信息论

1. 熵和互信息量的基本概念
2. 熵和互信息量的基本性质
3. 信源的无失真编码
4. 信道容量——代价函数
5. 最佳接收和错误概率的估计
6. 信道编码定理
7. 信源的率失真函数和限失真信源编码
8. 非离散信源和信道

3 信源的无失真编码

目录

3.1 引 言

- 信源分类
- 信源编码模型
- 无失真信源编码

3.1.1 信源分类

- 根据信源输出在时间上是否连续，以及取值是否连续，可分为：

  离散信源

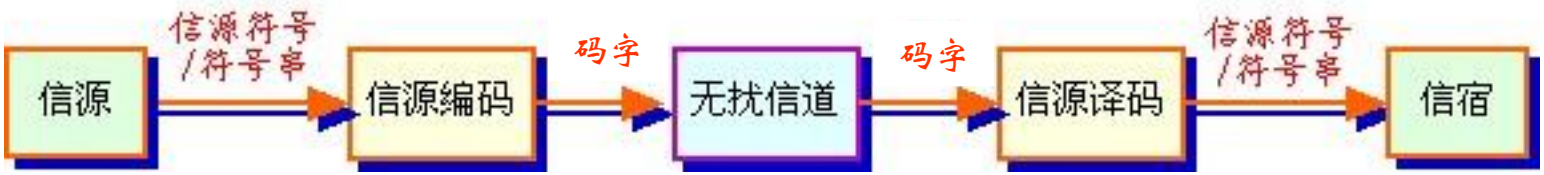
  非离散信源

- 按信源的记忆特性，离散信源又分为：

  离散无记忆信源

  离散有记忆信源

3.1.2 信源编码模型



例：五个不同的信源符号可表示为：

{000，001，011，110，101} → 等长编码

{01，011，0111，01111，011111} → 非等长编码

3.1.3 无失真信源编码

- **唯一可译码**
  - 每个信源符号、符号串都至少有一个码字与之对应
  - 不同的信源符号、符号串对应不同的码字
  - 非等长编码能自动识别一个码字的结束
- 唯一可译码在无扰信道中传输时，其译码错误概率为0，故称为**无失真编码**

目录

- 引言
- 无失真信源编码定理
- 非等长信源编码
- 一种实用编码方法:LZ编码

3.2 无失真信源编码定理

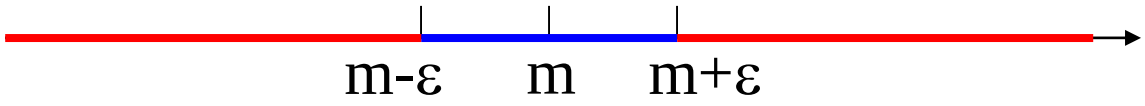
- 引理3.1：Chebyshev Inequality
- 引理3.2：离散无记忆信源的渐近均衡性
- 定理3.1：无失真信源编码定理

3.2.1 Chebyshev Inequality

- 引理3.1: (a) 随机变量x , 均值m , 方差 σ^2,

$$\text{则} \quad P_r\left\{\left|x-m\right|>\varepsilon\right\}\leq\frac{\sigma^2}{\varepsilon^2}$$

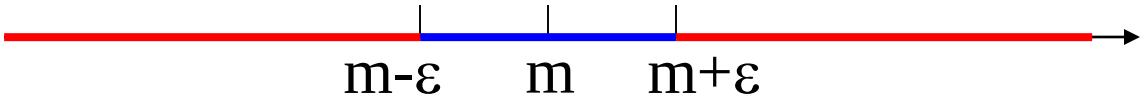
其中 $\varepsilon>0$.



3.2.1 Chebyshev Inequality

证明：
$$\sigma^2 = \int_{-\infty}^{+\infty} (x-m)^2 p(x)dx$$
$$\geq \int_{-\infty}^{m-\varepsilon} (x-m)^2 p(x)dx + \int_{m+\varepsilon}^{+\infty} (x-m)^2 p(x)dx$$
$$\geq \int_{-\infty}^{m-\varepsilon} \varepsilon^2 p(x)dx + \int_{m+\varepsilon}^{+\infty} \varepsilon^2 p(x)dx$$
$$= \varepsilon^2 P_r\{|x-m| > \varepsilon\}$$

因此 $$P_r\{|x-m| > \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$$



3.2.1 Chebyshev Inequality

- 引理3.1: (b) 设 $z_1, z_2, \ldots, z_N$是独立同分布随机变量
  (i.i.d-independent and identically distributed),
  有均值 $\overline{Z}$, 方差 σ^2 .
  对于$\forall\varepsilon>0$， 有

$$P_r\left\{\left|\frac{1}{N}\sum_{n=1}^{N}z_n-\overline{z}\right|>\varepsilon\right\}\le\frac{\sigma^2}{N\varepsilon^2}$$

  (可证明 $\frac{1}{N}\sum_{n=1}^{N}z_n$ 的均值为$\overline{Z}$, 方差为$\frac{\sigma^2}{N}$)

3.2 无失真信源编码定理

- 引理3.1：Chebyshev Inequality
- 引理3.2：离散无记忆信源的渐近均衡性
- 定理3.1：无失真信源编码定理

3.2.2 DMS 渐近均衡性

引理3.2: 设离散无记忆信源每个符号的熵为H(u).

给定 $\forall\varepsilon>0$,

考虑信源输出的长度为N的符号串的子集:

$$S(N,\varepsilon) = \{\mathbf{u}: 2^{-N[H(u)+\varepsilon]} \le p(\mathbf{u}) \le 2^{-N[H(u)-\varepsilon]}\}$$

则这个子集中的每个符号串可以用长度为L_N的二进制序列无失真地表示，L_N满足:

$$N[H(u)+\varepsilon] \le L_N < N[H(u)+\varepsilon]+1 \qquad (1)$$

且 $$P_r[\mathbf{u} \notin S(N,\varepsilon)] \le \frac{\sigma^2}{N\varepsilon^2} \qquad (2)$$

其中 $$\sigma^2 = \sum_u \left[\log\frac{1}{p(u)} - H(u)\right]^2 p(u)$$

3.2.2 DMS 渐近均衡性

证明: (1) S(N,ε)是长度为N的符号串的子集, 有

$$1=\sum_{\mathbf{u}} p(\mathbf{u}) \geq \sum_{\mathbf{u}\in S(N,\varepsilon)} p(\mathbf{u}) \geq \sum_{\mathbf{u}\in S(N,\varepsilon)} 2^{-N[H(u)+\varepsilon]}$$

$$= 2^{-N[H(u)+\varepsilon]} \cdot \left|S(N,\varepsilon)\right|$$

因此 $\left|S(N,\varepsilon)\right| \leq 2^{N[H(u)+\varepsilon]}$

子集S(N,ε)中符号串的数量

为了实现对S(N,ε)中符号串的无失真二进制编码, 码长为L_N的码字个数应满足:

$$2^{L_N} \geq \left|S(N,\varepsilon)\right|$$

3.2.2 DMS 渐近均衡性

证明: (1) (cont.)可取L_N满足:

$$L_N \geq N[H(u)+\varepsilon],$$

$$L_N-1 < N[H(u)+\varepsilon].$$

因此所有$S(N,\varepsilon)$中的符号串可以用长为L_N的二进制序列无失真地表示:

$$N[H(u)+\varepsilon] \leq L_N < N[H(u)+\varepsilon]+1.$$

3.2.2 DMS 渐近均衡性

证明: (2)定义 $F_N = P_r\{\mathbf{u} \notin S(N,\varepsilon)\}$

根据定义知:

$S(N,\varepsilon) = \{\mathbf{u}: 2^{-N[H(u)+\varepsilon]} \le p(\mathbf{u}) \le 2^{-N[H(u)-\varepsilon]}\}.$

$$= \{\mathbf{u}: -N[H(u)+\varepsilon] \le \log p(\mathbf{u}) \le -N[H(u)-\varepsilon]\}$$

$$= \{\mathbf{u}: -N\varepsilon \le \log p(\mathbf{u}) + NH(u) \le N\varepsilon\}$$

$$= \left\{\mathbf{u}: \left|\log \frac{1}{p(\mathbf{u})} - NH(u)\right| \le N\varepsilon\right\}$$

$$= \left\{\mathbf{u}: \left|\frac{1}{N}\log \frac{1}{p(\mathbf{u})} - H(u)\right| \le \varepsilon\right\}$$

3.2.2 DMS 渐近均衡性

证明：（2）(cont.)

$$\overline{S(N,\varepsilon)}=\left\{\mathbf{u}:\left|\frac{1}{N}\log\frac{1}{p(\mathbf{u})}-H(u)\right|>\varepsilon\right\}$$

$$=\left\{\mathbf{u}:\left|\frac{1}{N}\sum_{n=1}^{N}\log\frac{1}{p(u_n)}-H(u)\right|>\varepsilon\right\}$$

定义 $z_n=\log\dfrac{1}{p(u_n)}$

离散无记忆信源

则：$E[z_n]=H(u)$

$$Var[z_n]=\sigma^2=\sum_u\left[\log\frac{1}{p(u)}-H(u)\right]^2p(u)$$

3.2.2 DMS 渐近均衡性

证明: （2）(cont.)

由引理3.1可得

$$P_r\left[\left|\frac{1}{N}\sum z_n - \bar{z}\right| > \varepsilon\right] \le \frac{\sigma^2}{N\varepsilon^2}$$

因此 $$F_N = P_r\left[\left|\frac{1}{N}\sum_{n=1}^{N}\log\frac{1}{p(u_n)} - H(u)\right| > \varepsilon\right] \le \frac{\sigma^2}{N\varepsilon^2}$$

可见, 当 $N\to\infty$ 时, $F_N\to 0$.

3.2.2 DMS 渐近均衡性

说明1: 定义

$$S(N,\varepsilon)=\left\{\mathbf{u}:\left|\frac{1}{N}\log\frac{1}{p(\mathbf{u})}-H(u)\right|\le\varepsilon\right\}=\left\{\mathbf{u}:\log\frac{1}{p(\mathbf{u})}\approx NH(u)\right\}$$

$$\overline{S(N,\varepsilon)}=\left\{\mathbf{u}:\left|\frac{1}{N}\log\frac{1}{p(\mathbf{u})}-H(u)\right|>\varepsilon\right\}$$

特点: (1) S(N,ε)中的典型序列近似等概

(2) $N\to\infty$时，$P_r\{u\in S(N,\varepsilon)\}\to 1$，$P_r\{u\in\overline{S(N,\varepsilon)}\}\to 0$

(3) S(N,ε)中序列数目所占比例不大

(4) 对S(N,ε)中序列等长编码时，$L_N\approx N\,H(u)$

3.2.2 DMS 渐近均衡性

说明2:

离散无记忆信源的渐进均衡性(AEP-Asymptotic Equipartition Property)与弱大数定理(the weak law of large numbers)等价

3.2.2 DMS 渐近均衡性

例：抛硬币正面出现（用0表示）的概率为p，如果试验次数N足够大，由弱大数定理可知

$$P_r\left\{\left|\frac{N(0)}{N}-p\right|>\varepsilon\right\}\le\delta$$

其中δ随着N的增大可变得任意小，

而ε是任意小的正数.

3.2.2 DMS 渐近均衡性

例（续）：序列**u**出现的概率为

$$p(\mathbf{u}) = p^{N(0)}(1-p)^{N-N(0)}$$

平均每个符号的自信息量为：

$$\frac{1}{N}\log\frac{1}{p(\mathbf{u})} = -\frac{N(0)}{N}\log p - \left[1-\frac{N(0)}{N}\right]\log(1-p)$$

当N足够大时，N(0)/N→p, 上式→H(u)

$$P_r\left\{\left|\frac{1}{N}\log\frac{1}{p(\mathbf{u})} - H(u)\right| > \varepsilon\right\} \le \delta$$

3.2.2 DMS 渐近均衡性

例：(续)

可以认为典型序列是长度为N的序列中，出现正/反面的比例为p和1−p的序列

若p=0.25, 则H(u)=0.81,

取ε=0.05, N=100

$$\frac{|S(N,\varepsilon)|}{2^N} \le \frac{2^{N[H(u)+\varepsilon]}}{2^N} \approx 6.1\times 10^{-5}$$

可见S(N,ε)中序列数目所占比例并不大，

但是随着$N \to \infty$，

$$P_r\{\mathbf{u} \in S(N,\varepsilon)\} \to 1, \quad P_r\{\mathbf{u} \in \overline{S(N,\varepsilon)}\} \to 0$$

注意个别非典型序列出现的概率可能高于典型序列！

3.2 无失真信源编码定理

- 引理3.1：Chebyshev Inequality
- 引理3.2：离散无记忆信源的渐近均衡性
- 定理3.1：无失真信源编码定理

3.2.3 无失真信源编码定理

定理3.1: 给定离散无记忆信源,

它的符号取值于集合U, 信源熵是H(u).

对于长度为N的信源符号串(N=1,2,…),

存在唯一可译的二进制码,

它的平均码长 $\langle L_N \rangle = \sum_{\mathbf{u}} p_N(\mathbf{u}) \cdot l_N(\mathbf{u})$

$\langle L_N \rangle$满足以下不等式:

$$NH(u) \le \langle L_N \rangle < N[H(u) + 0(N)]$$

(当$N \to \infty$时, $0(N) \to 0$)

3.2.3 无失真信源编码定理

证明：这里只证明$\langle L_N \rangle \leq N[H(u)+0(N)]$

利用信源渐进均衡性引理的结论.

（1）对于$\mathbf{u}\in S(N,\varepsilon)$

用长为L_N（$L_N=\lceil NH(u)+N\varepsilon \rceil$）的二进制码来表示,

它是唯一可译的二进制码

码前冠以“0”代表其后的L_N位码表示的$\mathbf{u}\in S(N,\varepsilon)$,

总码长为L_N+1

则 $L_N+1<N[H(u)+\varepsilon]+2$

3.2.3 无失真信源编码定理

证明：（2）对于$\mathbf{u} \in \overline{S(N,\varepsilon)}$

用长为L'_N($L'_N<\log_2(r^N)$)的二进制码表示

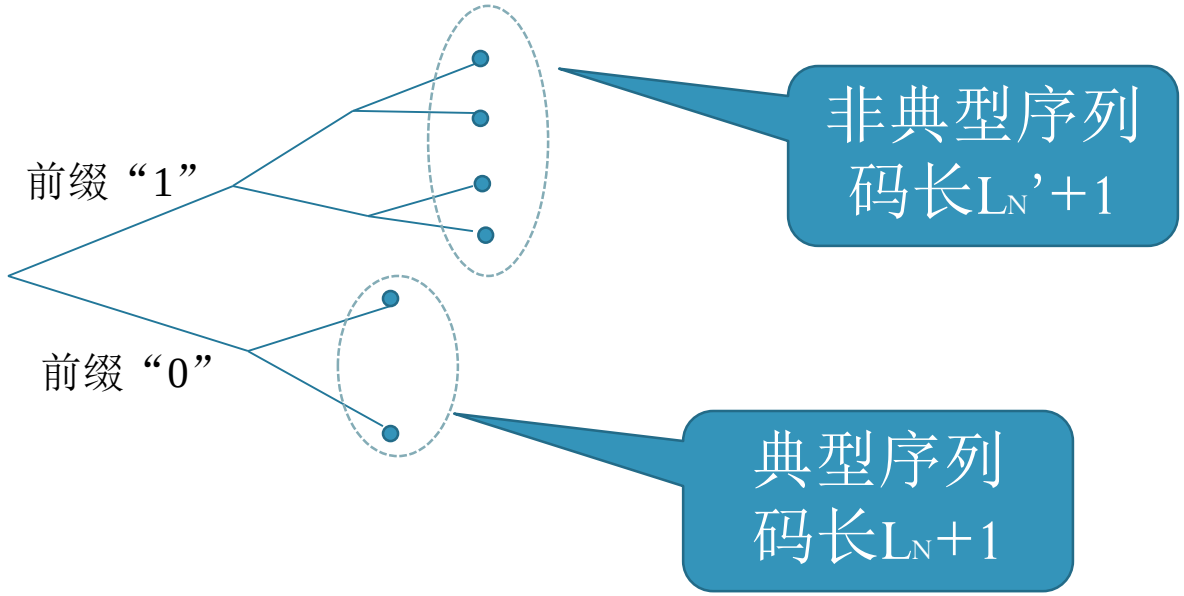
这里规定每个信源符号有r种可能的取值

它是唯一可译的二进制码

码前冠以“1”表示其后的L'_N位码表示 $\mathbf{u} \in \overline{S(N,\varepsilon)}$

总码长为 L'_N+1

则 $L'_N+1< N \log r +1$


前缀“1”
前缀“0”
非典型序列
码长$L_N'+1$
典型序列
码长L_N+1

3.2.3 无失真信源编码定理

证明: (cont.)因此，对于信源输出的长为N的每个符号串，都可以用一个唯一可译二进制码表示, 并且

$$\langle L_N \rangle = (L_N + 1)P_r\{\mathbf{u} \in S(N,\varepsilon)\} + (L'_N + 1)P_r\{\mathbf{u} \in \overline{S(N,\varepsilon)}\}$$

$$< [N(H(u)+\varepsilon)+2]\cdot 1 + [N\log r + 1]\cdot\frac{\sigma^2}{N\varepsilon^2}$$

$$= N\left[H(u)+\varepsilon+\frac{2}{N}+\left(\log r+\frac{1}{N}\right)\cdot\frac{\sigma^2}{N\varepsilon^2}\right]$$

3.2.3 无失真信源编码定理

证明: 取$\varepsilon = N^{-\frac{1}{3}}$

$$\langle L_N \rangle < N\left\{H(u)+\frac{2}{N}+\left[\left(\log r+\frac{1}{N}\right)\sigma^2+1\right]\frac{1}{N^{\frac{1}{3}}}\right\}$$

$$= N\left[H(u)+0(N)\right]$$

即 $\langle L_N \rangle < N\left[H(u)+0(N)\right]$

3.2.3 无失真信源编码定理

- The strong converse source coding theorem:

  As long as we require probability of error strictly less than 1, asymptotically, we cannot encode at rates below the entropy.

- The weak converse source coding theorem:

  Error probability cannot vanish if the compression rate is below the entropy.

目录

- 引言
- 无失真信源编码定理
- 非等长信源编码
- 一种实用编码方法:LZ编码

3.3 非等长信源编码

- 基本概念
- McMillan不等式
- Kraft不等式
- 非等长信源编码的平均码长
- 几种非等长信源编码方案

3.3.1 基本概念

- 例：信源符号集Au={0, 1, 2, 3}

  相应概率分布**p**={1/2, 1/4,1/8, 1/8}

  则H(U)=H_2(1/2, 1/4, 1/8, 1/8)=1.75bits

Au	p	编码 A	编码 B	编码 C	编码 D
0	1/2	0	0	0	0
1	1/4	0	1	01	10
2	1/8	1	00	011	110
3	1/8	10	11	0111	111
平均码长（bits)		1.125	1.25	1.875	1.75

编码D的平均码长：

<L>=1.75bits

效率：η=100%

3.3.1 基本概念

- 编码准则：
  - 出现概率大的信源符号对应短码
  - 出现概率小的信源符号对应长码

  例：Morse电报码，e对应最短码
- 编码效率： $\eta = H_s(U)/\langle L_s \rangle$
- 与等长信源编码比较：
  - 优点：编码效率高
  - 缺点：译码复杂

3.3.1 基本概念

- 异前缀码
  - 任何一个码字不允许是另一个码字的前缀
  - 满足唯一可译码要求

3.3 非等长信源编码

- 基本概念
- McMillan不等式
- Kraft不等式
- 非等长信源编码的平均码长
- 几种非等长信源编码方案

3.3.2 唯一可译码的约束不等式——McMillan不等式

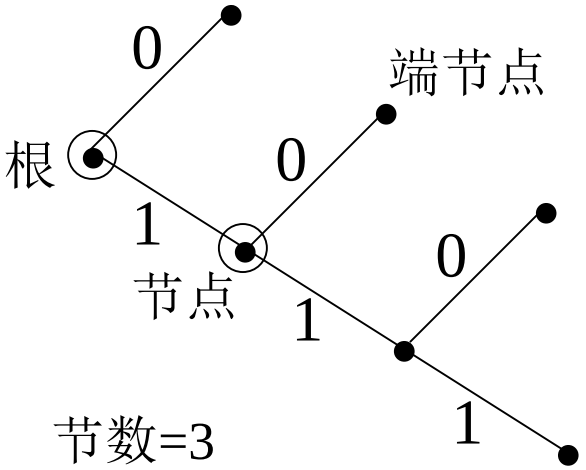
定理3.2: 如果C={$\sigma_0, \sigma_1, \ldots, \sigma_{r-1}$}是S进制的唯一可译码，其中$\sigma_i$表示第i个码字，$n_i = |\sigma_i|$是它的长度，

则

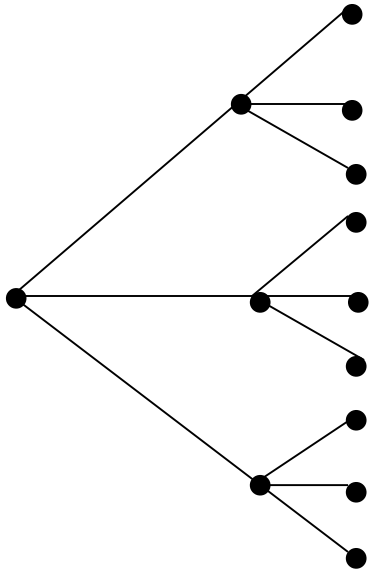
$$\sum_{i=0}^{r-1} s^{-n_i} \le 1$$

3.3.2 唯一可译码的约束不等式——McMillan不等式

码树：



二进制码树



三进制码树

满树：s进制N节码树的所有枝都被使用时，共有s^N个码字

3.3.2 唯一可译码的约束不等式——McMillan不等式

- 利用码树说明异前缀码满足McMillan不等式：
  - 设n_i中最大的值为N
  - 长度为n_i的异前缀码在N节满树上占用一个n_i节端节点
  - 从该端节点开始，到第N节上有s^{N-ni}个枝不能再用，否则就不是异前缀码
  - 总共不用的枝数为 $\sum_{i=0}^{r-1} s^{N-ni}$
  - 而N节满树的第N节上的总枝数为s^N
  - 必有 $\sum_{i=0}^{r-1} s^{N-ni} \le s^N$
  - 因而 $\sum_{i=0}^{r-1} s^{-ni} \le 1$

3.3 非等长信源编码

- 基本概念
- McMillan不等式
- Kraft不等式
- 非等长信源编码的平均码长
- 几种非等长信源编码方案

3.3.3 唯一可译码的约束不等式——Kraft不等式

- 定理3.3: 如果$\sum_{i=0}^{r-1} s^{-n_i} \leq 1$, 则存在一个唯一可译码, 相应的码长是$n_i$

3.3.3 唯一可译码的约束不等式——Kraft不等式

- 证明：假设 $n_0 \le n_1 \le n_2 \le \ldots \le n_{r-1}$.

  定义r个S进制小数 w_j :

  $w_0 = 0$ (n_0位),

  $w_j = \sum_{i=0}^{j-1} s^{-n_i}$ (n_j位) $1 \le j \le r-1$.

  则构造了一个S进制**异前缀码**w_j ,

  $j = 0, 1, \ldots, r-1$

3.3.3 唯一可译码的约束不等式——Kraft不等式

- 例题：s=3, $\{n_i\}=\{1, 1, 2, 2, 3, 3, 4, 4, 4\}$

满足 $\sum_{i=0}^{r-1} s^{-n_i} \le 1$

$w_0 = 0(n_0位)$

$w_j = \sum_{i=0}^{j-1} s^{-n_i} (n_i位)$

$1 \le j \le r-1$

$n_0=1$	$w_0=0$	0
$n_1=1$	$w_1=1/3$	1
$n_2=2$	$w_2=2/3$	20
$n_3=2$	$w_3=2/3+1/9$	21
$n_4=3$	$w_4=2/3+2/9$	220
$n_5=3$	$w_5=2/3+2/9+1/27$	221
$n_6=4$	$w_6= 2/3+2/9+2/27$	2220
$n_7=4$	$w_7= 2/3+2/9+2/27+1/81$	2221
$n_8=4$	$w_8= 2/3+2/9+2/27+2/81$	2222

3.3.3 唯一可译码的约束不等式——Kraft不等式

- 证明(续）：此码具有下面的性质，当k>j时，

$$\left(w_k - w_j\right)\cdot s^{n_j} = \left(\sum_{i=0}^{k-1} s^{-n_i} - \sum_{i=0}^{j-1} s^{-n_i}\right)\cdot s^{n_j} = s^{n_j}\cdot\sum_{i=j}^{k-1} s^{-n_i} \geq 1$$

反证法：如果不是S进制异前缀码

假设当k>j时，w_j是w_k的前缀

有 $\left(w_k - w_j\right)\cdot s^{n_j} < 1$

与前面的性质矛盾，假设不成立

异前缀码是一种唯一可译码，因此得证.

3.3 非等长信源编码

- 基本概念
- McMillan不等式
- Kraft不等式
- 非等长信源编码的平均码长
- 几种非等长信源编码方案

3.3.4 非等长信源编码的平均码长

- 定理3.4: 如果C是唯一可译码, 它的S进制编码的平均长度一定大于等于信源的S进制的熵. 即

$$\overline{n} \geq H_s(\mathbf{p}) = \sum_{i=0}^{r-1} p_i \log_s \frac{1}{p_i}$$

3.3.4 非等长信源编码的平均码长

- 证明：

$$H_s(\mathbf{p})-\bar{n}=\sum_{i=0}^{r-1}p_i\left[\log_s\frac{1}{p_i}-n_i\right]$$

$$=\sum_{i=0}^{r-1}p_i\left[\log_s\frac{1}{p_i}-\log_s s^{n_i}\right]=\sum_{i=0}^{r-1}p_i\left[\log_s\frac{s^{-n_i}}{p_i}\right]$$

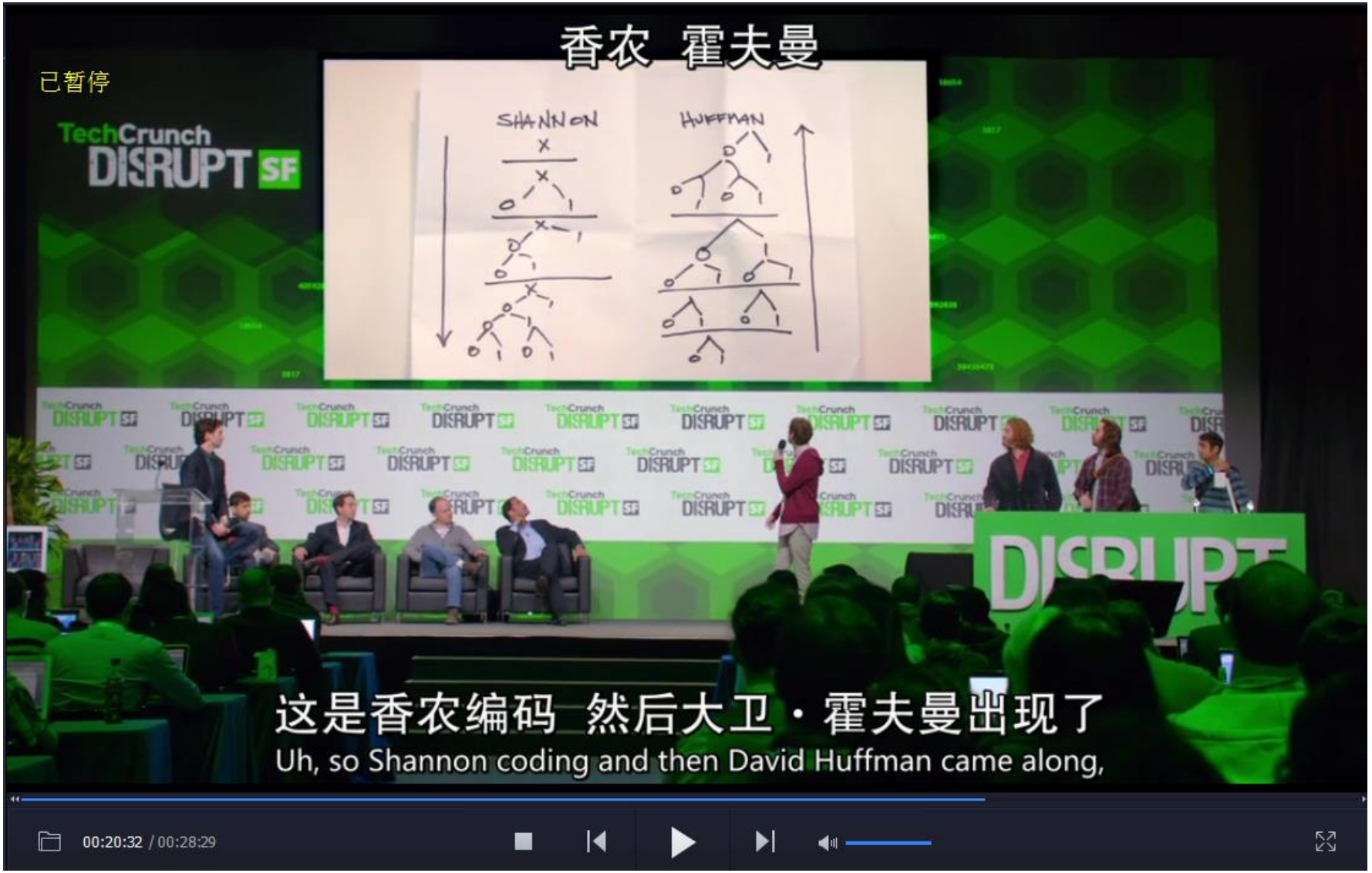
$$\le\log_s\left[\sum_{i=0}^{r-1}p_i\frac{s^{-n_i}}{p_i}\right]$$

$$=\log_s\left[\sum_{i=0}^{r-1}s^{-n_i}\right]\le\log_s 1=0$$

3.3 非等长信源编码

- 基本概念
- McMillan不等式
- Kraft不等式
- 非等长信源编码的平均码长
- 几种非等长信源编码方案

美剧 Silicon Valley



3.3.5 几种非等长信源编码方案

- 非等长信源编码方案主要有：
  - Shannon编码
  - Fano编码
  - Shannon-Fano-Elias编码
  - Huffman编码

3.3.5几种非等长信源编码方案
____Shannon编码

- 证明Kraft不等式时引入的编码方法被称为Shannon编码
- Shannon编码选择每个码字长度n_i为：

$$n_i = \lceil \log_s p_i^{-1} \rceil \qquad (i=1, 2, \ldots, r-1)$$

- 平均码长 $\bar{n} \le H_s(\mathbf{p}) + 1$

⌘一般情况下， Shannon编码的平均码长并不是最短的

3.3.5几种非等长信源编码方案

____Shannon编码

- 证明：　取$n_i = \lceil \log_s p_i^{-1} \rceil$

  即 $\log_s p_i^{-1} \le n_i < \log_s p_i^{-1} + 1,\ \ i=1, 2, \ldots, r-1.$

  由$\log_s p_i^{-1} \le n_i$

  有 $s^{-ni} \le p_i$,

$$\sum_{i=0}^{r-1} s^{-n_i} \le \sum_{i=0}^{r-1} p_i = 1$$

  故存在唯一可译码, 它的长度是n_i

3.3.5几种非等长信源编码方案
____Shannon编码

- 证明：

由 $n_i < \log_s p_i^{-1} + 1,$

有 $$\sum_{i=0}^{r-1} p_i n_i \le \sum_{i=0}^{r-1} p_i \log \frac{1}{p_i} + \sum_{i=0}^{r-1} p_i$$

$$\bar{n} \le H_s(\mathbf{p}) + 1$$

3.3.5 几种非等长信源编码方案

- 非等长信源编码方案主要有：
  - Shannon编码
  - Fano编码
  - Shannon-Fano-Elias编码
  - Huffman编码

3.3.5几种非等长信源编码方案
____Fano编码

- 方法(以二进制编码为例）：
  - 将信源符号以概率递减的次序排列
  - 将排好的信源符号划分为两组，使每组概率和近于相等，并分别编码为“0”和“1”
  - 进一步将每组中的信源符号再分为两组，使其概率和近于相等，并分别编码为“0”和“1”
  - 依次下去…
  - 直至每组只剩一个信源符号为止

3.3.5几种非等长信源编码方案

____ Fano编码

- 例题：

信源符号	概率p_i	编码过程				码字	码长n_i
a_1	0.32	0	0			00	2
a_2	0.22		1			01	2
a_3	0.18	1	0			10	2
a_4	0.16		1	0		110	3
a_5	0.08			1	0	1110	4
a_6	0.04				1	1111	4

3.3.5几种非等长信源编码方案
____Fano编码

- Fano编码实际上是构造码树的一种方法
- 它需要考虑信源的统计特性
- 这种编码方法的结果不一定是唯一的
- 有可能没有充分利用短码，从而增加了平均码长
- Fano编码的平均码长：

$$\overline{n} \le H_s(\mathbf{p}) + 2$$

3.3.5 几种非等长信源编码方案

- 非等长信源编码方案主要有：
  - Shannon编码
  - Fano编码
  - Shannon-Fano-Elias编码
  - Huffman编码

3.3.5几种非等长信源编码方案
Shannon-Fano-Elias编码

- 根据信源符号的累积分布函数来确定码字
- 定义累积分布函数：

$$F(a_k)=\sum_{i=1}^{k}p(a_i)\quad a_i,a_k\text{为信源符号}$$

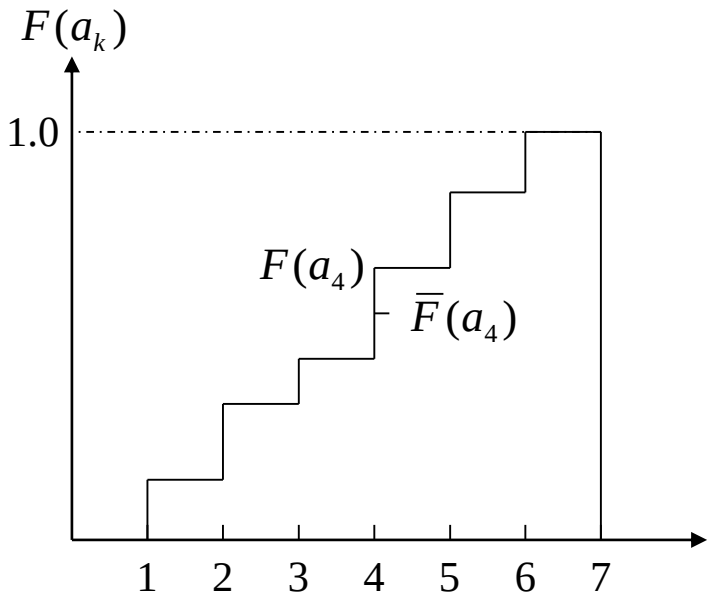
定义修正的累积分布函数：

$$\overline{F}(a_k)=\sum_{i=1}^{k-1}p(a_i)+\frac{1}{2}p(a_k)\quad a_i,a_k\text{为信源符号}$$

当j≠k时

$$F(a_j)\neq F(a_k)$$

$$\overline{F}(a_j)\neq \overline{F}(a_k)$$



3.3.5几种非等长信源编码方案
Shannon-Fano-Elias编码

- 因此可以利用 $\bar{F}(a_k)$的数值作为符号a_k的码字
- 取$n_k = \lceil \log p_k^{-1} \rceil + 1$，将表示$\bar{F}(a_k)$的二进制数截短为$n_k$位，则得到$a_k$的码字
- 这种编码方法的平均码长：

$$H_s(\mathbf{p}) + 1 \le \bar{n} < H_s(\mathbf{p}) + 2$$

- 优点：不要求信源符号按概率大小次序排列

3.3.5几种非等长信源编码方案
____ Shannon-Fano-Elias编码

- 例题

信源符号	概率p_i	累积分布函数$F(s_i)$	$\overline{F}(s_i)$	$\overline{F}(s_i)$ 的二进制数	码长n_i	码字
a1	0.25	0.25	0.125	0.001	3	001
a2	0.5	0.75	0.5	0.10	2	10
a3	0.125	0.875	0.8125	0.1101	4	1101
a4	0.125	1.0	0.9375	0.1111	4	1111

3.3.5几种非等长信源编码方案
____ Shannon-Fano-Elias编码

- 例题（续）：

  平均码长 $\bar{n} = 2.75 bits$

  信源的熵 H(S)=1.75*bits*

  而Shannon编码和Fano编码，以及将要介绍的Huffman编码都能达到该信源的熵

- Shannon-Fano-Elias编码虽然不是最佳的，但是由它发展而来的算术编码方法，目前广泛应用于image和video的数据压缩

3.3.5 几种非等长信源编码方案

- 非等长信源编码方案主要有：
  - Shannon编码
  - Fano编码
  - Shannon-Fano-Elias编码
  - Huffman编码

3.3.5介绍非等长信源编码方案 ____ Huffman编码

- D. Huffman在1952年提出
- 最佳非等长信源编码方法
- 在不同领域得到广泛应用
  - 例1-国际数字传真编码标准所采用的MH码是一种改进的Huffman编码
  - 例2-美国HDTV

3.3.5几种非等长信源编码方案
____ Huffman编码

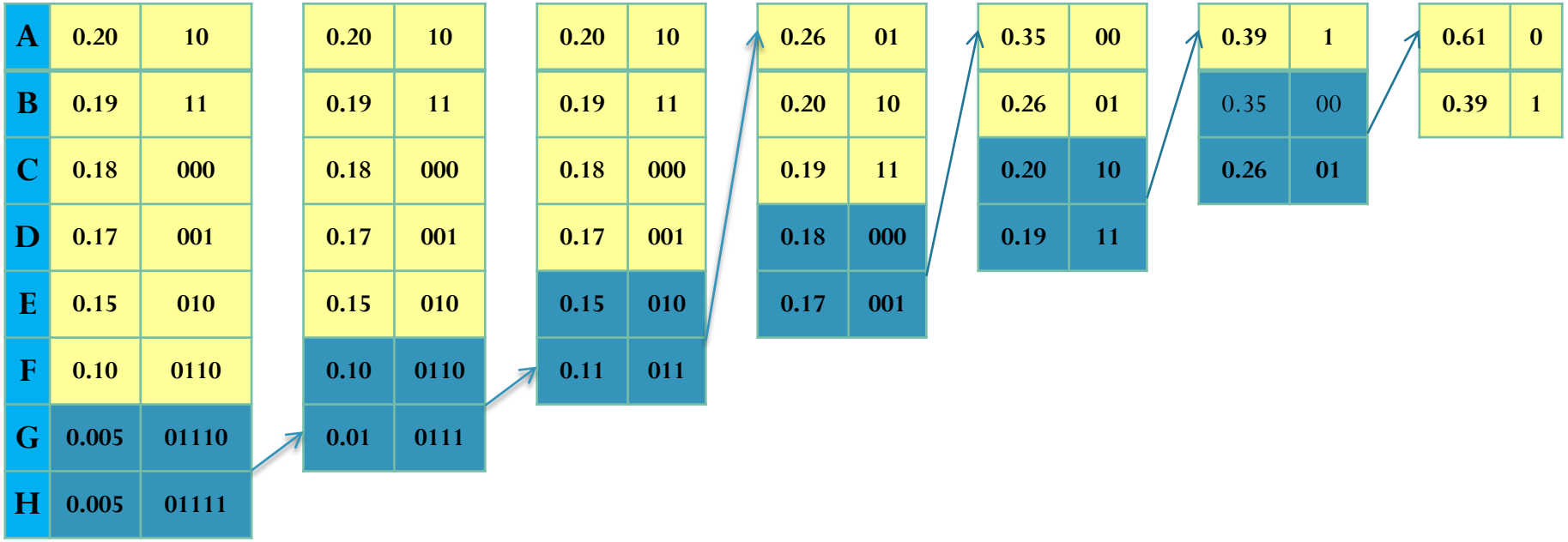
- 例3.2: 设信源符号有8种字母，相应概率如下，试进行二进制和三进制Huffman编码.

X	P(x)
A	0.20
B	0.19
C	0.18
D	0.17
E	0.15
F	0.10
G	0.005
H	0.005

3.3.5几种非等长信源编码方案

Huffman编码

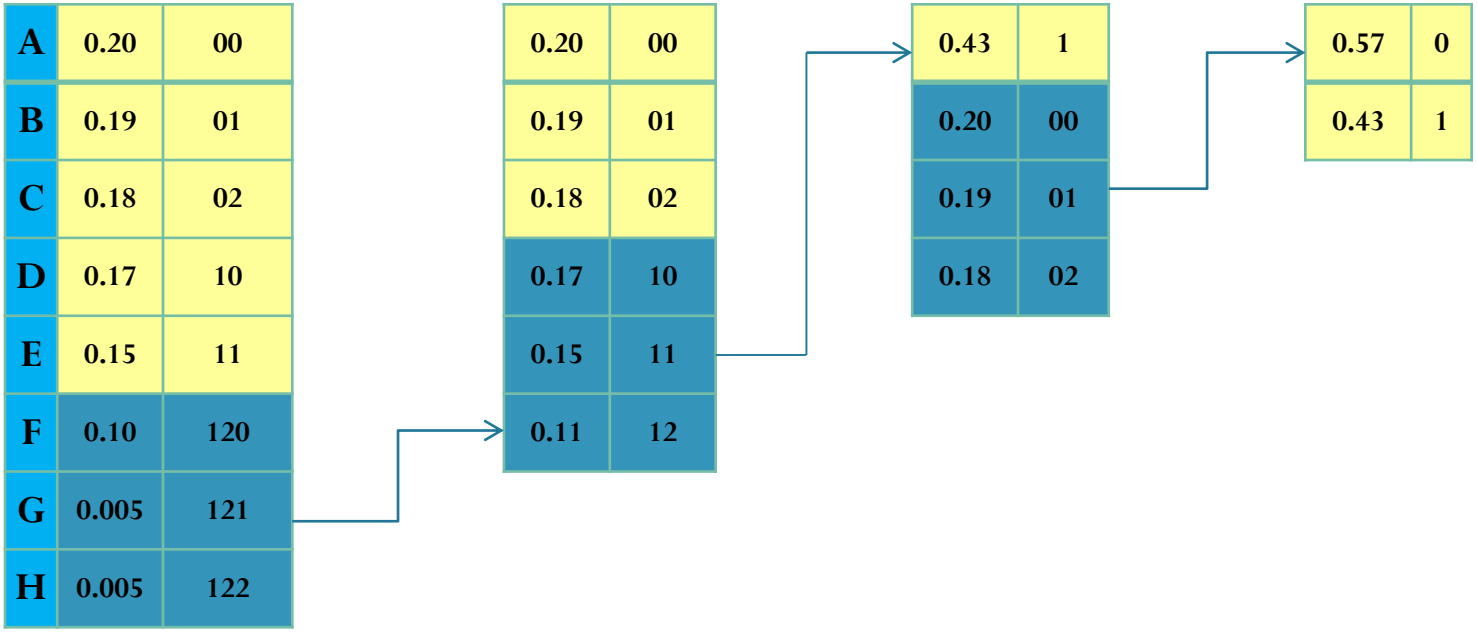
- 二进制Huffman编码：



3.3.5几种非等长信源编码方案

Huffman编码

- 三进制Huffman编码：（错误！）

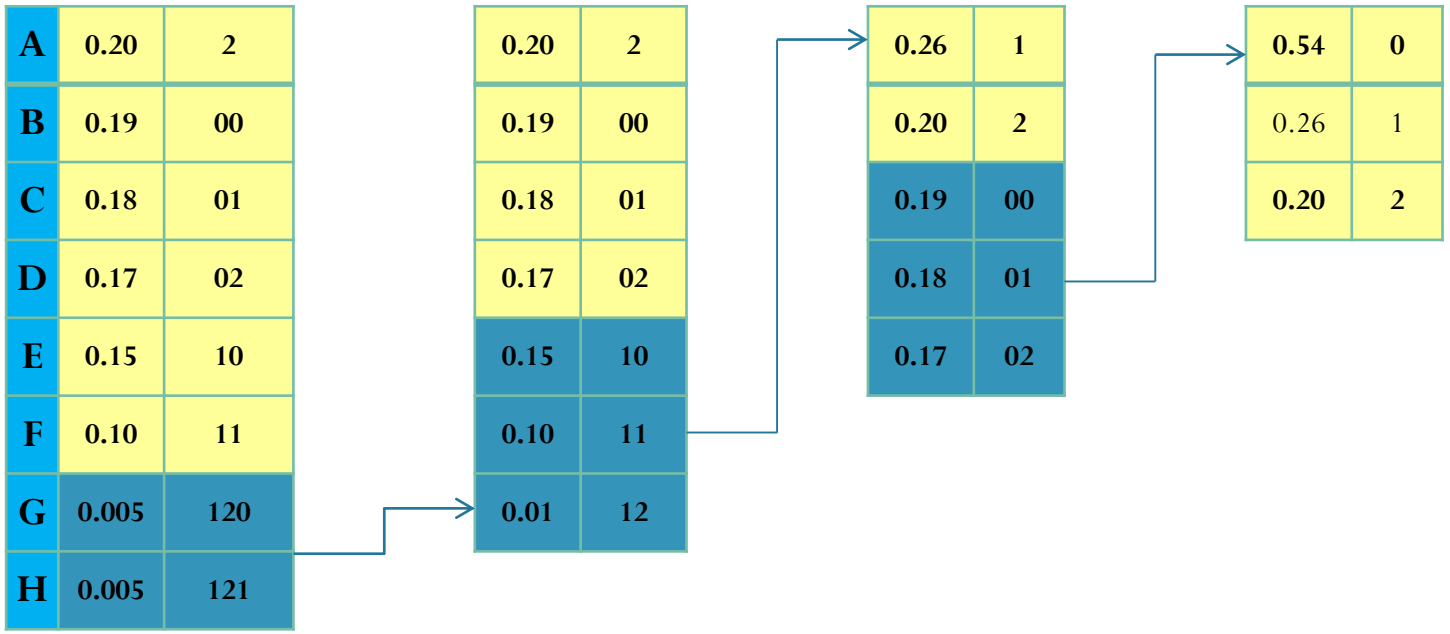


最短码“2”没被采用！

3.3.5几种非等长信源编码方案

Huffman编码

- 三进制Huffman编码：（正确！）



3.3.5几种非等长信源编码方案
____ Huffman编码

- 编码要求:

(1) 信源符号出现的概率越大, 编码后的码字越短
信源符号出现的概率越小, 编码后的码字越长

(2)S进制编码, 第一次合并的符号数s'={2, 3, …,s}

$s' = r \bmod (s-1)$

(3) 概率最小的s'个码字, 区别仅在于最后一位

3.3.5几种非等长信源编码方案
____Huffman编码

- 问题：

(1)要求了解信源的统计分布

(2)算法复杂度随着信源符号串长度的增加而迅速增长

目录

- 引言
- 无失真信源编码定理
- 非等长信源编码
- 一种实用编码方法:LZ编码

3.4一种实用编码方法:LZ编码

- 由A. Lempel和J. Ziv在1976-1978年期间提出
- 一种通用信源编码方法(Universal Source Coding)
  - 不需要了解信源的统计特性
- 属于变长⇒定长的信源编码方法
- 广泛应用于计算机文件的数据压缩

[1] J. Ziv and A. Lempel, ``A Universal Algorithm for Sequential Data Compression," *IEEE Transactions on Information Theory*, Vol. 23, pp. 337--342, 1977.
[2] J. Ziv and A. Lempel, ``Compression of Individual Sequences Via Variable-Rate Coding," *IEEE Transactions on Information Theory*, Vol. 24, pp. 530--536, 1978.

3.4一种实用编码方法:LZ编码

- 方法：
  - 将离散信源的输出序列分解成长度可变的分组，称为**码段(phrases)**
  - 每当信源输出字符组在最后位置加上一个字符后与前面已有码段都不相同时，把它作为一种新的码段引入…

【例】

信源输出序列：

101011010010011101010000110011101011000110 11…

分解为码段：

1,0,10,11,01,00,100,111,010,1000,011,001,110,101,10001,1011,…

3.4一种实用编码方法:LZ编码

- 方法：
  - 这些码段列入一个位置字典，用于记载已有码段的位置
  - 在对一个新的码段编码时，只要指出字典中现有码段的位置，把新字符附在后面
  - 该码的信源解码器在通信系统的接收端构造一个完全相同的表，对接收序列作相应的解码

	字典位置	字典内容	码字
1	0001	1	00001
2	0010	0	00000
3	0011	10	00010
4	0100	11	00011
5	0101	01	00101
6	0110	00	00100
7	0111	100	00110
8	1000	111	01001
9	1001	010	01010
10	1010	1000	01110
11	1011	011	01011
12	1100	001	01101
13	1101	110	01000
14	1110	101	00111
15	1111	10001	10101
16		1011	11101

3.4一种实用编码方法:LZ编码

- 如何选择码表的总长度？
  - 一般而言，无论表有多大，总会溢出……
  - 为此，信源编解码器必须达成一致，将无用的码段从各自的字典中删去，在它们留下的位置上换上新的码段

3.4一种实用编码方法:LZ编码

The following table compares an adaptive version of the Huffman code (implemented in the Unix ``compact'' program) and an implementation of the Lempel-Ziv algorithm (Unix “compress” program).

	Adaptive Huffman	Lempel-Ziv
LaTeX file	66%	44%
Speech file	65%	64%
Image file	94%	88%
Size of compressed file as percentage of the original file		

练习

1. (15分)信源输出符号A1，A2，…，A10的概率分别为1/8, 1/8, 1/8, 1/8, 1/10, 1/10, 1/10, 1/10, 1/20, 1/20. 试给出三进制Huffman编码，并计算平均码长和编码效率.
(注：要求计算出结果，保留三位有效数字.
参考数据：lg2≈0.301, lg3≈0.477, lg5≈0.699)

作业：习题3.1，3.2，3.3