{>0}\left[k(r-1)-(q-r)\right]}$$

4.4、r元Huffman码构造 – 例题5.11（续）

$$\begin{bmatrix} S \\ P(S) \end{bmatrix} = \begin{bmatrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \\ 0.4 & 0.18 & 0.1 & 0.1 & 0.07 & 0.06 & 0.05 & 0.04 \end{bmatrix}$$

信源符号	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8
码字	2	10	11	12	01	02	000	001
码长	1	2	2	2	2	2	3	4

$$H(S) = -\sum_{i=1}^{8} p(s_i)\log p(s_i) = 2.55(\text{比特/信源符号})$$

$$\overline{L} = \sum_{i=1}^{8} p(s_i)l_i = 1.69(\text{码元/信源符号})$$

$$\eta = \frac{H(S)/\overline{L}}{\log r} = \frac{2.55/1.69}{\log 3} = 95.2\%$$

4.4、Huffman码的补充说明 － 码字构成

信源符号	概率	缩减信源				码字	码长
		s_1	s_2	s_3	s_4		
x_1	0.4					0	1
x_2	0.18					10	2
x_3	0.1					11	2
x_4	0.1					12	2
x_5	0.07					21	2
x_6	0.06					22	2
x_7	0.05					200	3
x_8	0.04					201	3

若码字构造方向错误，则无法保证即时码的成立

4.4、Huffman码为紧致码

- 直观理解
  - 每次缩减总是选择概率最小的r个进行；
  - 缩减的顺序是从树的终端节点到树根；
  - 整个实现过程保证概率大的一定位于树的高层，即，对应码长较短；
- 证明
  - 每次缩减前后的码长差异只与参与本次缩减的符号的概率有关。

4.3、变长无失真信源编码定理 – 前言

- 平均码长界限定理，只考虑单个信源符号S_i
- 信源S发出的消息多为消息序列
  - 平均码长是否能进一步缩短？有无下限？
- 香农第一定理
  - 对输入消息序列直接编码，降低平均码长。
  - 对于有记忆信源（考虑符号相关性），降低平均码长。

考虑信源序列的变长编码 - 以无记忆为例

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0.9 & 0.1 \end{bmatrix} \quad [0 \quad 1] \quad \overline{L} = 1$$

$$\boxed{\overline{L} = \sum_i p(s_i) l_i}$$

$$\begin{bmatrix} X^2 \\ P \end{bmatrix} = \begin{bmatrix} 00 & 01 & 10 & 11 \\ 0.81 & 0.09 & 0.09 & 0.01 \end{bmatrix} \quad \overline{L_2} = 1.29 \Rightarrow \overline{L} = 0.645$$

$$[0 \quad 10 \quad 110 \quad 111]$$

$$\begin{bmatrix} X^3 \\ P \end{bmatrix} = \begin{bmatrix} 000 & 001 & 010 & 011 & 100 & 101 & 110 & 111 \\ 0.729 & 0.081 & 0.081 & 0.009 & 0.081 & 0.009 & 0.009 & 0.001 \end{bmatrix}$$

$$[0 \quad 100 \quad 101 \quad 1110 \quad 110 \quad 11110 \quad 111110 \quad 111111]$$

$$\overline{L_3} = 1.599 \Rightarrow \overline{L} = 0.533$$

1. 平均码长有无下限？

2. 下限与何有关？

4.3、无失真变长信源编码定理 – 定理5.8

- 香农第一定理（无失真变长信源编码定理）：★★

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。总可以找到一种编码方法，构成唯一可译码，使平均码长满足：

$$\frac{H(S)}{\log r} \le \frac{\overline{L}_N}{N} < \frac{H(S)}{\log r} + \frac{1}{N}$$

- 当 $N \to \infty$ 时，有 $\lim\limits_{N\to\infty} \frac{\overline{L}_N}{N} = H_r(S)$

- 香农第一定理推广到一般离散信源：$\lim\limits_{N\to\infty} \frac{\overline{L}_N}{N} = \frac{H_\infty}{\log r}$

4.3、变长无失真信源编码定理 － 证明

- 用码符号集*X*直接对离散无记忆信源*S*的*N*次扩展信源的每一个符号 α_j（信源*S*的符号序列）进行一一对应的无失真信源编码，根据平均码长界限定理，有：

$$\boxed{\frac{H(S)}{\log r} \le \bar{L} < \frac{H(S)}{\log r} + 1}$$

$$\frac{H(S^N)}{\log r} \le \bar{L}_N < \frac{H(S^N)}{\log r} + 1$$

$$\Rightarrow \frac{H(S^N)}{N\log r} \le \frac{\bar{L}_N}{N} < \frac{H(S^N)}{N\log r} + \frac{1}{N}$$

$$\Rightarrow \frac{H(S)}{\log r} \le \frac{\bar{L}_N}{N} < \frac{H(S)}{\log r} + \frac{1}{N}$$

4.3、变长无失真信源编码定理 – 说明

- 无记忆信源 - 当扩展次数N足够大时，每一个信源符号所需的平均码符号数，即平均码长，就可以无限接近下界值。
- 有记忆信源 - 必须考虑到信源符号间的依赖性（极限熵），以减少信源每发一个符号所携带的平均信息量，缩短平均码长。
- 编码效率★

$$\boxed{\eta = \frac{H(S)/\overline{L}}{\log r}} = \frac{H_r(S)}{\overline{L}} = \frac{H(S)}{\overline{L} * \log r}$$

4.3、变长无失真信源编码定理 – 例题5.12

$$\begin{bmatrix} S \\ P \end{bmatrix} = \begin{bmatrix} s_1 & s_2 \\ 3/4 & 1/4 \end{bmatrix} \quad \begin{bmatrix} S^2 \\ P \end{bmatrix} = \begin{bmatrix} s_1s_1 & s_1s_2 & s_2s_1 & s_2s_2 \\ 9/16 & 3/16 & 3/16 & 1/16 \end{bmatrix}$$

即时码：**0** **1** 即时码：**0** **10** **110** **111**

	未扩展	二次扩展	三次扩展	四次扩展
信源熵$\boldsymbol{H(S)}$	**0.811**	**0.811**		
平均码长 $\bar{L}$	**1**	**27/32**		
编码效率 η	**0.811**	**0.961**	**0.985**	**0.991**
信息传输率R	**0.811**	**0.961**	**0.985**	**0.991**

$$\boxed{\bar{L}_2 = 27/16}$$

$$\eta = \frac{H(S)/\bar{L}}{\log r} \qquad R = \frac{H(S)}{\bar{L}}\text{(比特/码符号)}$$

4.4、变长码 － Fano码

- 编码步骤如下：

1. 将概率按从大到小的顺序排列，令

$$p(s_1) \geq p(s_2) \geq \cdots \geq p(s_q)$$

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"。
4. 将每一分组再按同样方法划分，重复步骤2和3，直至概率不再可分为止。

4.4、变长码 – Fano码 – 例题 5.13

- 设单符号离散信源如下，要求对信源编二进制*Fano*码

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

4.4、变长码 – Fano码 – 例题 5.13（续）

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

从树根到终端节点的码树生成过程

4.4、变长码 – Fano码 – 例题 5.13（续）

- 平均码长为

$$\overline{L}=\sum_{i=1}^{7}p(s_i)l_i=2.74\ \text{码元符号/信源符号}$$

- 信源熵为

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

- 编码效率为

$$\eta=\frac{H(S)/\overline{L}}{\log r}=\frac{2.61/2.74}{1}=95.3\%$$

4.4、变长码 － 香农编码方法

- 编码步骤如下：**P.247 （8.12）**

1. 将信源符号按概率从大到小顺序排列，

$$p(s_1) \geq p(s_2) \geq \cdots \geq p(s_q)$$

2. 按下式计算第***i***个符号对应的码字的码长，

$$-\log p(s_i) \leq l_i < -\log p(s_i) + 1$$

3. 计算第***i***个符号的累加概率*Pi*：$P_i = \sum_{k=1}^{i-1} p(s_k)$

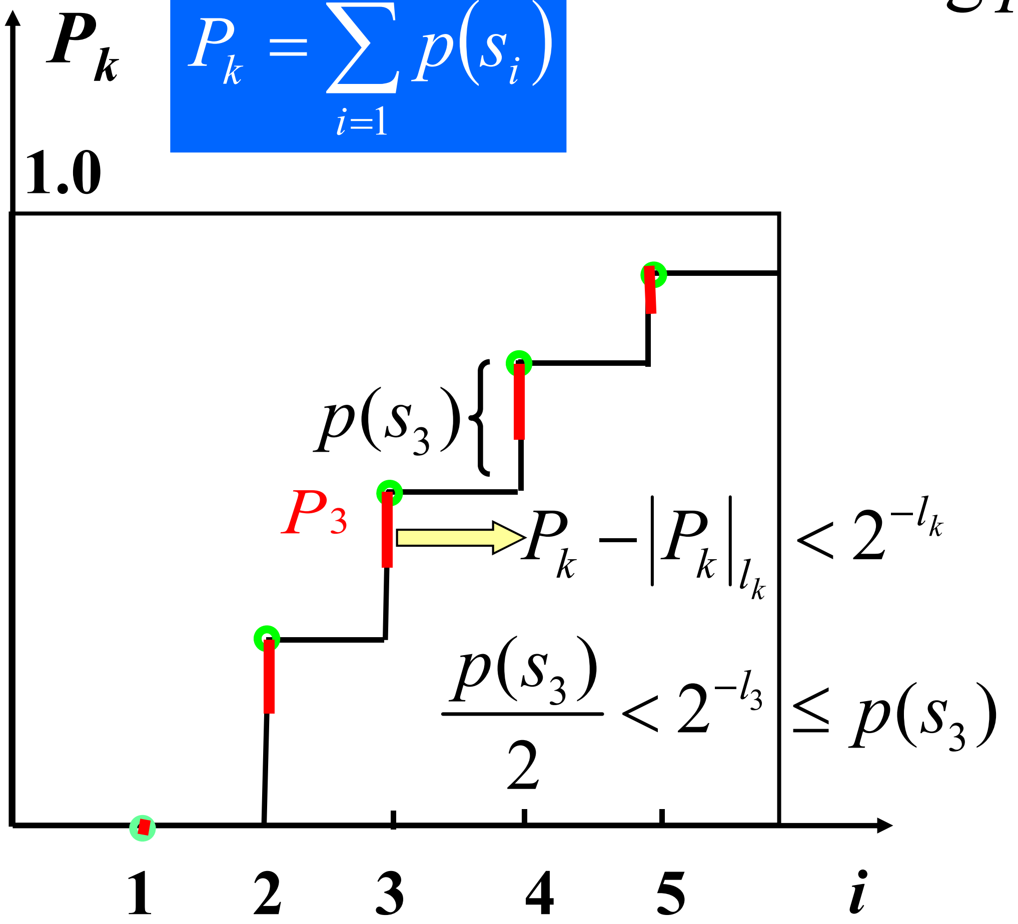
4. 将累加概率变换成二进制小数，取小数点后***l***l_i位数作为第***i***个符号的码字。

注：教材8.3节的“香农-费诺-埃利斯码”为辅助阅读。

4.4、变长码 － 香农编码方法 － 原理说明

累积概率示意图

$$P_k=\sum_{i=1}^{k-1}p(s_i)$$



码字区间分析

$$-\log p(s_i)\le l_i<-\log p(s_i)+1$$

$$\frac{p(s_i)}{2}<2^{-l_i}\le p(s_i)$$

对于 P_i，只取小数点后　　位，我们以符号 $|P_i|_{l_i}$ 表示

$$\boxed{P_{i-1}<|P_i|_{l_i}\le P_i}$$

4.4、变长码 – 香农编码方法 – 例题 5.14

- 设单符号离散信源如下，要求对信源编香农码。

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

4.4、香农编码方法 – 例题 5.14（续）

信源符号	符号概率	累加概率	$-\log p(s_i)$	码长	码字
s_1	0.20	0	2.34	3	000
s_2	0.19	0.2	2.41	3	001
s_3	0.18	0.39	2.48	3	011
$\boldsymbol{s_4}$	**0.17**	**0.57**	**2.56**	**3**	**100**
s_5	0.15	0.74	2.74	3	101
s_6	0.10	0.59	3.34	4	1110
s_7	0.01	0.99	6.66	7	1111110

$$P_4 = 0.57 \rightarrow 0*2^0 + 1*2^{-1} + 0*2^{-2} + 0*2^{-3} + 1*2^{-4} + \cdots \leftrightarrow 0.\underline{1001}\cdots$$

4.4、香农编码方法 – 例题 5.14（续）

平均码长

$$\bar{L}=\sum_{i=1}^{q}p(s_i)l_i=3.14 \quad \text{码元符号/信源符号}$$

信源熵

$$H(S)=-\sum_{i=1}^{q}p(s_i)\log p(s_i)=2.61 \quad \text{比特/信源符号}$$

编码效率

$$\eta=\frac{2.61/3.14}{1}=83.1\%$$

结论：

1. $\dfrac{H(S)}{\log r}\le\bar{L}<\dfrac{H(S)}{\log r}+1$

2. 香农码是即时码，但冗余度稍大，不是最佳码。

4.4、霍夫曼码、费诺码、香农码总结

- 霍夫曼码、费诺码、香农码都考虑了信源的统计特性，编码时将常出现的信源符号对应于较短的码字，使信源的平均码长缩短，从而可实现对信源的压缩。
- 霍夫曼码对信源的统计特性没有特殊要求，编码效率比较高，对编码设备的要求也比较简单，因此综合性能优于香农码和费诺码。霍夫曼编码也可用做决策树。
- 费诺码比较适合于对分组概率相等或接近的信源编码。
- 三种编码所得编码结果均不唯一！

第五+八章主要内容回顾

- 编码器
  - 编码器概论
  - 信源编码器与基本术语
- 分组码
  - 唯一可译性、即时码
- 定长码和定长编码定理
- 变长码
  - 变长无失真信源编码定理 – 香农第一定理
  - 几种变长编码方法

作业

- 编码理论
  - 唯一可译码的判别？ - 练习表格方法
    - 补充题：C = {0，10，1100，1110，1011，1101}
  - **5.1、5.2**
  - **5.4、5.6**（最后一问不做）

- 编码算法
  - **8.3**（**N=4**时不用求解）、**8.4、8.5**
  - **8.9、8.11**

补充作业

1、【定长编码定理相关】设无记忆二元信源，$P(1)=0.005$，$P(0)=0.995$。信源输出N=100的二元序列。如果只对该序列中“含有3个‘1’或小于3个‘1序列”序列进行一一对应的二元定长编码。

1）求码字所需的最小长度；

2）计算此时对应的编码误差。

2、【编码相关】设有一个信源发出符号A和B，他们是相互独立地发出，并已知$P(A)=1/4$，$P(B)=3/4$。

1)计算该信源的熵；

2）若用二进制代码组传输消息，A->0，B->1,求$P(0)$, $P(1)$；

3）该信源发出二重扩展消息，采用费诺编码，求编码后的信息传输率和$P(0)$, $P(1)$；

4）该信源发出三重扩展消息，采用霍夫曼编码，求编码后的信息传输率和$P(0)$, $P(1)$。