信息理论

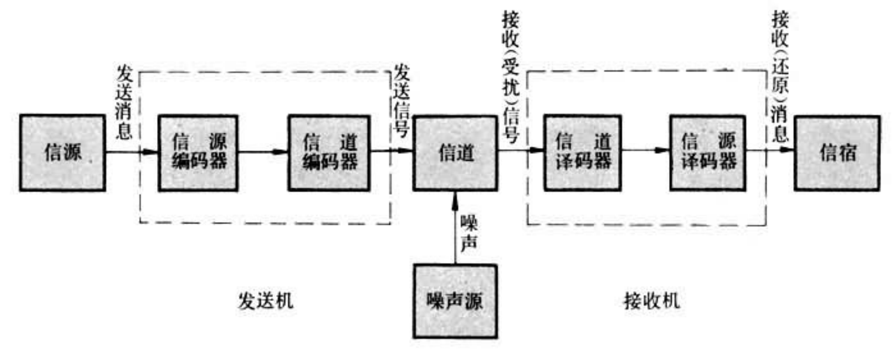
第二部分：信息的无损压缩

余官定　教授

浙江大学
信息与电子工程学院

信息论中的信源问题

信源



信息传输模型

信源是产生消息（包括消息序列）的源：图像、声波、符号，DNA序列等等。

信息论中的信源问题

- 构成描述信源的模型：随机变量序列或随机过程
- 计算信源输出的信息量，或者说信源的熵
- 如何有效地表示信源的输出，即信源编码

信源编码的目标

在代价最小的意义上来最有效地表示一个信源。
代价最小: 最少的比特数；即到底用多少个比特可以对一个信源进行编码?

信息论中的信源问题

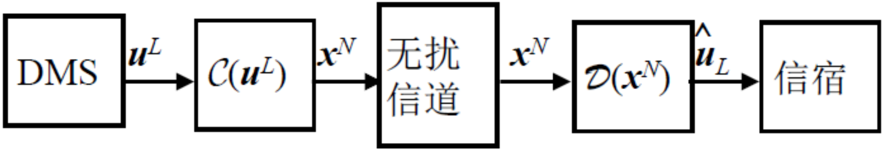
信源编码目标

在代价最小的意义上来最有效地表达一个信源。包括量化、压缩、映射、变换、自然语言翻译等许多具体和抽象的过程。

$$
\text{信源编码}\begin{cases}\text{无损编码}\begin{cases}\text{绝对无差错编码} & P_e^{(n)} = 0, \forall n \\ \text{渐进无差错编码} & P_e^{(n)} \xrightarrow{n\to\infty} 0\end{cases} \\ \\ \text{有损编码（限失真编码）}\end{cases}
$$

离散无记忆信源(DMS)

信源编码译码方框图



- $u^L = (u_1, u_2, \cdots, u_L)$: 长度为$L$的消息序列
- $x^N = (x_1, x_2, \cdots, x_N)$: 长度为$N$的码字
- $\mathcal{C}$: 码书、码字的集合（标号的集合）
- $\mathcal{D}$：译码器

第一讲：等长编码

- 了解离散无记忆信源等长编码的基本概念
- 掌握香农信源编码定理
- 了解随机序列的渐进等分性质
- 了解典型列的基本概念和意义

等长编码

离散无记忆信源

字符表: $\mathcal{A} = \{a_1, a_2, \ldots a_K,\}$
概率分布: $\{p_1, p_2, \ldots, p_K\}$
输出长度为L的消息序列$\boldsymbol{u}^L = \{u_1, u_2, \ldots, u_K\}$, 这样的序列一共有$K^L$个

编码

编码字符表: $\mathcal{B} = \{b_1, b_2, \ldots, b_D\}$
编码长度: N
无损编码: $D^N \geq K^L \Rightarrow N \geq \dfrac{L \log K}{\log D}$

等长编码

整数编码:对整数进行编码

- $\mathcal{A} = \{0, 1, \ldots, 9\}, L = 1$
  编码字符表: $\mathcal{B} = \{0, 1\}$, 则$N \geq \log 10, N = 4$
- $\mathcal{A} = \{0, 1, \ldots, 9\}, L = 2$或者$\mathcal{A} = \{0, 1, \ldots, 99\}, L = 1$
  则$N \geq \log 100, N = 7$,每个字符要3.5比特
- $\mathcal{A} = \{0, 1, \ldots, 9\}, L = \infty$
  $N_0 = \frac{N}{L} \geq \log 10 = 3.322$

信源输出序列越长, 编码效率越高, 接近$\log K$。
实际实现过程中, 编码序列越长, 译码时延也越长。

香农编码定理

问题

码字长度与熵的关系$N \geq \frac{L \log K}{\log D}, H(U) \leq \log K$。

有没有可能$N = \frac{LH(U)}{\log D}$?

香农编码定理

当$N > \frac{L\left(H(U) + \epsilon_L\right)}{\log D}$时, 可以实现无损编码。

当$N < \frac{L\left(H(U) - \epsilon_L\right)}{\log D}$时, 不存在无损编码。

香农编码定理

- 编码速率: $R = \dfrac{N}{L} \log D \to H(U)$。
- 无损编码是指信源编码的错误概率可以任意小, 但并非为零。
- 通常是对非常长的消息序列进行编码，特别当消息序列长度L趋于无穷时，才能实现Shannon编码。

香农编码定理

例如: $\mathcal{A} = \{0, 1, \ldots, 9\}, L = 1$

- 若$p_0 \to 0.5, p_1 \to 0.5, p_i \to 0, i > 2$
  则$H(A) \to 1$,每个符号的编码长度$R \to 1$。如何编?
- 若$p_0 \to 1, p_i \to 0, i > 1$,
  则$H(A) \to 0$, 每个符号的编码长度$N \to 1, R \to 0$. 如何编?
- 对于一般的$\{p_1, p_2, \ldots, p_K\}$, 如何实现香农定理?

香农编码定理的直观说明

长度为L的信源输出序列，个数为K^L。当L非常大时，根据大数定理，输出序列中符号a_i的个数约为Lp_i，具有这样构成成分的序列称为典型列。
典型列出现的概率为:

$$\Pi_{i=1}^{K} p_i^{Lp_i} = 2^{-LH(U)}$$

典型列的个数:

$$M = \frac{L!}{(Lp_1)!(L-Lp_1)!} \cdot \frac{(L-Lp_1)!}{(Lp_2)!(L-Lp_1-Lp_2)!} \cdots = \frac{L!}{\Pi_{i=1}^{K}(Lp_i)!}$$

香农编码定理的直观说明

$$M = \frac{L!}{(Lp_1)!(L-Lp_1)!} \cdot \frac{(L-Lp_1)!}{(Lp_2)!(L-Lp_1-Lp_2)!} \cdots = \frac{L!}{\Pi_{i=1}^{K}(Lp_i)!}$$

根据斯特林公式:$x! \approx \left(\frac{x}{e}\right)^x \sqrt{2\pi x}$, 有:

$$\begin{aligned}
\log M =& L\left(\log L - 1\right) + \frac{1}{2}\log(2\pi L) \\
& - \sum_{i=1}^{K} Lp_i \left(\log L + \log p_i - 1\right) - \sum_{i=1}^{K} \frac{1}{2}\log(2\pi L p_i) \\
\frac{\log M}{L} =& H(U) - \frac{1}{2L}\left((K-1)\log(2\pi L) + \sum_{i=1}^{K}\log p_i\right)
\end{aligned}$$

则:

$$L \to \infty, M = 2^{LH(U)}$$

香农编码定理的直观说明

典型列出现的概率: $\Pi_{i=1}^{K} p_i^{Lp_i} = 2^{-LH(U)}$
典型列出现的个数: $M = 2^{LH(U)}$
说明:

- 典型列几乎等概
- 当$L \to \infty$, 输出非典型列的可能性趋于零

香农编码定理

对典型列编码，编码速率为$R = H(U)$

渐进等分性质

长度为L的输出序列:$\boldsymbol{u}^L = (u_1, u_2, \ldots, u_L)$
序列发生的概率:$p(\boldsymbol{u}^L) = \Pi_{l=1}^{L} p(u_l)$
序列的自信息:$I(\boldsymbol{u}^L) = -\log p(\boldsymbol{u}^L) = \sum_{l=1}^{L} I(u_l)$

定义随机变量:$I_L \triangleq \dfrac{I(\boldsymbol{u}^L)}{L} = \sum_{l=1}^{L} \dfrac{I(u_l)}{L}$
则:

$$E(I_L) = E\left\{\frac{1}{L}\sum_{l=1}^{L} I(u_l)\right\} = E\left\{I(u)\right\} = H(U),$$

$$D(I_L) = D\left\{\frac{1}{L}\sum_{l=1}^{L} I(u_l)\right\} = \frac{1}{L^2} D\left\{\sum_{l=1}^{l} I(u^L)\right\} = \frac{1}{L} D\left\{I(u)\right\} \triangleq \frac{\sigma_I^2}{L}$$

渐进等分性质

由切比雪夫不等式:

$$P\{|\xi - E(\xi)| > \epsilon\} \leq \frac{\mathrm{Var}(\xi)}{\epsilon^2}, \forall \text{随机变量}\xi\text{和}\epsilon$$

可得

$$P\{|I_L - H(U)| > \epsilon\} \leq \frac{\sigma_I^2}{L\epsilon^2}$$

给定ϵ, 当L充分大时, $\dfrac{\sigma_I^2}{L\epsilon^2} < \epsilon$

故: $P\{|I_L - H(U)| > \epsilon\} \leq \epsilon$,
或: $P\{|I_L - H(U)| < \epsilon\} \geq 1 - \epsilon$。

典型列

令$H(U)$为一个离散无记忆信源DMS$\{U,p(\cdot)\}$的熵，ϵ为任意正数

$$A_\epsilon^{(L)}(U)=\left\{u^L:\left|-\frac{1}{L}\log p(u^L)-H(U)\right|<\epsilon\right\}$$

为给定DMS输出长度为L的ϵ典型列集合，简称典型列集，其中$\boldsymbol{u}^L\in U^L$。

典型列性质

1. 当L足够大时，

$$Pr\left(\boldsymbol{u}^L\in A_\epsilon^{(L)}(U)\right)>1-\epsilon$$

$$\begin{aligned}Pr\left(A_\epsilon^{(L)}(U)\right)&=\sum_{\boldsymbol{u}^L\in A_\epsilon^{(L)}(U)}p(\boldsymbol{u}^L)=\sum_{\boldsymbol{u}^L\in U^L}p(\boldsymbol{u}^L)I\left(\boldsymbol{u}^L\in A_\epsilon^{(L)}(U)\right)\\&=E\left\{I\left(\boldsymbol{u}^L\in A_\epsilon^{(L)}(U)\right)\right\}=Pr\left(\boldsymbol{u}^L\in A_\epsilon^{(L)}(U)\right)>1-\epsilon\end{aligned}$$

典型列

典型列性质

2.
$$2^{-L(H(U)+\epsilon)} \le p(\boldsymbol{u}^L) \le 2^{-L(H(U)-\epsilon)}, \text{if } \boldsymbol{u}^L \in A_\epsilon^{(L)}(U)$$

3.
$$(1-\epsilon)2^{L(H(U)-\epsilon)} \le \left|A_\epsilon^{(L)}(U)\right| \le 2^{L(H(U)+\epsilon)}$$

证明:

$$\begin{aligned}
1 &= \sum_{\boldsymbol{u}^L \in U^L} p(\boldsymbol{u}^L) \ge \sum_{\boldsymbol{u}^L \in A_\epsilon^{(L)}(U)} p(\boldsymbol{u}^L) \\
&\ge \sum_{\boldsymbol{u}^L \in A_\epsilon^{(L)}(U)} 2^{-L(H(U)+\epsilon)} \\
&= \left|A_\epsilon^{(L)}(U)\right| \cdot 2^{-L(H(U)+\epsilon)} \\
\therefore \left|A_\epsilon^{(L)}(U)\right| &\le 2^{L(H(U)+\epsilon)}
\end{aligned}$$

$$\begin{aligned}
1-\epsilon &\le \sum_{\boldsymbol{u}^L \in A_\epsilon^{(L)}(U)} p(\boldsymbol{u}^L) \\
&\le \sum_{\boldsymbol{u}^L \in A_\epsilon^{(L)}(U)} 2^{-L(H(U)-\epsilon)} \\
&= \left|A_\epsilon^{(L)}(U)\right| \cdot 2^{-L(H(U)-\epsilon)} \\
\therefore \left|A_\epsilon^{(L)}(U)\right| &\ge (1-\epsilon)2^{L(H(U)-\epsilon)}
\end{aligned}$$

典型列

离散无记忆源的输出序列分为两类:

- $A_\epsilon^{(L)}(U)$: 典型列集合, 高概率集
- $\overline{A_\epsilon^{(L)}}(U)$: 非典型列集合, 低概率集

- 个别非典型列出现的概率不一定比典型列出现概率小。
  $p_0 = p > 0.5, p_1 = 1 - p < 0.5$
  全0序列的自信息为$-L\log p \neq H(U)$, 因此不是典型列, 但是全0序列出现的概率为$p^L > p^{Lp}(1-p)^{L(1-P)}$。
  全1序列也不是典型列, 但是全1序列出现的概率小于典型列出现概率。
- 非典型列的数目不一定比典型列的数目少。
  $p = 0.25, H(U) = 0.81$,当$L = 100$时, $\left|A_\epsilon^{(L)}(U)\right| = 2^{81}$, 仅占所有序列的$\frac{1}{2^{19}}$。

香农编码定理证明

香农编码定理

当$N > \frac{L(H(U)+\epsilon)}{\log D}$时, 可以实现无损编码；当$N < \frac{L(H(U)-\epsilon)}{\log D}$时, 不存在无损编码。

证明：由于典型列的个数$\left|A_\epsilon^{(L)}(U)\right| \leq 2^{L(H(U)+\epsilon)}$, 所以当$N > \frac{L(H(U)+\epsilon)}{\log D}$, 可以对所有的典型列进行编码, 而对所有的非典型列用统一的一个序列(比如全D序列)编码, 当接收端收到全D序列时, 声称译码错误

$$p_e = p\left\{\hat{\boldsymbol{u}}^L \neq \boldsymbol{u}^L\right\} = p\left\{\boldsymbol{u}^L \notin A_\epsilon^{(L)}(U)\right\} < \epsilon$$

香农编码定理证明

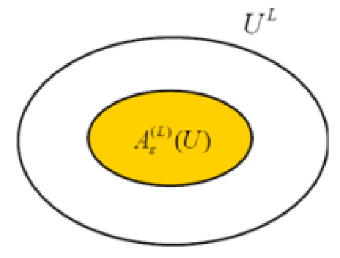
逆定理

当$N < \dfrac{L(H(U)-\epsilon)}{\log D}$, 记

$$D^N = 2^{L(H(U)-\epsilon-\epsilon_1)},$$

则每个典型列找到编码序列的概率为:

$$\frac{D^N}{A_\epsilon^{(L)}(U)} \leq \frac{2^{L(H(U)-\epsilon-\epsilon_1)}}{(1-\epsilon)2^{L(H(U)-\epsilon)}} = \frac{2^{-L\epsilon_1}}{1-\epsilon} \xrightarrow{L\to\infty} 0$$



第一讲 小结

- 信源等长编码
- 香农编码定理
- 随机序列的渐进等分性质
- 典型列的概念和意义

不等长编码

为什么要不等长编码

回到绝对无差错编码!
例子:
假设一个无记忆信源$\{a_1, a_2, a_3\}$, 概率密度为 $\{0.5, 0.25, 0.25\}$, $H(U) = 1.5bit$.

- 如果用等长编码, 则L很大时才能达到最佳编码的效率。
- 如果用不等长编码,将$\{a_1, a_2, a_3\}$分别编码成$\{1, 00, 01\}$,译码器在收到1时,译码成a_1,收到00时译码成a_2, 01时译码成a_3, 比如:
  $10001100001101 \xrightarrow{\text{译}} 1,00,01,1,00,00,1,1,01 = a_1a_2a_3a_1a_2a_2a_1a_1a_3$.
  平均码字长度：$\bar{n} = \sum_{k=1}^{K} p_k n_k = 1.5bit$!

DMS的不等长编码

消息集	概率分布	码字	码长
a_1	p_1	$\boxed{b_{11}b_{12}\cdots b_{1n_1}}$	n_1
a_2	p_2	$\boxed{b_{21}b_{22}\cdots b_{2n_2}}$	n_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$
a_K	p_K	$\boxed{b_{K1}b_{K2}\cdots b_{Kn_K}}$	n_K

$$\overline{n} = \sum_{k=1}^{K} p_k n_k$$

信源消息	出现概率	码A	码B	码C	码D
a_1	0.5	0	0	0	0
a_2	0.25	0	1	01	10
a_3	0.125	1	00	011	110
a_4	0.125	10	11	0111	1110

DMS的不等长编码

- 唯一可译性
- 即时可译性
- $\mathcal{B} = \{101, 00111, 10111, 11001\}$

只有当15个bit出现时才可以译码

10111,00111,0011$\boxed{1}$,101

101,11001,11001,1$\boxed{0}$111

译码延时无限大

$101110011100111001 11\cdots$

$10111,00111,00111,00111,\cdots$

$101,11001,11001,11001,\cdots$

唯一可译性

- 非奇异性

$$x_i \neq x_j \Rightarrow C(x_i) \neq C(x_j)$$

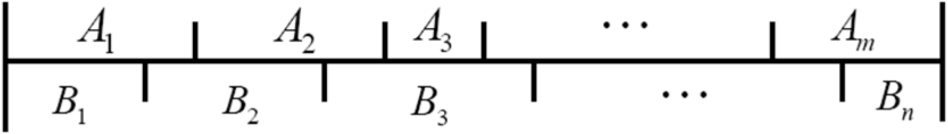
- 码扩展

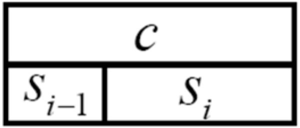
$$C^*(x_1x_2x_3\cdots x_n) = C(x_1)C(x_2)\cdots C(x_n)$$

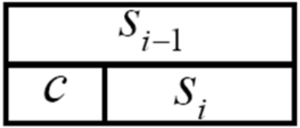
如果码的任意扩展都是非奇异的, 则称码是唯一可译的.

Sardinas & Petterson 判据

后缀分解







$S_{i-1}S_i = c$

$S_{i-1} = cS_i$

$$S_0 = c, s_i \in S_i$$

Sardinas & Petterson 判据

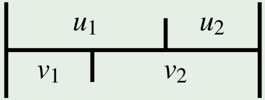
后缀分解集

$S_0 = C$	S_1	S_2	S_3	S_4	S_5	…
0	2	1	0	1	0	…
10			2	2	2	…
12			12		12	…
21			122		122	…
112					1	…
1122						…

Sardinas & Petterson 判据

一个码是唯一可译码的充分必要条件是除S_0外没有任何一个后缀分解集中包含码字.

$t = u_1 u_2 = v_1 v_2,\ u_1 \neq v_1$



不妨设$|u_1| > |v_1|$,则必有$u_1 = v_1 w$

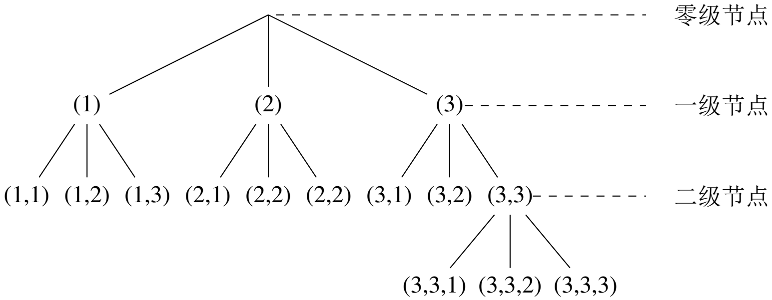
$\Rightarrow w u_2 = v_2$

$\Rightarrow \mathcal{D} \bigcap S_2 \neq \varnothing$

- 若$S_1 = \varnothing$, 则该码是即时可译并且唯一可译的.
- 若$S_n = \varnothing, n \geq 1$, 则该码是唯一可译, 且译码延时有限.

异字头码

如果一个码中没有任何一个码字是其它码字的前缀，则称该码是异字头码，即$S_1 = \varnothing$.

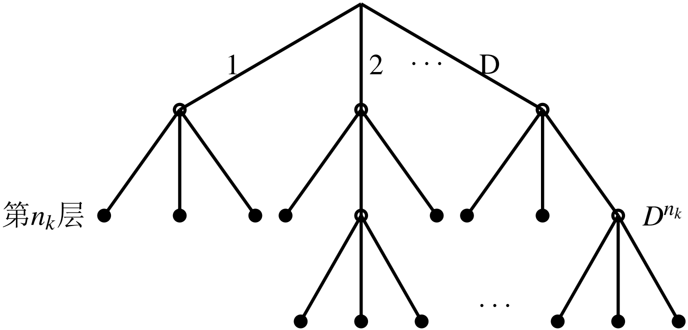


异字头码的树形表示：所有码字只出现在叶结点上.

Kraft不等式

存在长度为$n_1, n_2, \cdots, n_K$的D元异字头码的充要条件为:

$$\sum_{k=1}^{K} D^{-n_k} \leq 1$$



Kraft不等式证明

必要性

设$n_1 < n_2 < \cdots < n_k$, 可将异字头码的所有码字放置在码树的叶节点上, 每放置一个长为n_k的码字, 相当于砍掉其下生长的子树的共$D^{n_K - n_k}$个叶节点. 为了保证放置过程可行, 必须:

$$\sum_{n_k=1}^{n_K} D^{n_K - n_k} \leq D^{n_K}, \text{ 即 } \sum_{n_k=1}^{n_K} D^{-n_k} \leq 1.$$

充分性

可将所有码字依次放置在码树的叶节点上, 方法如下:
对长为n_k的码字在第n_k层任选一个可用的叶节点, 砍掉其下生长的子树, 相当于砍掉第n_K叶节点的$\frac{1}{D_{n_k}}$, 如果$\sum_{n_k=1}^{n_K} D^{-n_k} \leq 1$, 则上述放置过程一直可以进行下去, 既可以构成一个异字头码.

唯一可译性与Kraft不等式

任何唯一可译码必然满足Kraft不等式

证明:

$$\left(\sum_{k=1}^{K} D^{-n_k}\right)^r = \left(\sum_{k_1=1}^{K} D^{-n_{k_1}}\right)\left(\sum_{k_2=1}^{K} D^{-n_{k_2}}\right)\cdots\left(\sum_{k_r=1}^{K} D^{-n_{k_r}}\right)$$

$$= \sum_{k_1=1}^{K}\sum_{k_2=1}^{K}\cdots\sum_{k_r=1}^{K} D^{-(n_{k_1}+n_{k_2}+\cdots+n_{k_r})}$$

$$= \sum_{i=1}^{rn_{\max}} A_i D^{-i}$$

唯一可译性 $\Rightarrow A_i \le D^i$

$$\sum_{k=1}^{K} D^{-n_k} \le \left(\sum_{i=1}^{rn_{\max}} 1\right)^{\frac{1}{r}} \le (rn_{\max})^{\frac{1}{r}} = 2^{\frac{1}{r}\log_2 rn_{\max}} \xrightarrow{r\to\infty} 1$$

唯一可译码与异字头码的关系

唯一可译码
$\Longrightarrow$ Kraft不等式成立
$\Longrightarrow$ 存在一个同样长度的异字头码

不等长编码定理

任何一个唯一可译码的平均码字长度必须满足$\bar{n} \geq \frac{H(U)}{\log D}$, 同时一定存在一个D元唯一可译码, 其平均长度满足$\bar{n} \leq \frac{H(U)}{\log D} + 1$.

证明:

$$(1).\ H(U) - \bar{n}\log D = -\sum_{k=1}^{K}(p_k \log p_k + p_k n_k \log D)$$
$$= \sum_{k=1}^{K} p_k \log \frac{D^{-n_k}}{p_k} \leq \sum_{k=1}^{K} p_k \left(\frac{D^{-n_k}}{p_k} - 1\right)$$
$$= \sum_{k=1}^{K} D^{-n_k} - 1 \leq 0$$

当且仅当$p_k = D^{-n_k}$时, 等号成立.

(2). 选择唯一的n_k, 使得
$$D^{-n_k} \le p_k \le D^{-(n_k-1)}$$
对左边不等式两侧求和, 得
$$\sum_{k=1}^{K} D^{-n_k} \le \sum_{k=1}^{K} p_k = 1$$
因此, 存在长度为n_k的异字头码.
对右侧不等式取对数, 并求期望值, 得:
$$\sum_{k=1}^{K} p_k \log p_k < -\sum_{k=1}^{K} p_k(n_k - 1)\log D$$
所以:
$$H(U) > (\overline{n} - 1)\log D,\ \overline{n} \le \frac{H(U)}{\log D} + 1$$

不等长编码定理的扩展

对于长为L的平稳信源U^L, 有:

$$\frac{H(U^L)}{\log D} \le \bar{n}(U^L) < \frac{H(U^L)}{\log D} + 1$$

$$\Longrightarrow \frac{H(U^L)}{L\log D} \le \bar{n} = \frac{\bar{n}(U^L)}{L} < \frac{H(U^L)}{L\log D} + \frac{1}{L}$$

$$\Longrightarrow \bar{n} \xrightarrow{r\to\infty} \frac{H(U)}{\log D}$$

编码速率: $R \triangleq \bar{n}\log D$, 则:

$$R \xrightarrow{L\to\infty} H(U)$$

编码效率:

$$\eta = \frac{H(U)}{R}$$

Huffman编码(最佳不等长编码)

最佳不等长编码:
给定信源分布, 在平均码长最短的意义上最佳.

二元最佳码:
给定信源分布,其最佳二元编码必然满足:

- 其出现概率越小的消息所对应的码长越长,
- 出现概率最小的两个消息所对应的码长相等, 且码字最后一位不同.

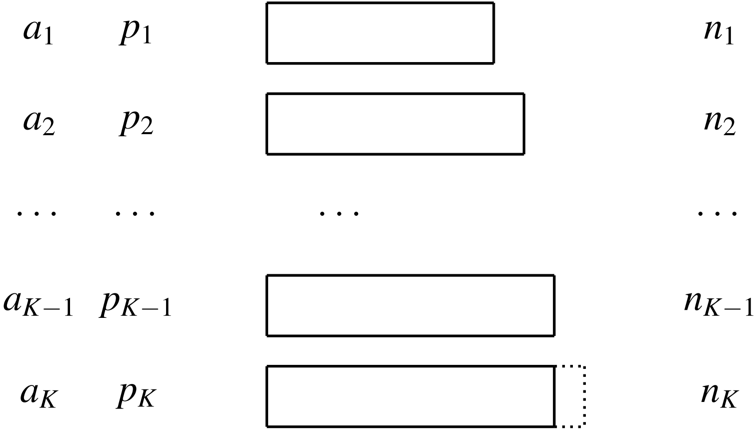
Huffman编码(最佳不等长编码)

1. $p_1 \geq p_2 \geq \cdots \geq p_K \Rightarrow n_1 \leq n_2 \leq \cdots \leq n_K$

$$
\begin{array}{cccc}
a_1 & p_1 & \boxed{b_{11}b_{12}\cdots b_{1n_1}} & n_1 \\
a_2 & p_2 & \boxed{b_{21}b_{22}\cdots\cdots b_{2n_2}} & n_2 \\
\cdots & \cdots & \cdots & \cdots \\
a_K & p_K & \boxed{b_{K1}b_{K2}\cdots\cdots b_{Kn_K}} & n_K
\end{array}
$$

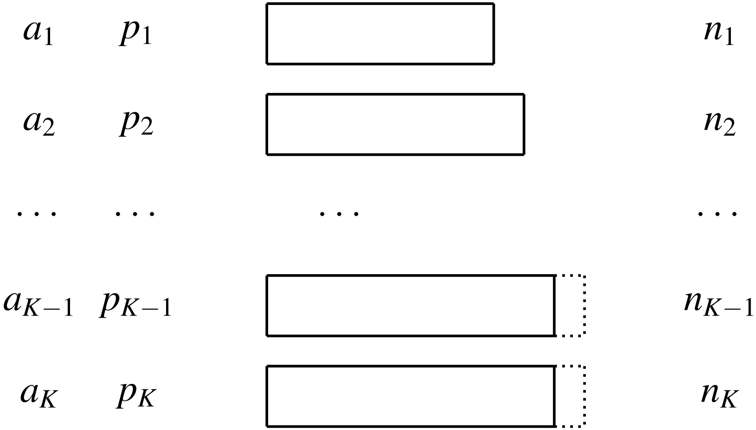
Huffman编码(最佳不等长编码)

2) $n_{K-1} = n_K$



Huffman编码(最佳不等长编码)

3) 对a_{K-1}和a_K，有$b_{K-1,n_K} \neq b_{K,n_K}$



辅助源

$$U \sim \begin{pmatrix} a_1 & a_2 & \cdots & a_{K-1} & a_K \\ p_1 & \geq p_2 & \cdots & \geq p_{K-1} & \geq p_K \end{pmatrix}$$

$$U' \sim \begin{pmatrix} a'_1 = a_1 & a'_2 = a_2 & \cdots & a'_{K-1} = a_{K-1} \cup a_K \\ p'_1 = p_1 & p'_2 = p_2 & \cdots & p'_{K-1} = p_{K-1} + p_K \end{pmatrix}$$

可递归编码原理

对辅助源U'的最佳编码也是对原始源的最佳编码.

证明:若$C'_1, C'_2, \cdots C'_{K-1}$是辅助源的最佳编码, 相应码长分别为$n'_1, n'_2, \cdots n'_{K-1}$. 对应地, U的码字$C_1, C_2, \cdots, C_K$的长度分别为:

$$n_k = n'_k, k = 1, 2, \cdots, K-2 \qquad n_k = n'_{K-1} + 1, k = K-1, K$$

故:

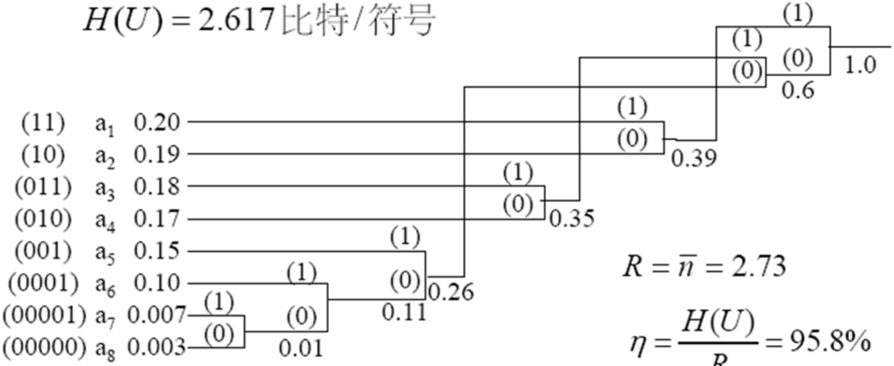
$$\bar{n} = \sum_{k=1}^{K} p_k n_k = \sum_{k=1}^{K-2} p_k n'_k + (p_{K-1} + p_K)(n'_{K-1} + 1) = \bar{n}' + (p_{K-1} + p_K)$$

所以由$\bar{n}'$最小可得出$\bar{n}$也是最小的.

Huffman编码

$$U = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 0.2 & 0.19 & 0.18 & 0.17 & 0.15 & 0.1 & 0.007 & 0.003 \end{pmatrix}$$

$H(U) = 2.617$ 比特/符号



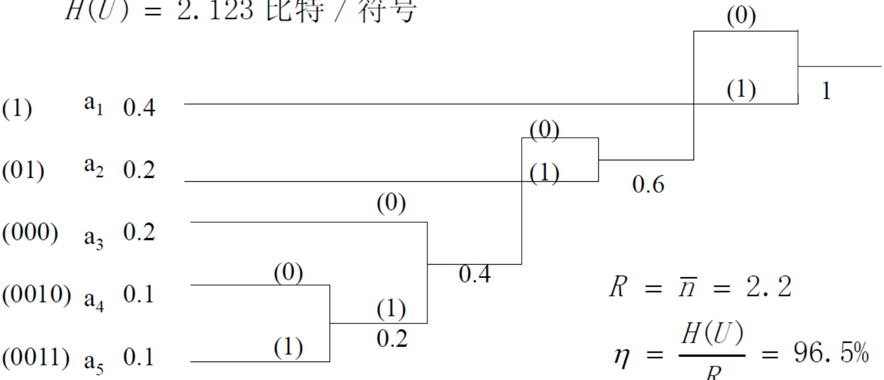
$R = \bar{n} = 2.73$

$$\eta = \frac{H(U)}{R} = 95.8\%$$

Huffman编码

$$U = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{pmatrix}$$

$$H(U) = 2.123 \text{ 比特 / 符号}$$

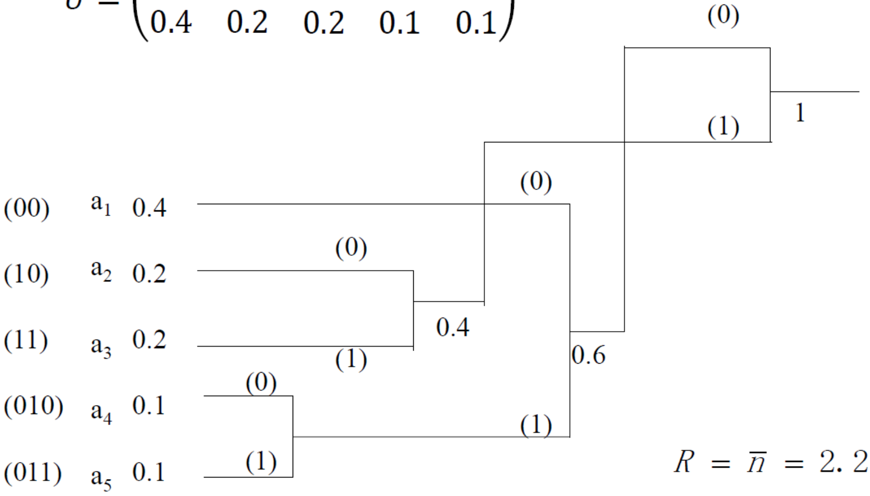


$$R = \overline{n} = 2.2$$

$$\eta = \frac{H(U)}{R} = 96.5\%$$

Huffman编码

$$U = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{pmatrix}$$

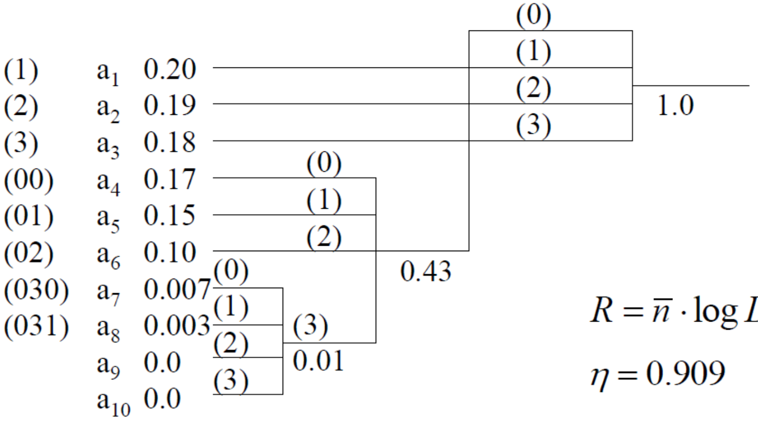


$$R = \bar{n} = 2.2$$

D元Huffman编码

若$K=(D-1)\cdot i+1$, 则每次均有D个消息要合并, 短标号得到充分利用;

若$K=(D-1)\cdot i+M$, 则必须增补$D-M$个概率为零的虚拟消息, 使得最后一次合并任有D个消息要合并, 从而充分利用短标号.



$$R=\overline{n}\cdot\log D=2.88$$

$$\eta=0.909$$

如何进一步增加编码效率?

第二讲小结

- 不等长编码的基本概念
- 唯一可译性和即时译码性(Sardinas & Petterson 判据)
- Kraft 不等式
- 不等长编码定理
- Huffman编码

作业:
3.1; 3.3; 3.4; 3.6; 3.11;

第三讲：几种不等长编码

- 最佳不等长编码(Huffman编码)
- Shannon编码
- Fano编码
- Shannon-Fano-Elias编码
- 算术编码
- 通用信源编码

Shannon编码

Shannon编码码长

a_k的编码长度为:

$$l_k = \left\lceil \log \frac{1}{p_k} \right\rceil, \quad 2^{-l_k} \le p_k < 2^{-l_k+1}$$

Shannon

$$U \sim \begin{pmatrix} a_1 & a_2 & \cdots & a_{K-1} & a_K \\ p_1 & \ge p_2 & \cdots & \ge p_{K-1} & \ge p_K \end{pmatrix}$$

编码方法

$$P_k = \sum_{i=1}^{k-1} p_i \qquad P_1 = 0$$

$$= 0.\underbrace{c_1 c_2 \cdots c_{l_k-1} c_{l_k}} c_{l_k+1} \cdots$$

$$l_k = \left\lceil \log \frac{1}{p_k} \right\rceil \quad 2^{-l_k} \le p_k < 2^{-l_k+1}$$

Shannon编码

p_k	0.5	0.25	0.125	0.125
l_k	1	2	3	3
P_k	0	0.5	0.75	0.875
二进制展开	0.000000	0.100000	0.1100000	0.1110000
码字	0	10	110	111

是一个前缀码, 且等同于最佳编码(Huffman编码)

Shannon编码

p_k	0.4	0.25	0.2	0.15
l_k	2	2	3	3
P_k	0	0.4	0.65	0.85
二进制展开	0.000000	$0.011\cdots$	$0.101\cdots$	$0.110\cdots$
码字	00	01	101	110
Huffman	1	01	000	001

是一个前缀码, 但不等同于最佳编码(Huffman编码)

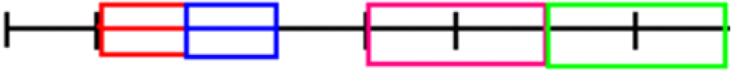
Shannon码是前缀码

如果把长度为l的二进制码字$\boldsymbol{z} = z_1 z_2 \cdots z_l$与一个区间$\left(0.z_1 z_2 \cdots z_l, 0.z_1 z_2 \cdots z_l + \frac{1}{2^l}\right]$对应, 则一个码是前缀码就等价于这些码字所对应的区间彼此不相交.

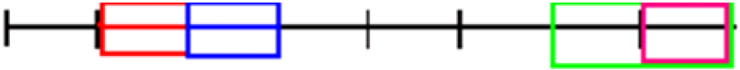
证明:如果是$\boldsymbol{z}^{(1)}$是$\boldsymbol{z}^{(2)}$的前缀, 且$\boldsymbol{z}^{(2)} = z_1 z_2 \cdots z_l$, 则$\boldsymbol{z}^{(1)} = z_1 z_2 \cdots z_k, k < l$, $\boldsymbol{z}^{(1)}$对应的区间包含$\boldsymbol{z}^{(2)}$. 同样, 如果两个区间相交, 则必然一个码是另一个码的前缀.

Shannon码是前缀码

例子：001,010,11,10 是前缀码



001,010,11,111 不是前缀码



Shannon码是前缀码

$$
\begin{aligned}
[P_{k+1}]_{l_{k+1}} &= [P_k + p_k]_{l_{k+1}} \geq \left[P_k + \frac{1}{2^{l_k}}\right]_{l_{k+1}} \\
&= [P_k]_{l_{k+1}} + \frac{1}{2^{l_k}} \geq [P_k]_{l_k} + \frac{1}{2^{l_k}}
\end{aligned}
$$

每一个码对应的区间不相交.

Shannon码是一个前缀码.

Shannon码的编码效率

$$H(U) \leq \bar{n} = \sum_{k=1}^{K} l_k p_k = \sum_{k=1}^{K} \left\lceil \log \frac{1}{p_k} \right\rceil p_k$$
$$\leq \sum_{k=1}^{K} \left(\log \frac{1}{p_k} + 1 \right) p_k = H(U) + 1$$

与Huffman码相比, Shannon码渐进收敛性能较差.

例子:

$$p_1 = 0.9999, p_2 = 0.0001$$
$$l_1 = 1\text{bit}, l_2 = 14\text{bit}$$

Shannon码的编码效率

Shannon码逼近Shannon信源编码定理.

例子: 两个符号, p_1, p_2.
序列$\boldsymbol{u}^k$含有k个1, $n-k$个0, 则序列出现的概率:

$$p\left(\boldsymbol{u}^k\right) = p^k(1-p)^{n-k}$$

编码长度: $l_k = \left\lceil \log \frac{1}{p(\boldsymbol{u}^k)} \right\rceil \leq k \log \frac{1}{p} + (n-k) \log \frac{1}{1-p} + 1$
平均码长:

$$\overline{L} \leq \sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} \left(k \log \frac{1}{p} + (n-k) \log \frac{1}{1-p} + 1 \right) = nH(U) + 1$$

Shannon码的编码效率

$$
\begin{aligned}
&\sum_{k=0}^{n} C_n^k p^k (1-p)^{n-k} \left(k \log \frac{1}{p} + (n-k) \log \frac{1}{1-p} + 1\right) \\
&= \log \frac{1}{p} \cdot (1-p)^n \sum_{k=0}^{n} C_n^k k \left(\frac{p}{1-p}\right)^k + \log \frac{1}{1-p} \cdot p^n \sum_{k=0}^{n} C_n^k k \left(\frac{1-p}{p}\right)^k + 1 \\
&= n \log \frac{1}{p} \cdot (1-p)^n \frac{p}{1-p} \left(\frac{p}{1-p} + 1\right)^{n-1} + n \log \frac{1}{1-p} \cdot p^n \frac{1-p}{p} \left(\frac{1-p}{p} + 1\right)^{n-1} + 1 \\
&= nH(U) + 1
\end{aligned}
$$

$$
\begin{aligned}
\sum_{k=0}^{n} C_n^k k x^k &= \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} k x^k = n \sum_{k=1}^{n} \frac{(n-1)!}{(k-1)!(n-k)!} x^k \\
&= nx \sum_{k=0}^{n-1} C_{n-1}^k x^k = nx(1+x)^{n-1}
\end{aligned}
$$

Fano编码

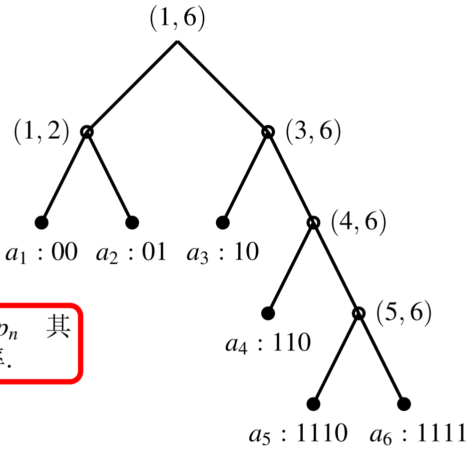
$$U = \left(\begin{array}{cccc|ccc} a_1 & a_2 & \cdots & a_k & a_{k+1} & \cdots & a_K \\ p_1 \geq & p_2 \geq & \cdots & p_k \geq & p_{k+1} \geq & \cdots & \geq p_K \end{array} \right)$$

概率和对分法:

$$\left| \sum_{i=1}^{k} p_i - \sum_{i=k+1}^{K} p_i \right| \to \min$$

Fano编码的效率

$$U = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \\ 0.3 & 0.25 & 0.2 & 0.15 & 0.05 & 0.05 \end{pmatrix}$$

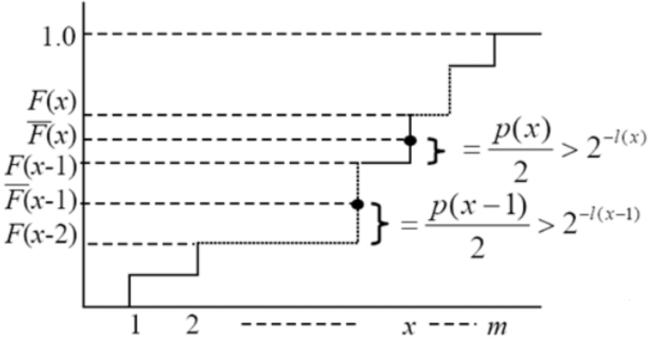


$\bar{n} \leq H(U) + 1 - 2p_n$ 其中p_n为最小符号概率.

Shannon-Fano-Elias编码

特点: 不需要对概率进行排序.

$$U = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \\ p(1) & p(2) & \cdots & p(m) \end{pmatrix}$$



$$\overline{F}(x) \triangleq \sum_{i<x} p(i) + \frac{1}{2}p(x)$$

$$F(x) \triangleq \sum_{i\leq x} p(i)$$

$$l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil + 1$$

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

$$\overline{F}(x) - \left\lfloor \overline{F}(x) \right\rfloor_{l(x)} < 2^{-l(x)}$$

$$\left\lfloor \overline{F}(x) \right\rfloor_{l(x)} = 0.z_1 z_2 z_3 \cdots z_{l(x)}$$

Shannon-Fano-Elias编码的效率

$$\bar{n} = \sum_x p(x)l(x) = \sum_x p(x)\left\{\left\lceil \log \frac{1}{p(x)} \right\rceil + 1\right\} < H(U) + 2$$

x	$p(x)$	$F(x)$	$\overline{F}(x)$	二进制表示	$l(x)$	码字	Huffman码
1	0.25	0.25	0.125	0.001	3	001	01
2	0.5	0.75	0.5	0.10	2	10	1
3	0.125	0.875	0.8125	0.1101	4	1101	001
4	0.125	1.0	0.9375	0.1111	4	1111	000

离散有记忆信源

令ϵ是任意小正数, 对平稳有记忆信源$\{u^L, p(u^L)\}$进行D元不等长编码, 则总可以找到一个$L(\epsilon)$, 当$L > L(\epsilon)$时, 平均编码码长$\bar{n}$ 满足:

$$\frac{H(U|U^\infty)}{\log D} \le \bar{n} < \frac{H(U|U^\infty)}{\log D} + \epsilon$$

编码方法:Shannon码.

$$H_\infty(U) = \sum_s q(s) H(U|s) = H(U|S)$$

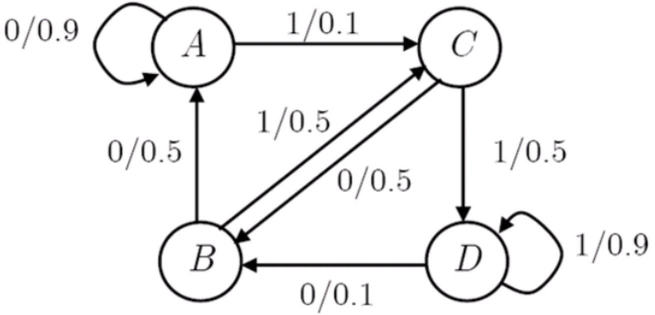
编码思路:

1. 用$\lceil \log |S| \rceil$个比特对初始状态S进行编码
2. 对状态的输出长度为L的消息序列进行不等长编码.

当L充分大时:

$$\frac{H(U|U^\infty)}{\log D} \le \bar{n} < \frac{H(U|U^\infty)}{\log D} + \frac{1}{L}$$

马尔可夫信源编码示例



$$q(A) = q(D) = \frac{5}{12}, q(B) = q(C) = \frac{1}{12}$$

$$H(X|S = A) = H(X|S = D) = 0.469$$

$$H(X|S = C) = H(X|S = B) = 1$$

$$\therefore H(X|S) = \sum_{S \in \{A,B,C,D\}} q(s) H(X|s) = 0.558$$

$\therefore$ 二元编码时

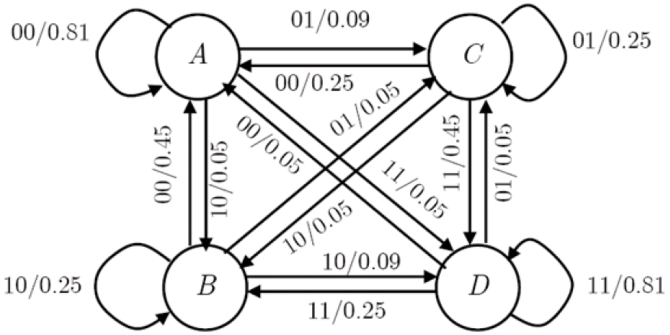
$$\bar{n}_{\text{opt}} = \frac{H_\infty}{\log 2} = H(X|S) = 0.558$$

马尔可夫信源编码示例

对上述马尔可夫信源输出的单符号序列进行二元编码, 则任何状态下都需要至少1比特来标记该输出符号, 因此平均编码码长:

$$\bar{n} \geq 1$$

输出的二符号序列进行编码



马尔可夫信源编码示例

消息	A	B	C	D
00	0(0.81)	0(0.45)	10(0.25)	111(0.5)
01	10(0.09)	111(0.05)	110(0.25)	110(0.05)
10	110(0.05)	110(0.25)	111(0.05)	10(0.09)
11	111(0.05)	10(0.25)	0(0.45)	0(0.81)
平均码长$\bar{n}_L$	1.29	1.85	1.85	1.29

$$\bar{n} = \frac{1}{L}\left(q(A)\times 1.29 + q(B)\times 1.85 + q(C)\times 1.29 + q(D)\times 1.29\right) = 0.6917$$

若进一步对更长的输出符号序列编码效率将更高.

第三讲小结

- Shannon 编码
- Fano 编码
- S-F-E编码
- 平稳信源编码
- 马尔可夫信源编码

作业:
3.10; 3.16;