



电子科技大学

University of Electronic Science and Technology of China

信息论导论

第三章

离散信源及其信息熵

但黎琳

通信抗干扰技术国家级重点实验室

授课内容



无失真编码的基本思路

无失真编码定理

香农编码

霍夫曼编码

费诺编码

信源编码-概述

信源发出消息需要传输或存储才能发挥作用。

为了满足信道特性，往往需要将 n 元的信源符号序列变换为 m 元的信源码序列，这一过程就是信源编码。

而对于传输或存储来说，现在二进制码元（**bit**）最具优势。

信源编码-概述

- 信息传输的3个关键要求：
有效性、可靠性、安全性
- 信源编码是为了提高通信的有效性为目的的编码。
- 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或信源的码率。
- 同样多的信息用较少的比特数来传送，使单位时间内传送的平均信息量增加，从而提高通信的有效性。

信源编码-概述

- 信源编码的基础是信息论中的两个编码定理：
 - 无失真编码定理
 - 限失真编码定理
- 无失真编码只适用于离散信源；对于连续信源，只能在失真受限制的情况下进行限失真编码

本章介绍离散平稳信源的编码定理和方法。

信源编码-概述

信源编码的3个问题是：

- ①怎样保证编码和译码过程不丢失信息，即怎样实现无失真编码？
- ②怎样有效地利用码字的每一个比特去携带信息，即在编码过程中怎样用最少的比特数去表示信源的信息熵？
- ③无失真编码的具体方法？

无失真编码的基本思路

无失真编码的含义：保证符号序列与码字一一对应，此时用码序列表示的信源信息熵保持不变。

$$\text{例：} \begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1 & x_2 & x_3 & x_4 \\ 0.5 & 0.3 & 0.15 & 0.05 \end{Bmatrix}$$

最简单的无失真编码：**4-2**进制变换，即

$$x_1 \rightarrow 00, x_2 \rightarrow 01, x_3 \rightarrow 10, x_4 \rightarrow 11$$

无失真编码的基本思路

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1 \rightarrow 00 & x_2 \rightarrow 01 & x_3 \rightarrow 10 & x_4 \rightarrow 11 \\ 0.5 & 0.3 & 0.15 & 0.05 \end{Bmatrix}$$

其信息熵

$$H(X) = -\sum_{i=1}^4 P(x_i) \lg P(x_i) \approx 1.648(\text{bit} / \text{symbol})$$

该编码的码长 **$K=2$** ，码字的每一个比特携带信息的效率即编码效率

$$\eta = \frac{H(X)}{K} = \frac{1.648}{2} \approx 82.4\%$$

思考：效率可以再提高吗？

无失真编码的基本思路

- 码字的比特数中约有**17.6%**未携带信息，属于冗余比特，传输效率不高。
- 信源无失真编码通过变换信源输出序列，减少信源输出符号序列的剩余度，提高符号的平均信息量。
- 编码目标：
 - 保证无失真恢复原符号序列
 - 码字尽量短，码元承载的平均信息量尽量大

无失真编码的基本思路

为了压缩比特数，可以考虑对信源符号进行不等长编码，如

$$x_1 \rightarrow 0, x_2 \rightarrow 1, x_3 \rightarrow 00, x_4 \rightarrow 01$$

可以无失真译码吗？

但该编码不能实现无失真译码，即不能保证符号元与码字的一一对应。

无失真编码的基本思路

一种能保证符号元与码字一一对应的不等长编码为 $x_1 \rightarrow 0, x_2 \rightarrow 10, x_3 \rightarrow 110, x_4 \rightarrow 111$

其平均码长

$$\bar{K} = 0.5 \times 1 + 0.3 \times 2 + 0.15 \times 3 + 0.05 \times 3 = 1.7$$

编码效率

$$\eta = \frac{H(X)}{\bar{K}} = \frac{1.648}{1.7} \approx 96.9\%$$

与第一种编码相比，码字压缩了**0.3**个比特，编码效率提高了**14.5%**。

无失真编码的基本思路

进一步，如果对该信源的二次扩展信源

$$\begin{bmatrix} X^2 \\ P(X^2) \end{bmatrix} = \left\{ \begin{array}{cccccccc} x_1x_1 & x_1x_2 & x_1x_3 & x_1x_4 & x_2x_1 & x_2x_2 & x_2x_3 & x_2x_4 \\ 0.25 & 0.15 & 0.075 & 0.025 & 0.15 & 0.09 & 0.045 & 0.015 \\ x_3x_1 & x_3x_2 & x_3x_3 & x_3x_4 & x_4x_1 & x_4x_2 & x_4x_3 & x_4x_4 \\ 0.075 & 0.045 & 0.0225 & 0.0075 & 0.025 & 0.015 & 0.0075 & 0.0025 \end{array} \right\}$$

进行无失真编码，一种能保证符号序列元与码字一一对应的不等长编码为

$$x_1x_1 \rightarrow 00, x_1x_2 \rightarrow 100, x_1x_3 \rightarrow 1100, x_1x_4 \rightarrow 11100$$

$$x_2x_1 \rightarrow 101, x_2x_2 \rightarrow 010, x_2x_3 \rightarrow 0110, x_2x_4 \rightarrow 111100$$

无失真编码的基本思路

$x_3x_1 \rightarrow 1101, x_3x_2 \rightarrow 0111, x_3x_3 \rightarrow 111110, x_3x_4 \rightarrow 1111110$

$x_4x_1 \rightarrow 11101, x_4x_2 \rightarrow 111101, x_4x_3 \rightarrow 11111110, x_4x_4 \rightarrow 11111111$

其平均码长

$$\begin{aligned}\bar{K} &= \sum_{i=1}^{16} P(a_i)k_i = 0.25 \times 2 + 0.15 \times 3 + \dots \\ &+ 0.0075 \times 8 + 0.0025 \times 8 \approx 3.328\end{aligned}$$

编码效率

$$\eta = \frac{H(X^2)}{\bar{K}} = \frac{2H(X)}{\bar{K}} = \frac{3.296}{3.328} \approx 99.0\%$$

无失真编码的基本思路-主要问题

与第二种编码相比，码字又压缩了约**0.04**个比特，编码效率提高了**2.1%**。

总结该例子，有以下几点结论与问题：

①一般采用不等长编码，使平均码长接近信息熵，从而在无失真前提下提高编码效率；编码的基本原则是大概率符号元编成短码，小概率符号元编成长码。

无失真编码的基本思路-主要问题

②由于码长不等，如何保证接收端从码序列中唯一地分割出对应与每一个符号元的码字，以实现无失真译码？

③对符号序列进行组(**block**)编码有助于使平均码长接近信息熵，但平均码长能否无限接近信息熵，从而使编码效率趋近1？如果能，对序列长度有什么要求？

授课内容



无失真编码的基本思路



无失真编码定理



香农编码



霍夫曼编码



费诺编码

二、无失真编码定理

定义:无须考虑后续的码符号,即可从码符号序列中译出码字,这样的唯一可译码称为即时码。

- 设一个码字 $\bar{W}_i = W_{i_1}W_{i_2} \cdots W_{i_l}$, 对于任意的 $1 \leq j \leq l$, 称码符号序列的前 j 个元素 $W_{i_1}W_{i_2} \cdots W_{i_j}$ 为码字 $\bar{W}_i$ 的前缀。

二、无失真编码定理

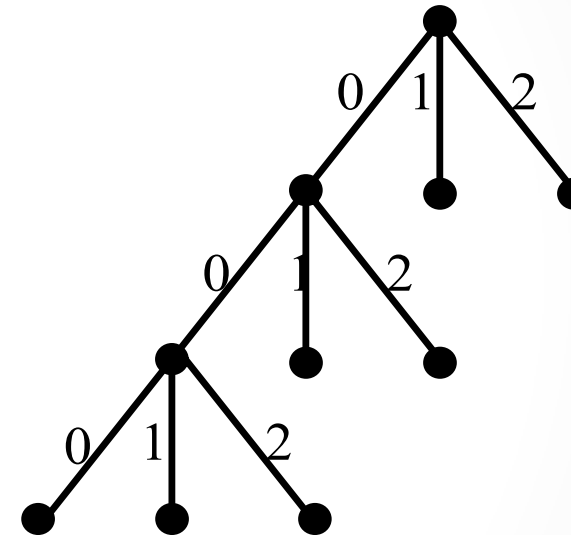
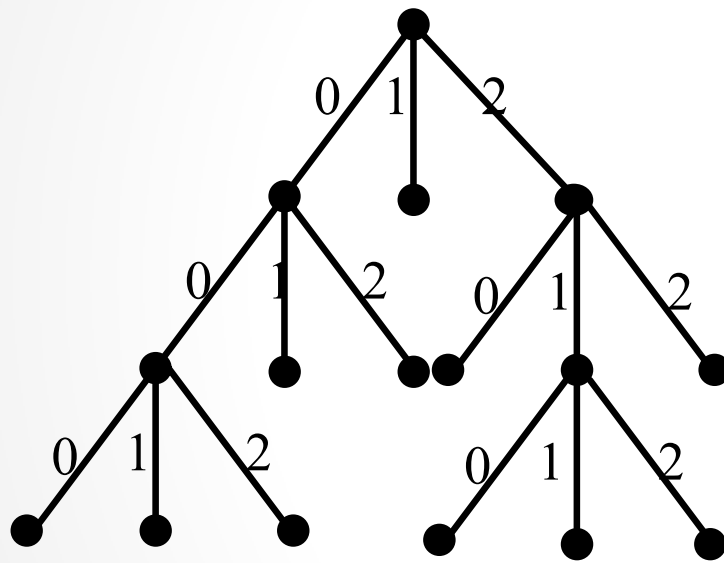
- 异前置码是一种单义可译即时码
- 如果所采用的不等长编码使接收端能从码序列中唯一地分离出对应与每一个符号元的码字，则称该不等长编码为单义可译码
- 单义可译码中，如果能在对应于每一个符号元的码字结束时立即译出的称为即时码，如果要等到对应与下一个符号元的码字才能译出的称为延时码。

例 判断以下码字是否可分离？

消息	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
		多译码	多译码	单义可译 延时码	单义可译 即时码 (异前置码)

异前置码-特点

异前置码的特点是：任何一个码字都不是其他码字的前缀，因此将该码称为异前置码。

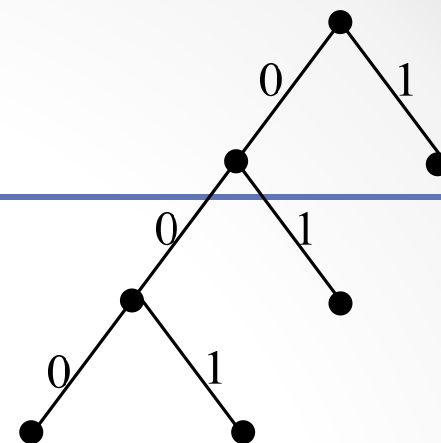


000,001,002,01,02,1,2

异前置码是从树根开始到每一个终节点的联枝
代表一个码字

克拉夫特(Kraft)不等式

码d所对应的二元码树图



由 n 个 m 元、长度分别为 k_i ($i=1,2,\dots,n$)的异前置码存在的充分必要条件是：

$$\sum_{i=1}^n m^{-k_i} \leq 1$$

该充要条件称为克拉夫特(Kraft)不等式。

克拉夫特(Kraft)不等式-证明

设 m 元异前置码第 i 个码字的长度为 k_i , $i=1,2,\dots,n$

考虑一个 N 级满树，在第 N 级共有 m^N 个节点，在第 k_i 级共有 m^{k_i} 个节点。

根据异前置码的定义，第 i 个码字后的节点不能再用，故第 N 级不能用的节点数为 m^{N-k_i}

克拉夫特(Kraft)不等式-证明

m^{N-k_i} : 即以 k_i 节点为子树根, 得到的子树在第 N 级的节点数)

构造异前置码的码树图第 N 级上总共不能用的节点总数

$$\sum_{i=1}^n m^{N-k_i} \leq m^N \quad \text{从而} \quad \sum_{i=1}^n m^{-k_i} \leq 1$$

【无失真编码定理】

如 N 维离散平稳信源的平均符号熵为 $H_N(X_1X_2\cdots X_N)$ ，对信源符号序列进行 m 元不等长组编码，一定存在一种无失真编码方法，当 N 足够大时，使得每个信源符号所对应码字的平均比特数 $\bar{K}$

$$H_N(X_1X_2\cdots X_N) \leq \frac{\bar{K}}{N} \log m < H_N(X_1X_2\cdots X_N) + \varepsilon$$

式中， ε 为任意给定的小正数。

【无失真编码定理】-证明

设不等长组编码对应于符号元 $\mathbf{a}_i = \mathbf{x}_{i1}\mathbf{x}_{i2}\dots\mathbf{x}_{iN}$ 的码字长度为 k_i

**Kraft
不等式**

取 k_i 使之满足 $m^{-k_i} \leq P(a_i)$

$$\text{由于} \sum_{i=1}^n m^{-k_i} \leq \sum_{i=1}^n P(a_i) = 1$$

说明该编码是异前置码

$$\because m^{-k_i} \leq P(a_i) \text{ 即 } m^{k_i} \geq \frac{1}{P(a_i)}$$

$$\therefore k_i \lg m \geq -\lg P(a_i), \quad \text{即 } k_i \geq -\frac{\lg P(a_i)}{\lg m}$$

【无失真编码定理】-证明

$$\text{取 } -\frac{\lg P(a_i)}{\lg m} \leq k_i < -\frac{\lg P(a_i)}{\lg m} + 1$$

其平均码长

$$-\sum_{i=1}^n P(a_i) \frac{\lg P(a_i)}{\lg m} \leq \sum_{i=1}^n P(a_i) k_i < -\sum_{i=1}^n P(a_i) \frac{\lg P(a_i)}{\lg m} + 1$$

$$\text{即 } \frac{H(X_1 X_2 \cdots X_N)}{\lg m} \leq \bar{K} < \frac{H(X_1 X_2 \cdots X_N)}{\lg m} + 1$$

$$\frac{H(X_1 X_2 \cdots X_N)}{N} \leq \frac{\bar{K}}{N} \lg m < \frac{H(X_1 X_2 \cdots X_N)}{N} + \frac{\lg m}{N}$$

【无失真编码定理】-证明

$$H_N(X_1 X_2 \cdots X_N) \leq \frac{\bar{K}}{N} l b m < H_N(X_1 X_2 \cdots X_N) + \frac{l b m}{N}$$

只要 $\varepsilon \geq \frac{l b m}{N}$

$$H_N(X_1 X_2 \cdots X_N) \leq \frac{\bar{K}}{N} l b m < H_N(X_1 X_2 \cdots X_N) + \varepsilon$$

无失真编码定理又叫香农第一定理，该定理从理论上阐明了编码效率

$$\eta = \frac{H_N(X_1 X_2 \cdots X_N)}{\bar{K} l b m / N} \rightarrow 1 \text{ 的理想无失真编码的存在性.}$$

【无失真编码定理】-说明

无失真编码的代价是取无限长的符号序列进行组编码，即只有 $N \rightarrow \infty$ 时

$$\eta = \frac{H_{\infty}}{\lim_{N \rightarrow \infty} \frac{\bar{K}}{N} \text{lb} m} = 1$$

可见，极限熵 H_{∞} 是一个界限，通常也称为(信源编码的)香农界。

刚才讨论了一般的离散平稳信源，对其他已学过的离散信源，根据无失真编码定理可得出什么结论呢？

- N 维离散平稳无记忆信源
- m 阶马尔科夫信源

N维离散平稳无记忆信源

由于其平均符号熵 $H_N(X_1X_2\cdots X_N) = H(X)$

故对信源符号序列进行 m 元不等长组编码，一定存在一种无失真编码方法，当 N 足够大时，使得每个信源符号所对应码字的平均比特数

$$H(X) \leq \frac{\bar{K}}{N} \log m < H(X) + \varepsilon$$

式中， ε 为任意给定的小正数。

此时香农界为 $H(X)$

m 阶马尔科夫信源

m 阶马尔科夫信源($m < N$), N 足够大时, 平均符号熵 $H_N(X_1 X_2 \dots X_N) = H_{m+1}$

对信源符号序列进行 m 元不等长组编码, 一定存在一种无失真编码方法, 使得每个信源符号对应码字的平均比特数

$$H_{m+1} \leq \frac{\bar{K}}{N} \log m < H_{m+1} + \varepsilon$$

式中, ε 为任意给定的小正数。

此时香农界为 H_{m+1} 。

二、无失真编码定理

平稳无记忆信源的香农界 $H(X)$ 大于 m 阶马尔科夫信源的香农界 H_{m+1} ，而 m 阶马尔科夫信源的香农界 H_{m+1} 又大于一般平稳信源的香农界 H_∞ 。

因此，对离散平稳信源进行无失真编码，每个信源符号所对应码字的平均比特数平稳无记忆信源最多， m 阶马尔科夫信源次之，一般平稳信源最少。

授课内容



无失真编码的基本思路

无失真编码定理

香农编码

霍夫曼编码

费诺编码

香农编码

香农编码是一种采用异前置码的***m***元编码方法。

设离散信源

设离散信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1 & x_2 & \cdots & x_n \\ p(x_1) & p(x_2) & \cdots & p(x_n) \end{Bmatrix}$$

设 $p(x_1) > p(x_2) > \cdots > p(x_n)$, 且 $\sum_{i=1}^n p(x_i) = 1$

香农编码的编码步骤

①将符号元 x_i 按概率进行降序排列；

②令 $p(x_0)=0$ ，计算第 i 个码字的累加概率

$$p_a(x_i) = \sum_{j=0}^{i-1} p(x_j) \quad i = 1, 2, \dots, n$$

③确定第 i 个码字的码长 k_i ， k_i 为满足下列不等式的最小整数：
 $-lbp(x_i) \leq k_i < -lbp(x_i) + 1$

④将 $p_a(x_i)$ 用二进制表示，取小数点后 k_i 位作为符号元 x_i 的码字。

单符号信源的香农编码-举例

例，对单符号离散信源编二进制香农码，并计算其编码效率。 $\left[\begin{matrix} X \\ P(X) \end{matrix} \right] = \left\{ \begin{matrix} x_1 & x_2 & x_3 & x_4 \\ 0.5 & 0.3 & 0.15 & 0.05 \end{matrix} \right\}$

解：①将 x_i 按概率进行降序排列

x_i	$p(x_i)$	$p_a(x_i)$	k_i	码字
x_1	0.5			
x_2	0.3			
x_3	0.15			
x_4	0.05			

单符号信源的香农编码-举例

②令 $p(x_0)=0$ ，计算第 i 个码字的累加概率

$$p_a(x_i) = \sum_{j=0}^{i-1} p(x_j) \quad i = 1, 2, \dots, n$$

$$p_a(x_1)=0$$

$$p_a(x_2)=0+0.5=0.5$$

$$p_a(x_3)=0.5+0.3=0.8 \quad p_a(x_4)=0.8+0.15=0.95$$

③确定第 i 个码字的码长 k_i :

$$\because -\lg p(x_1) = \lg 2 = 1 \quad \therefore k_1 = 1$$

$$\because -\lg p(x_2) = -\lg 0.3 \approx 1.74, -\lg p(x_2) + 1 = 2.74, \quad \therefore k_2 = 2$$

单符号信源的香农编码-举例

$$\because -\lg p(x_3) = -\lg 0.15 \approx 2.74, -\lg p(x_3) + 1 = 3.74, \therefore k_3 = 3$$

$$\because -\lg p(x_4) = -\lg 0.05 \approx 4.32, -\lg p(x_4) + 1 = 5.32, \therefore k_4 = 5$$

④将 $p_a(x_i)$ 用二进制表示，取小数点后 k_i 位作为 x_i 的码字

$$p_a(x_1) = 0.0 = (0.00)_2 \rightarrow 0$$

$$p_a(x_2) = 0.5 = (0.100)_2 \rightarrow 10$$

$$p_a(x_3) = 0.8 = (0.1100\dots)_2 \rightarrow 110$$

$$p_a(x_4) = 0.95 = (0.111100\dots)_2 \rightarrow 11110$$

单符号信源的香农编码-举例

x_i	$p(x_i)$	$p_a(x_i)$	k_i	码字
x_1	0.5	0	1	0
x_2	0.3	0.5	2	10
x_3	0.15	0.8	3	110
x_4	0.05	0.95	5	11110

$$\begin{aligned}\because H(X) &= -\sum_{i=1}^4 p(x_i) \lg p(x_i) \\ &= -0.5 \lg 0.5 - 0.3 \lg 0.3 - 0.15 \lg 0.15 - 0.05 \lg 0.05 \\ &\approx 1.648(\text{bit} / \text{symbol})\end{aligned}$$

单符号信源的香农编码-举例

$$\begin{aligned}\bar{K} &= \sum_{i=1}^4 p(x_i)k_i \\ &= 0.5 \times 1 + 0.3 \times 2 + 0.15 \times 3 + 0.05 \times 5 = 1.8(\text{bit} / \text{symbol})\end{aligned}$$

$$\therefore \eta = \frac{H(X)}{\bar{K}} = \frac{1.648}{1.8} \approx 0.916 = 91.6\%$$

N维扩展信源的香农编码-举例

例2，分别对下列单符号离散信源和该信源的二次扩展信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1 & x_2 & x_3 \\ 0.5 & 0.3 & 0.2 \end{Bmatrix}$$

编二进制香农码，并计算其编码效率。

解：(1)对单符号离散信源编码

$$p_a(x_1)=0$$

$$p_a(x_2)=0+0.5=0.5$$

$$p_a(x_3)=0.5+0.3=0.8$$

N 维扩展信源的香农编码-举例

$$\because -\text{lb}p(x_1) = \text{lb}2 = 1 \quad \therefore k_1 = 1$$

$$\because -\text{lb}p(x_2) = -\text{lb}0.3 \approx 1.74, -\text{lb}p(x_2) + 1 = 2.74, \quad \therefore k_2 = 2$$

$$\because -\text{lb}p(x_3) = -\text{lb}0.2 \approx 2.32, -\text{lb}p(x_3) + 1 = 3.32, \quad \therefore k_3 = 3$$

$$\mathbf{p_a(x_1)=0.0=(0.0)_2 \rightarrow 0}$$

$$\mathbf{p_a(x_2)=0.5=(0.10)_2 \rightarrow 10}$$

$$\mathbf{p_a(x_3)=0.8=(0.110\dots)_2 \rightarrow 110}$$

N维扩展信源的香农编码-举例

$\mathbf{x_i}$	$\mathbf{p(x_i)}$	$\mathbf{p_a(x_i)}$	$\mathbf{k_i}$	码字
$\mathbf{x_1}$	0.5	0	1	0
$\mathbf{x_2}$	0.3	0.5	2	10
$\mathbf{x_3}$	0.2	0.8	3	110

$$\therefore H(X) = -\sum_{i=1}^3 p(x_i) \lg p(x_i) = -0.5 \lg 0.5 - 0.3 \lg 0.3 - 0.2 \lg 0.2 \approx 1.485$$

$$\bar{K} = \sum_{i=1}^3 p(x_i) k_i = 0.5 \times 1 + 0.3 \times 2 + 0.2 \times 3 = 1.7$$

$$\therefore \eta = \frac{H(X)}{\bar{K}} = \frac{1.485}{1.7} \approx 0.874 = 87.4\%$$

N维扩展信源的香农编码-举例

(2)对二次扩展信源编码

$$\begin{bmatrix} X^2 \\ P(X^2) \end{bmatrix} = \begin{Bmatrix} x_1x_1 & x_1x_2 & x_1x_3 & x_2x_1 & x_2x_2 & x_2x_3 & x_3x_1 & x_3x_2 & x_3x_3 \\ 0.25 & 0.15 & 0.1 & 0.15 & 0.09 & 0.06 & 0.1 & 0.06 & 0.04 \end{Bmatrix}$$

将符号元 a_i 按概率进行降序排列

$$\begin{bmatrix} X^2 \\ P(X^2) \end{bmatrix} = \begin{Bmatrix} x_1x_1 & x_1x_2 & x_2x_1 & x_1x_3 & x_3x_1 & x_2x_2 & x_2x_3 & x_3x_2 & x_3x_3 \\ 0.25 & 0.15 & 0.15 & 0.1 & 0.1 & 0.09 & 0.06 & 0.06 & 0.04 \end{Bmatrix}$$

$$p_a(a_1)=0$$

$$p_a(a_2)=0+0.25=0.25$$

$$p_a(a_3)=0.25+0.15=0.4$$

$$p_a(a_4)=0.4+0.15=0.55$$

N 维扩展信源的香农编码-举例

$$p_a(a_5)=0.55+0.1=0.65$$

$$p_a(a_6)=0.65+0.1=0.75$$

$$p_a(a_7)=0.75+0.09=0.84$$

$$p_a(a_8)=0.84+0.06=0.9$$

$$p_a(a_9)=0.9+0.06=0.96$$

$$\because -\text{lb}p(a_1) = -\text{lb}0.25 = 2 \quad \therefore k_1 = 2$$

$$\because -\text{lb}p(a_2) = -\text{lb}0.15 \approx 2.74, -\text{lb}p(a_2) + 1 = 3.74$$

$$\text{取 } k_2 = 3$$

N维扩展信源的香农编码-举例

同理 $\because -\text{lb}p(a_3) = -\text{lb}0.15 \approx 2.74, -\text{lb}p(a_3) + 1 = 3.74$

取 $k_3 = 3$

$\because -\text{lb}p(a_4) = -\text{lb}0.1 \approx 3.32, -\text{lb}p(a_4) + 1 = 4.32$

取 $k_4 = 4$

$\because -\text{lb}p(a_5) = -\text{lb}0.1 \approx 3.32, -\text{lb}p(a_5) + 1 = 4.32$

取 $k_5 = 4$

$\because -\text{lb}p(a_6) = -\text{lb}0.09 \approx 3.47, -\text{lb}p(a_6) + 1 = 4.47$

取 $k_6 = 4$

N维扩展信源的香农编码-举例

$$\because -\text{lb}p(a_7) = -\text{lb}0.06 \approx 4.06, -\text{lb}p(a_7) + 1 = 5.06$$

取 $k_7 = 5$

$$\because -\text{lb}p(a_8) = -\text{lb}0.06 \approx 4.06, -\text{lb}p(a_8) + 1 = 5.06$$

取 $k_8 = 5$

$$\because -\text{lb}p(a_9) = -\text{lb}0.04 \approx 4.64, -\text{lb}p(a_9) + 1 = 5.64$$

取 $k_9 = 5$

$$p_a(a_1) = 0.0 = (0.00)_2 \rightarrow 00$$

$$p_a(a_2) = 0.25 = (0.010)_2 \rightarrow 010$$

$$p_a(a_3) = 0.4 = (0.011\dots)_2 \rightarrow 011$$

N 维扩展信源的香农编码-举例

$$p_a(a_4)=0.55=(0.1000\dots)_2 \rightarrow 1000$$

$$p_a(a_5)=0.65=(0.1010\dots)_2 \rightarrow 1010$$

$$p_a(a_6)=0.75=(0.1100)_2 \rightarrow 1100$$

$$p_a(a_7)=0.84=(0.11010\dots)_2 \rightarrow 11010$$

$$p_a(a_8)=0.9=(0.11100\dots)_2 \rightarrow 11100$$

$$p_a(a_9)=0.96=(0.11101\dots)_2 \rightarrow 11101$$

N 维扩展信源的香农编码, 举例

a_i	$p(a_i)$	$p_a(a_i)$	k_i	码字
a_1	0.25	0	2	00
a_2	0.15	0.25	3	010
a_3	0.15	0.4	3	011
a_4	0.1	0.55	4	1000
a_5	0.1	0.65	4	1010
a_6	0.09	0.75	4	1100
a_7	0.06	0.84	5	11010
a_8	0.06	0.9	5	11100
a_9	0.04	0.96	5	11101

N 维扩展信源的香农编码-举例

$$\because H(X^2) = 2H(X) = 2 \times 1.485 = 2.97$$

$$\begin{aligned}\bar{K} &= \sum_{i=1}^9 p(a_i)k_i = 0.25 \times 2 + 2 \times 0.15 \times 3 + 2 \times 0.1 \times 4 \\ &+ 0.09 \times 4 + 2 \times 0.06 \times 5 + 0.04 \times 5 = 3.36\end{aligned}$$

$$\therefore \eta = \frac{H(X^2)}{\bar{K}} = \frac{2.97}{3.36} \approx 0.884 = 88.4\%$$

N 维扩展信源的香农编码-举例

$$\text{或} \because H(X) = 1.485$$

$$\frac{\bar{K}}{N} = \frac{3.36}{2} = 1.68$$

$$\therefore \eta = \frac{H(X)}{\frac{\bar{K}}{N}} = \frac{1.485}{1.68} \approx 0.884 = 88.4\%$$

$N=1$: 单符号的编码效率为87.4%

$N=2$: 编码效率为88.4%，即码率更接近于熵率，冗余减少。

香农编码的特点

- 先确定码字，后确定码长
- 码字确定原则：
 - 平均码长不超过联合熵一个比特， N 越大平均码长越接近联合熵
 - 保证大概率符号序列编为短码，小概率符号序列编为长码
- 码字确定原则：根据累加概率确定码字，具有唯一性，且第一个符号序列的对应的码字总是0、00、...形式

授课内容



无失真编码的基本思路

无失真编码定理

香农编码

霍夫曼编码

费诺编码

赫夫曼(Huffman)编码

霍夫曼编码也是一种采用异前置码的**m**进制编码方法。

霍夫曼编码的编码效率高于香农编码。

设离散信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1 & x_2 & \cdots & x_n \\ p(x_1) & p(x_2) & \cdots & p(x_n) \end{Bmatrix}$$

设 $p(x_1) > p(x_2) > \cdots > p(x_n)$, 且 $\sum_{i=1}^n p(x_i) = 1$

赫夫曼编码-编码步骤

- ①将信源符号按概率由大到小顺序排队；
 - ②给两个概率最小的符号各分配一个码位:为概率最小的符号元分配一个码元1，概率次小的符号元分配一个码元0；
 - ③将概率最小的两个符号元合并成一个新的符号元，用两者概率之和作为该新符号元的概率；
- 重复以上三个步骤，直到最后合并出一个以1为概率的符号元，结束编码。

单符号离散信源的赫夫曼编码-举例

例1，对单符号离散信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1 & x_2 & x_3 & x_4 \\ 0.5 & 0.3 & 0.15 & 0.05 \end{Bmatrix}$$

编二进制霍夫曼码，并
计算其编码效率。

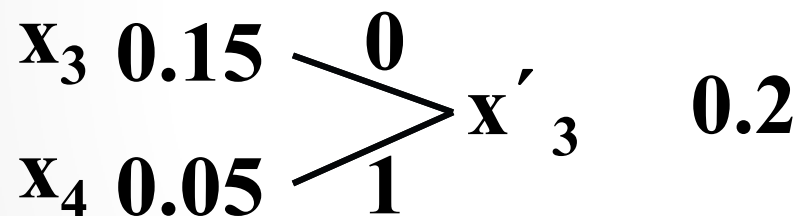
解：

①将符号元按概率进行降
序排列

符号元	概率
x_1	0.5
x_2	0.3
x_3	0.15
x_4	0.05

单符号离散信源的赫夫曼编码-举例

②为概率最小的符号元分配一个码元1，概率次小的符号元分配一个码元0；

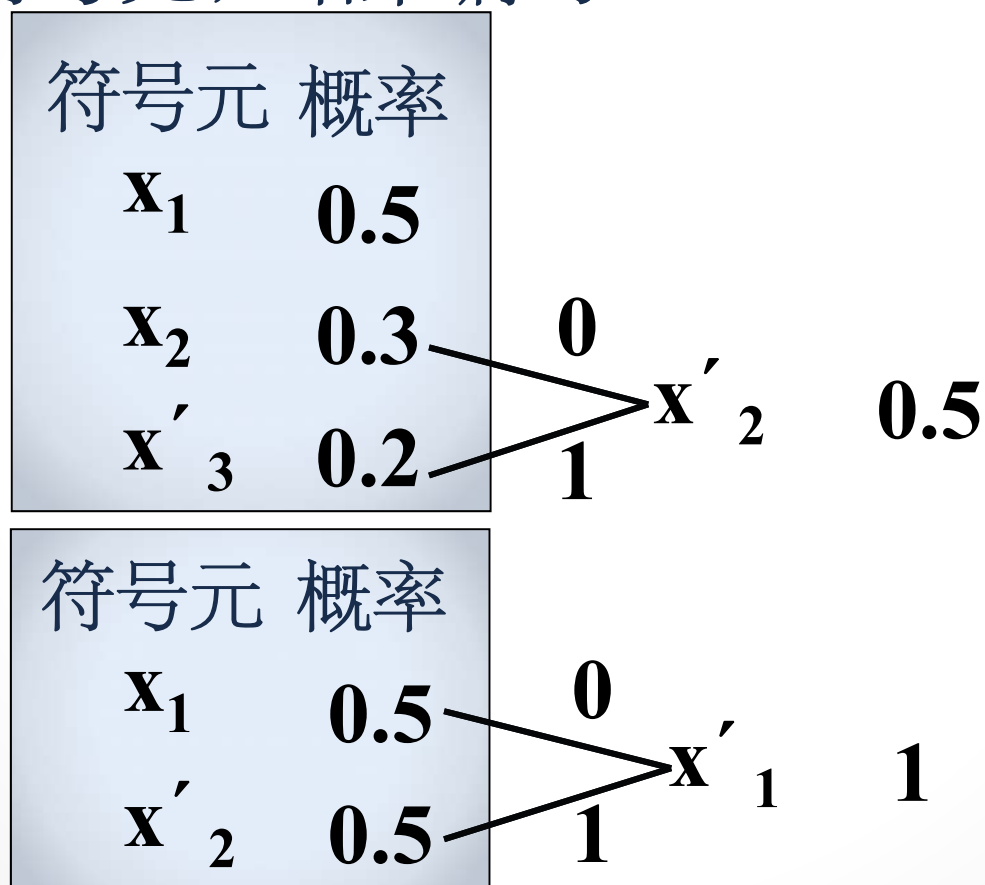


符号元	概率
$\mathbf{x_1}$	0.5
$\mathbf{x_2}$	0.3
$\mathbf{x_3}$	0.15
$\mathbf{x_4}$	0.05

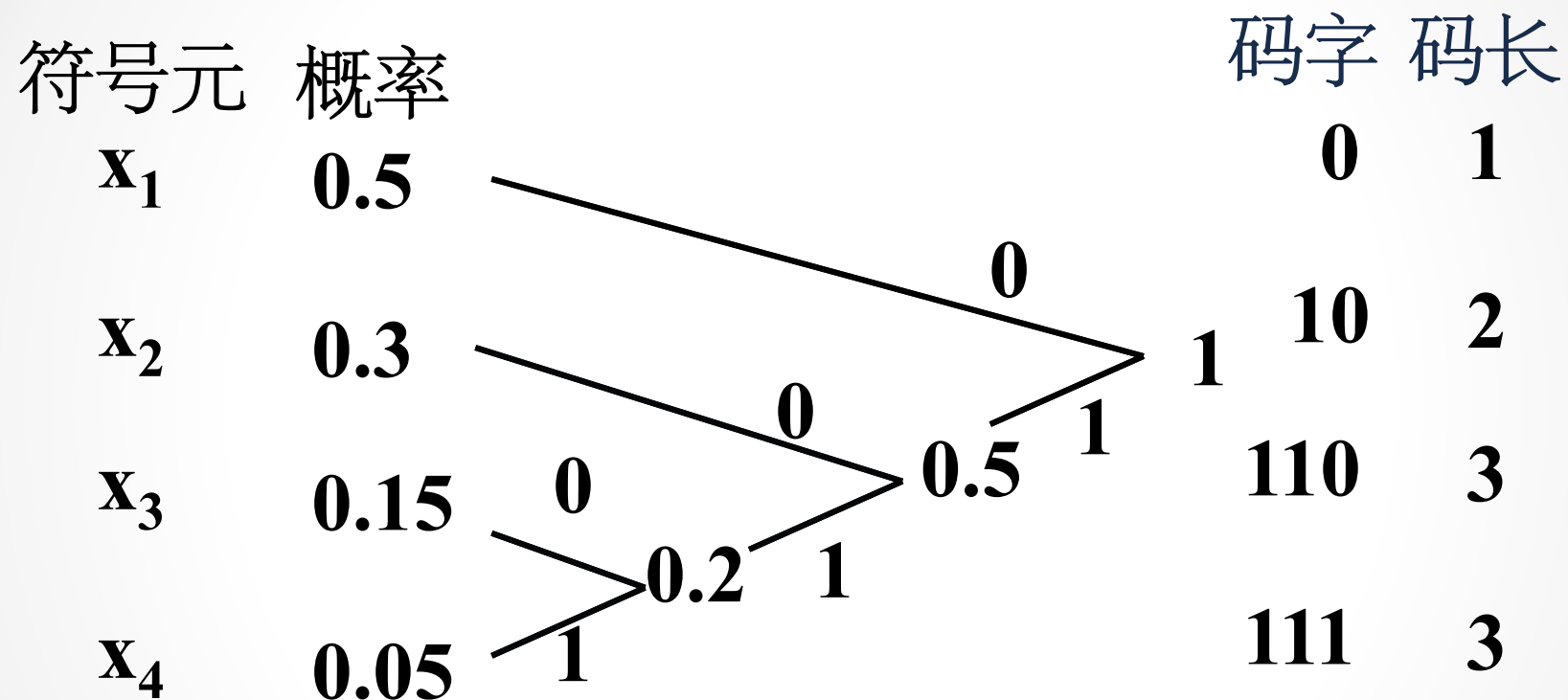
③将概率最小的两个符号元合并成一个新的符号元，用两者概率之和作为该新符号元的概率；

单符号离散信源的赫夫曼编码-举例

重复以上三个步骤，直到最后合并出一个以1为概率的符号元，结束编码。



单符号离散信源的赫夫曼编码-举例



显然，所编霍夫曼码构成二元码树，为异前置码。

单符号离散信源的赫夫曼编码-举例

$$\begin{aligned}\because H(X) &= -\sum_{i=1}^4 p(x_i) \lg p(x_i) \\ &= -0.5 \lg 0.5 - 0.3 \lg 0.3 - 0.15 \lg 0.15 - 0.05 \lg 0.05 \\ &\approx 1.648(\text{bit} / \text{symbol})\end{aligned}$$

$$\begin{aligned}\bar{K} &= \sum_{i=1}^4 p(x_i) k_i \\ &= 0.5 \times 1 + 0.3 \times 2 + 0.15 \times 3 + 0.05 \times 3 = 1.7(\text{bit} / \text{symbol})\end{aligned}$$

$$\therefore \eta = \frac{H(X)}{\bar{K}} = \frac{1.648}{1.7} \approx 0.969 = 96.9\%$$

N维扩散信源的赫夫曼编码-举例

例2，分别对下列单符号离散信源和该信源的二次扩展信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1 & x_2 & x_3 \\ 0.5 & 0.4 & 0.1 \end{Bmatrix}$$

编二进制霍夫曼码，并计算其编码效率。

解：(1)对单符号离散信源编码

x_i	$P(x_i)$		码字	码长
x_1	0.5		0	1
x_2	0.4		10	2
x_3	0.1		11	2

N维扩散信源的赫夫曼编码-举例

$$\because H(X) = -\sum_{i=1}^3 p(x_i) \lg p(x_i)$$

$$= -0.5 \lg 0.5 - 0.4 \lg 0.4 - 0.1 \lg 0.1$$

$$\approx 1.36(\text{bit} / \text{symbol})$$

$$\bar{K} = \sum_{i=1}^3 p(x_i) k_i$$

$$= 0.5 \times 1 + 0.4 \times 2 + 0.1 \times 2 = 1.5(\text{bit} / \text{symbol})$$

$$\therefore \eta = \frac{H(X)}{\bar{K}} = \frac{1.36}{1.5} \approx 90.7\%$$

N维扩散信源的赫夫曼编码-举例

(2)对二次扩展信源编码

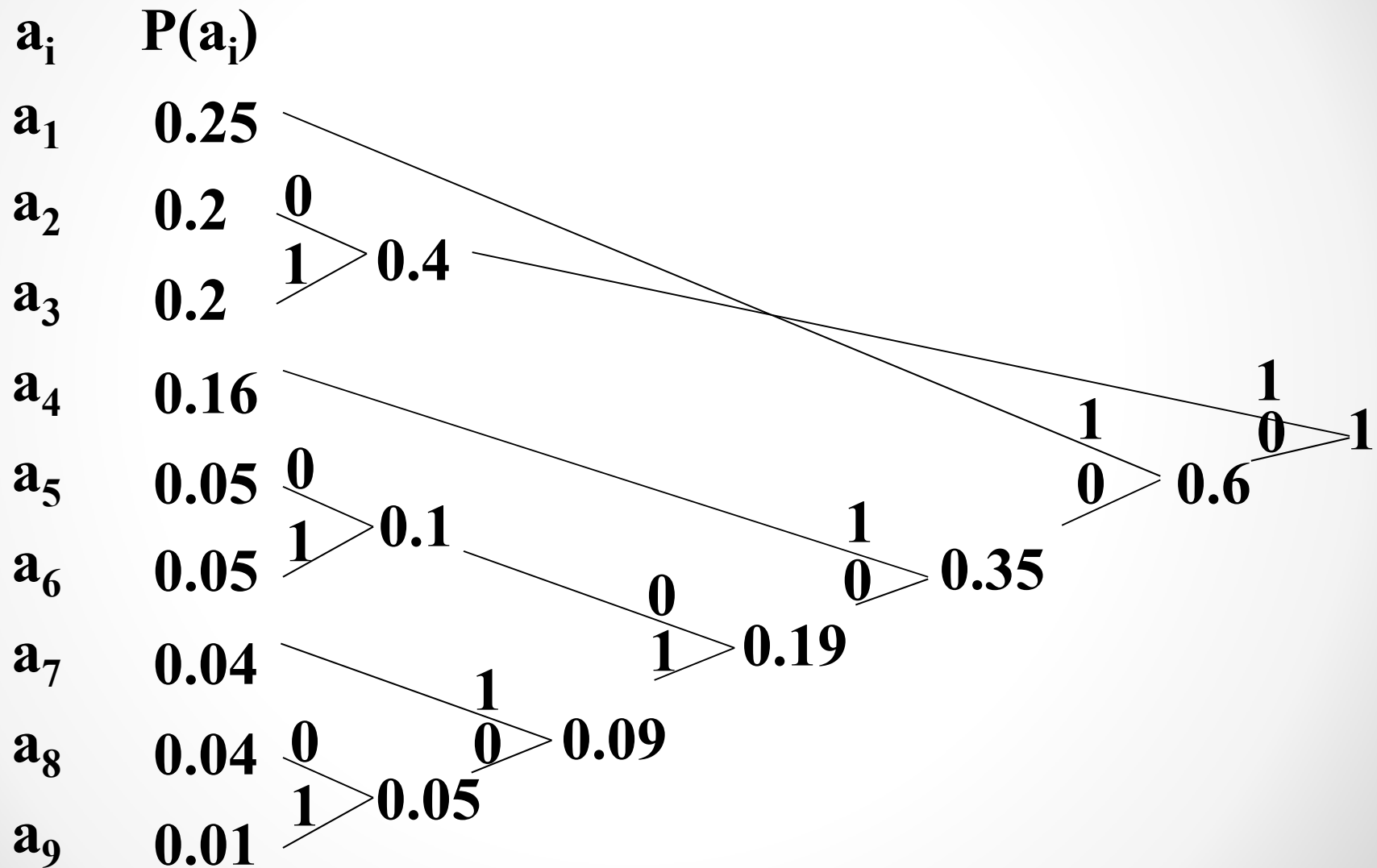
$$\begin{bmatrix} X^2 \\ P(X^2) \end{bmatrix} = \left\{ \begin{array}{ccccccccc} x_1x_1 & x_1x_2 & x_1x_3 & x_2x_1 & x_2x_2 & x_2x_3 & x_3x_1 & x_3x_2 & x_3x_3 \\ 0.25 & 0.2 & 0.05 & 0.2 & 0.16 & 0.04 & 0.05 & 0.04 & 0.01 \end{array} \right\}$$

将符号元 $\mathbf{ai}$ 按概率进行降序排列

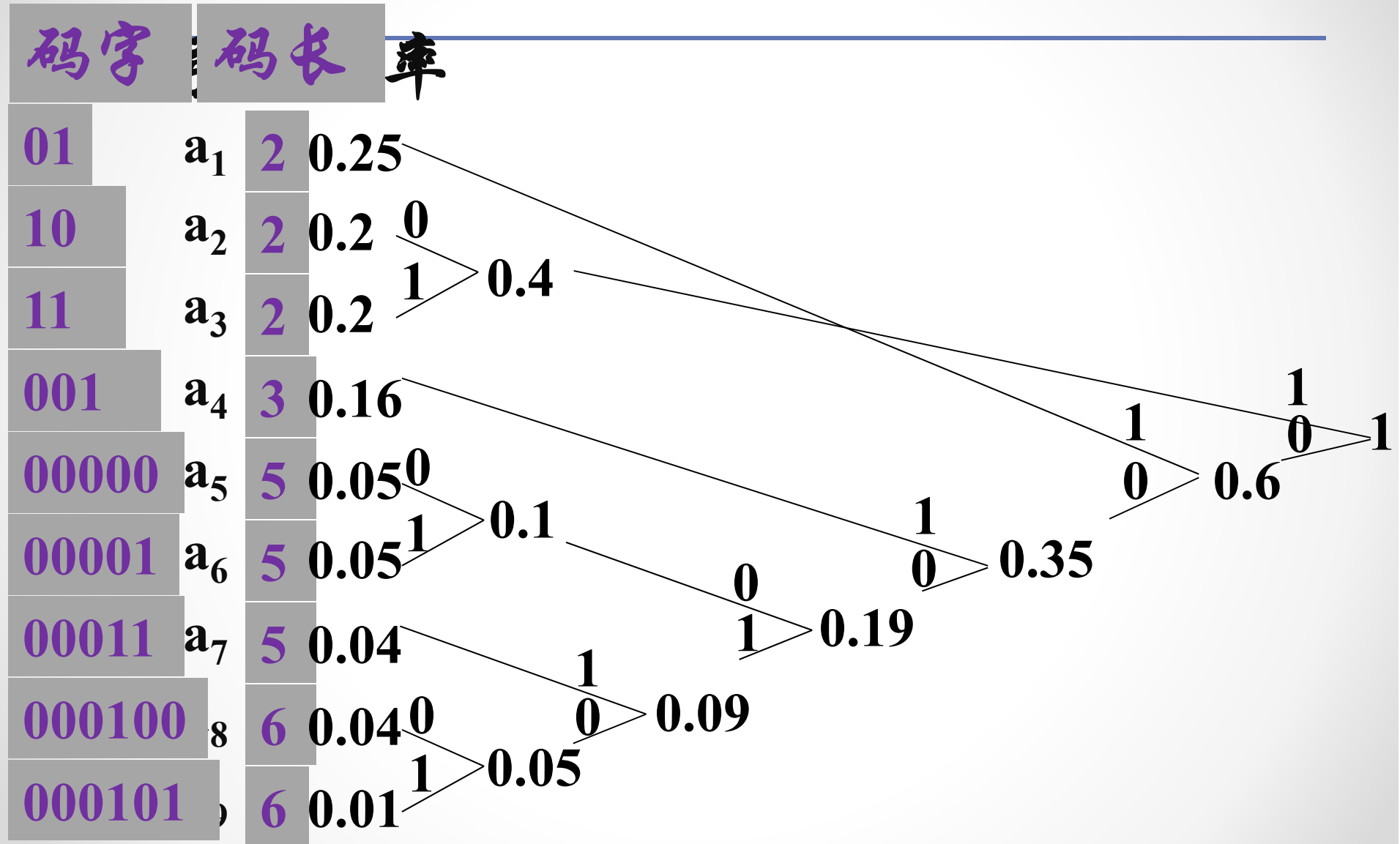
$$\begin{bmatrix} X^2 \\ P(X^2) \end{bmatrix} = \left\{ \begin{array}{ccccccccc} x_1x_1 & x_1x_2 & x_2x_1 & x_2x_2 & x_1x_3 & x_3x_1 & x_2x_3 & x_3x_2 & x_3x_3 \\ 0.25 & 0.2 & 0.2 & 0.16 & 0.05 & 0.05 & 0.04 & 0.04 & 0.01 \end{array} \right\}$$

霍夫曼编码过程为

N维扩散信源的赫夫曼编码-举例



N维扩散信源的赫夫曼编码-举例



N 维扩散信源的赫夫曼编码-举例

$$\begin{aligned} K &= \sum_{i=1}^9 P(a_i)k_i \\ &= (0.25 + 2 \times 0.2) \times 2 + 0.16 \times 3 + (2 \times 0.05 + 0.04) \times 5 + (0.04 + 0.01) \times 6 \\ &= 2.78(\text{bit} / \text{message}) \end{aligned}$$

$$\therefore H(X) \approx 1.36 \text{bit/symbol}$$

$$\eta = \frac{H(X^2)}{K} = \frac{2 \times 1.36}{2.78} = \frac{2.72}{2.78} \approx 97.91\%$$

赫夫曼编码非唯一性-举例

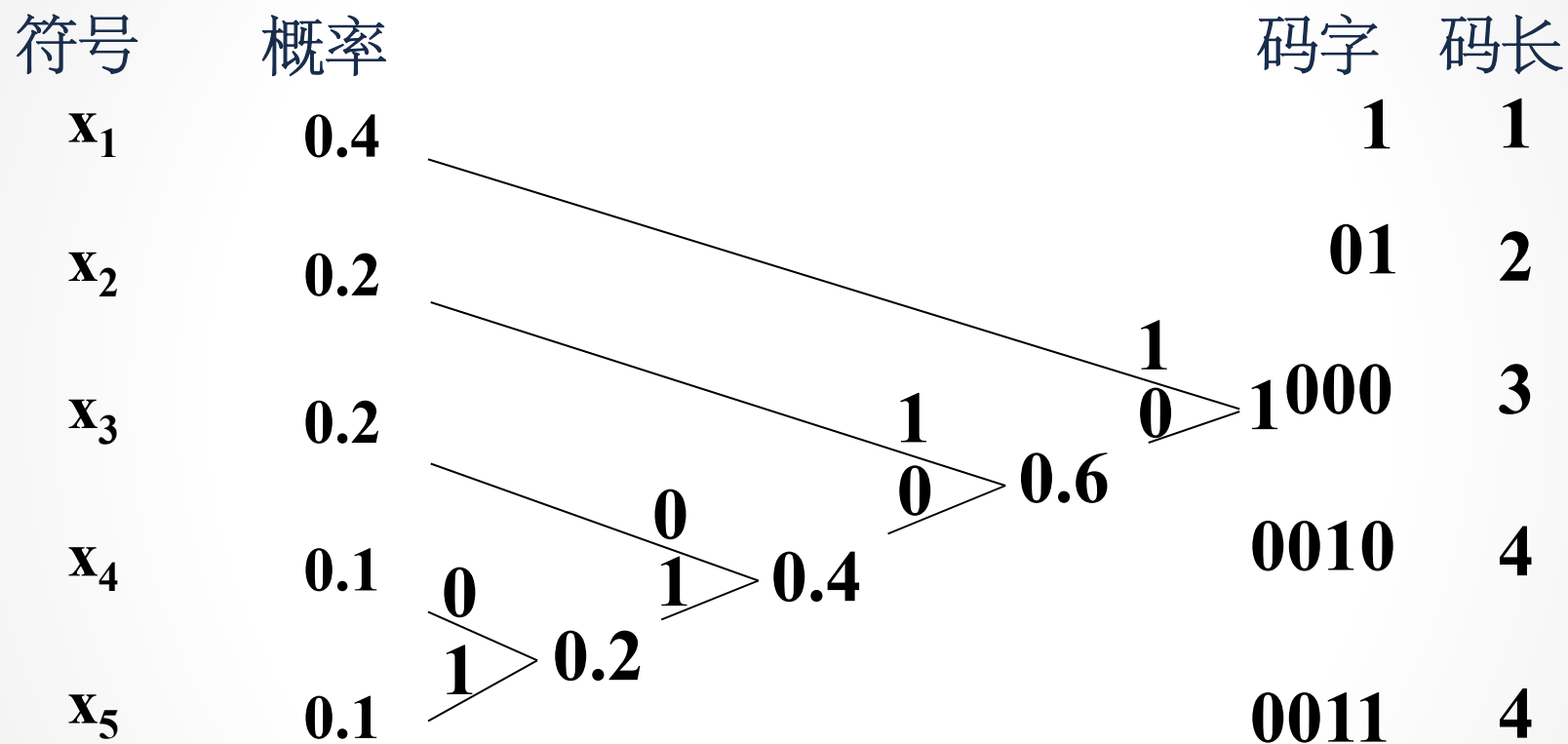
例3，对单符号离散信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 \\ 0.4 & 0.2 & 0.2 & 0.1 & 0.1 \end{Bmatrix}$$

用两种方法对其进行二进制霍夫曼编码，并计算其平均码长。

解：(1)将合并后的新符号排在其他相同概率符号后
编霍夫曼码的过程如下：

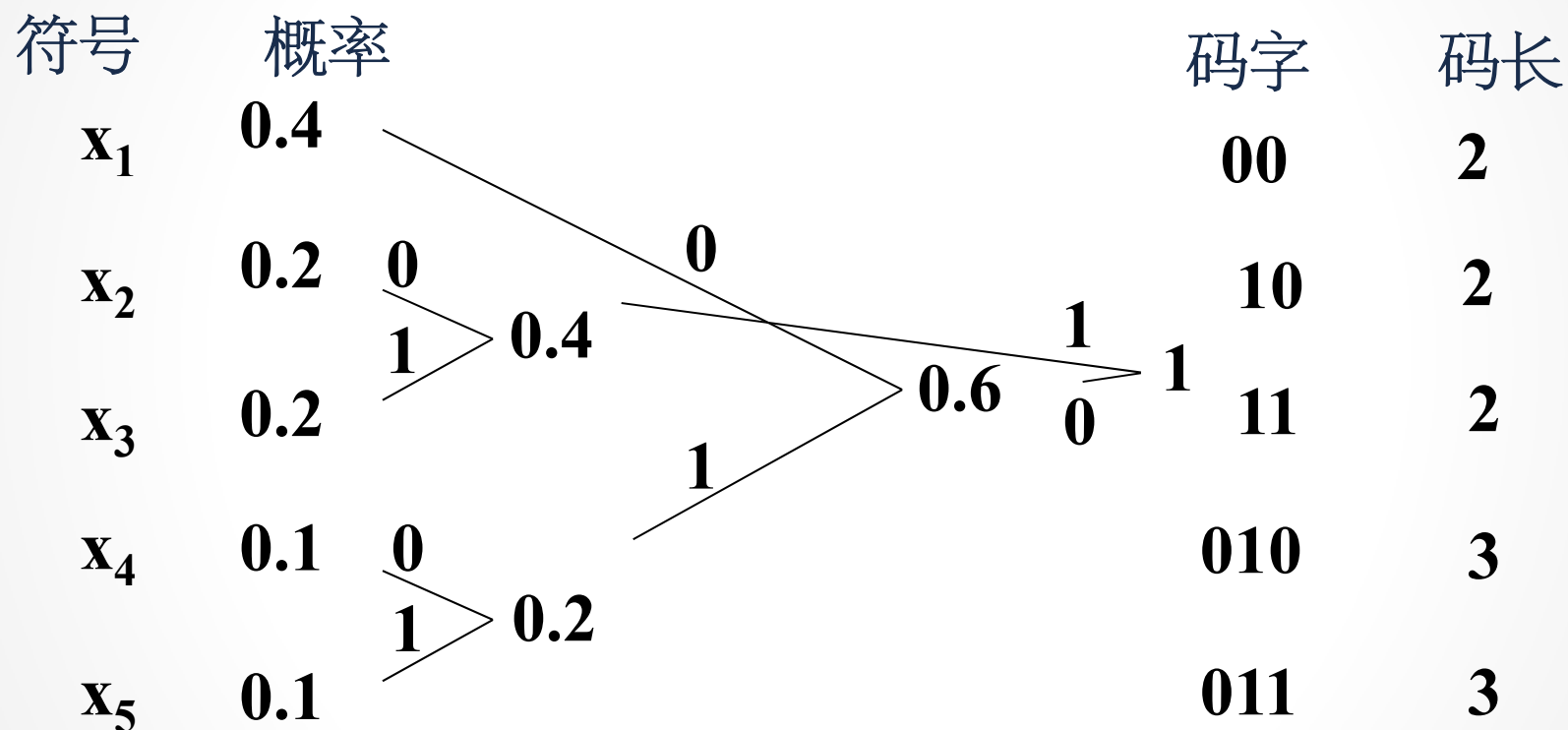
赫夫曼编码非唯一性-举例



$$\bar{K} = \sum_{i=1}^5 p(a_i)k_i = 0.4 \times 1 + 0.2 \times 2 + 0.2 \times 3 + 2 \times 0.1 \times 4 = 2.2(\text{bit/symbol})$$

赫夫曼编码非唯一性-举例

(2) 将合并后的新符号排在其他相同概率符号之前



$$\bar{K} = \sum_{i=1}^5 p(a_i)k_i = 0.4 \times 2 + 0.2 \times 2 \times 2 + 0.1 \times 3 \times 2 = 2.2(\text{bit/symbol})$$

赫夫曼编码非唯一性-举例

两种不同的编码方法得到的码字和码长的对比

信源符号 x_i	概率 $p(x_i)$	码字 W_{i1}	码长 K_{i1}	码字 W_{i2}	码长 K'_{i2}
x_1	0.4	1	1	00	2
x_2	0.2	01	2	10	2
x_3	0.2	000	3	11	2
x_4	0.1	0010	4	010	3
x_5	0.1	0011	4	011	3

赫夫曼编码非唯一性-举例

平均码长和编码效率：

$$\bar{K} = \sum_{i=1}^5 p(a_i)k_i = 2.2(\text{bit/symbol})$$

$$\eta = \frac{NH(X)}{\bar{K}} = 0.965 = 96.5\%$$

码长方差：

$$\sigma_l^2 = E\left[(k_i - \bar{K})^2\right] = \sum_{i=1}^Q p(a_i)(k_i - \bar{K})^2$$

$$\sigma_{l1}^2 = 1.36 \quad \sigma_{l2}^2 = 0.16$$

赫夫曼编码非唯一性-说明

- **Huffman**码的编码方法不是唯一的。
- 采用不同排列方法编出的赫夫曼码，其码字和码长可能完全不相同，但平均码长一定相等，编码效率不会因排列方法而改变。
- 不同编码方式影响码字的长度方差，一般将合并的概率放在上面，这样可获得较小的码方差。

赫夫曼编码非唯一性-结论

每次分配码元规律一致，使得码字可分离。

码字始于树根，止于终节点，符合异前置码原则。

进行赫夫曼编码时，为得到码方差最小的码，应使合并的信源符号位于缩减信源序列尽可能高的位置上，以减少再次合并的次数，充分利用短码。

霍夫曼编码特点

- 先确定码字，后确定码长
- 码字确定原则：
 - 用局部累加概率代替累加概率，根据合并概率逐级分配码元，保证大概率符号序列编为短码，小概率符号序列为长码
 - 不具有唯一性，但不同赫夫曼码本的编码效率相同
- 平均码长不超过联合熵一个比特， N 越大平均码长越接近联合熵

授课内容



无失真编码的基本思路



无失真编码定理



香农编码



霍夫曼编码



费诺编码

费诺码的编码步骤

费诺码基于概率匹配的思想，也是采用异前置码实现的无失真不等长编码。

费诺码的编码步骤：

- ①将符号序列 a_i 按概率降序排列；
 - ②进行分组，使每组概率尽可能相等；
 - ③给每个分组分配一个二进制码元0和1；
- 重复② ③步骤，直到不可分为止。

费诺编码-举例

例，对单符号离散信源进行费诺编码，并计算编码效率。

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{Bmatrix} x_1 & x_2 & x_3 & x_4 \\ 0.5 & 0.3 & 0.15 & 0.05 \end{Bmatrix}$$

x_i	$P(x_i)$	分组1	分组2	分组3	码字
x_1	0.5				
x_2	0.3				
x_3	0.15				
x_4	0.05				

费诺编码-举例

第一次分组:

$$P(x_1) = 0.5$$

$$P(x_2) + P(x_3) + P(x_4) = 0.3 + 0.15 + 0.05 = 0.5$$

x_i	$P(x_i)$	分组1	分组2	分组3	码字
x_1	0.5	0			
x_2	0.3	1			
x_3	0.15				
x_4	0.05				

费诺编码-举例

第二次分组:

$$P(x_2) = 0.3$$

$$P(x_3) + P(x_4) = 0.15 + 0.05 = 0.2$$

x_i	$P(x_i)$	分组1	分组2	分组3	码字
x_1	0.5	0			
x_2	0.3	1	0		
x_3	0.15		1		
x_4	0.05				

费诺编码-举例

第三次分组:

$$P(x_3) = 0.15$$

$$P(x_4) = 0.05$$

x_i	$P(x_i)$	分组1	分组2	分组3	码字
x_1	0.5	0			0
x_2	0.3	1	0		10
x_3	0.15		1	0	110
x_4	0.05			1	111

费诺编码-举例

$$\begin{aligned} K &= \sum_{i=1}^4 P(x_i)k_i \\ &= 0.5 \times 1 + 0.3 \times 2 + 0.15 \times 3 + 0.05 \times 3 \\ &= 1.7(\text{bit} / \text{symbol}) \end{aligned}$$

$$\eta = \frac{H(X)}{K} = \frac{1.648}{1.7} \approx 96.94\%$$

费诺码的特点

- 先确定码字，后确定码长
- 大概率符号序列分解次数少，编为短码，小概率符号序列分解次数多，编为长码
- 具有唯一性
- 平均码长不超过联合熵一个比特， N 越大平均码长越接近联合熵

作业:

• Pp166 5.1 , 5.3

• 5.5(5) 设无记忆二进制信源 $\begin{pmatrix} X \\ P(X) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0.1 & 0.9 \end{pmatrix}$
把信源序列变成数字0,1,2,3, 在替换成二进制变长码字

序列	数字	二元码字
1	0	100
01	1	101
001	2	110
000	3	0