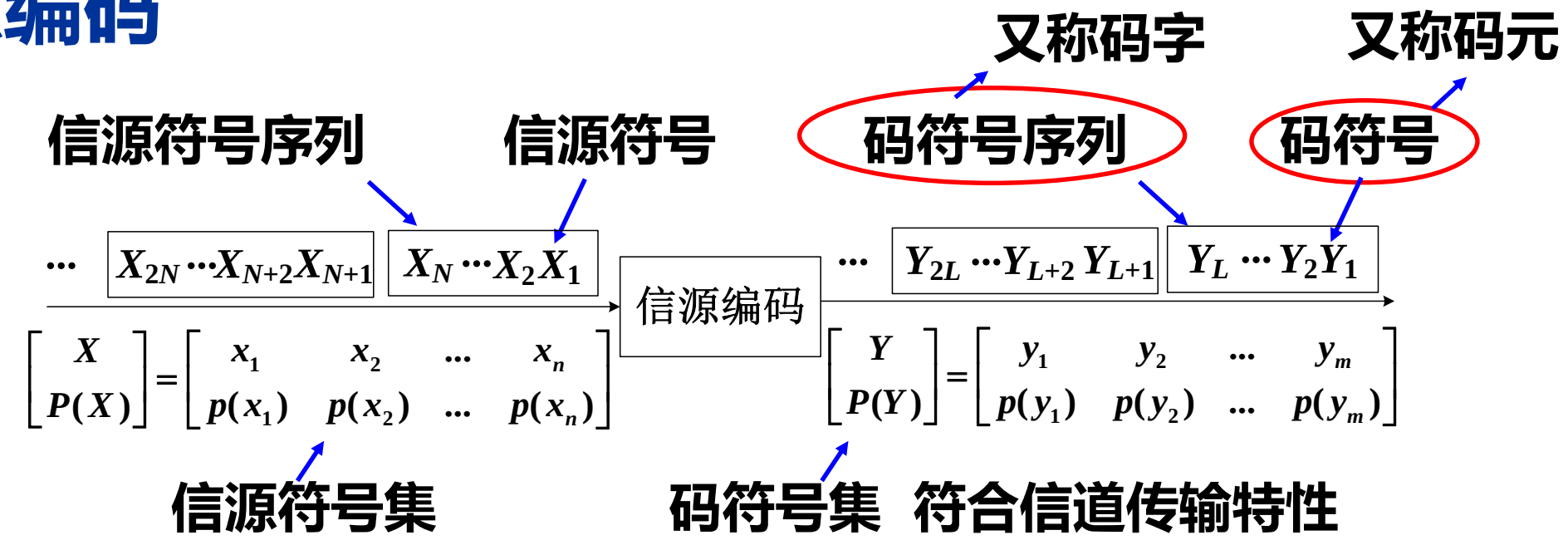
基础信息论

定长编码定理

华中科技大学电信学院

信源编码



信源符号序列集合 $S=\{s_1,s_2,\cdots,s_q\}\quad q\le n^N$

码符号序列集合 又称码书，码 $W=\{\omega_1,\omega_2,\cdots,\omega_q\}$

$s_i=\{x_{i_1}x_{i_2}\cdots x_{i_N}\}$

$i_1,i_2,\cdots i_N\in\{1,2,\cdots n\}$

$\omega_i=\{y_{i_1}y_{i_2}\cdots y_{i_{l_i}}\}$

$i_1,i_2,\cdots\in\{1,2,\cdots m\}$　　l_i: 码字长度

学习目标

- 定义唯一可译码条件
- 说明唯一可译定长码的存在条件
- 定义编码效率
- 解释定长编码定理的物理含义
- 应用定长编码定理，给定误码率要求，计算定长码的最小长度

唯一可译定长码的存在条件

唯一可译要求: 码的任意有限次扩展码应为非奇异码。

定长码: 每个码字长度相等，所以只要定长码是非奇异码，则必为唯一可译码。

- 对一个信源X进行定长编码，信源X存在唯一可译定长码的条件是:

$$n \leq m^K$$

  - n是信源X的符号个数，m是码符号数（码元个数），K是定长码的码长（码字长）。

L次扩展信源的定长码

- 对L次扩展信源X^L进行定长编码，若要编得的定长码是唯一可译码，则必须满足：

$$n^L \leq m^K$$

两边取以2为底的对数，有

$$L\log n \leq K\log m$$

或

$$\frac{K}{L} \geq \frac{\log n}{\log m} = \log_m n$$

唯一可译定长码的存在条件 - 举例

- 英文电报信源有32个符号，26个英文字母加上6个标点符号，对此信源的每个符号进行二元编码。如何实现唯一可译定长码?

分析:

信源符号数为n = 32，码元个数为m = 2，码字长度?

$$K \geq \log_m n = \log n = \log 32 = 5$$

这就是说，每个英文电报符号至少要用5位二元符号进行编码才能得到唯一可译码。

定长信源编码定理 - 引入

$$\frac{K}{L} \geq \frac{\log n}{\log m} = \log_m n$$

- 满足上述条件的定长编码，可保证无失真的编码

问题：平均码长很大，编码的效率很低。

- **定长编码定理：讨论了编码的有关参数对译码差错的限制关系。**

信源编码（主要内容）

- 信源编码定理（定长、变长编码定理）
  - 信源编码的相关概念：
  - 定长编码定理
  - 变长编码定理（香农第一定理）
  - 香农第三定理
- 信源编码方法
  - 离散信源编码
  - 连续信源编码
  - 相关信源编码
  - 变换编码

唯一可译定长码的存在条件

定长信源编码定理

定长编码定理

正定理：

一个熵为$H(X)$的离散无记忆信源，若对长度为L的信源符号序列进行等长编码，设码字是从m个码符号集中选取的K个码元组成。对于任意的$\varepsilon > 0$和$1 > \delta > 0$，

只要满足：

$$\frac{K \cdot \log m}{L} \geq H(X) + \varepsilon$$

则当L足够长，必可使译码差错小于δ。

δ：所要求的译码差错概率

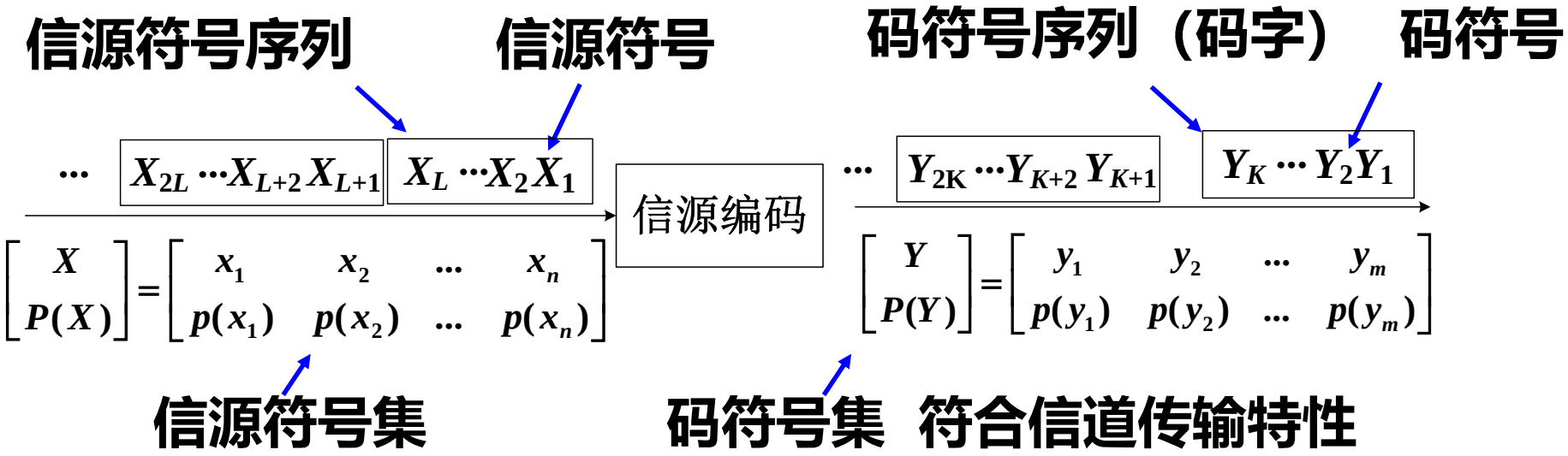
逆定理：

反之，当： $\frac{K \cdot \log m}{L} \leq H(X) - 2\varepsilon$ **，译码差错一定大于δ。**

当$L \to \infty$时，译码差错趋近于1。

定长编码定理的物理意义

编码场景



n：信源符号个数
L：信源符号序列长

m：码符号个数
K：码符号序列长

定长编码定理的物理意义

正定理： $\dfrac{K\cdot\log m}{L}\geq H(X)+\varepsilon$

若 ⇒ <，说明K不够。

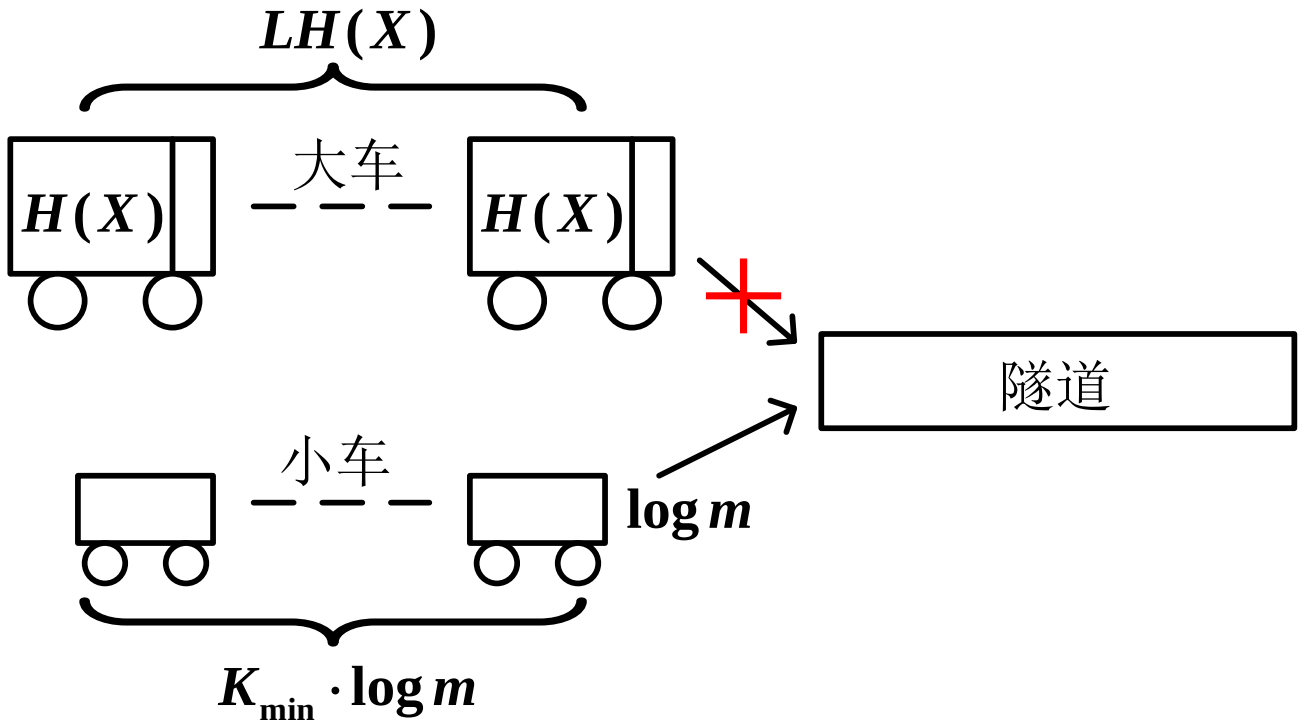
$K\cdot\log m$： K长码符号序列的最大可能载荷信息量。

$\dfrac{K\cdot\log m}{L}$： 折算后，平均每个信源符号的最大可能载信量。

$H(X)$： 每个信源符号的实际载信量。

意义：定长编码后，所能携带的最大信息量，一定要大于信源所携带的平均信息量(熵)

定长编码定理的物理意义



$$\frac{K_{\min} \cdot \log m}{L} = H(X) + \varepsilon$$

定长编码定理 - 提高效率

唯一可译定长码的存在条件 $\frac{K}{L} \geq \frac{\log n}{\log m} = \log_m n$

问题：平均码长很大，编码的效率很低。

- 分析：

  一般情况下，信源符号非等概率分布，且相互关联

  ➡ 信源极限熵

$$H_\infty(\mathrm{X}) \ll H_{\max}(\mathrm{X}) = \log \mathrm{n}$$

进行定长编码可使每个信源符号平均所需的码符号长大大减少，从而提高效率

定长编码定理 - 提高效率

例：英文电报的二元编码：

已知信源极限熵$H_\infty(X)$ ≈1.4比特/符号

由 $$\frac{K \cdot \log m}{L} \geq H(X) + \varepsilon$$

可以得出：即(K/L)>1.4；

平均每个英文信源符号只需近似用1.4个二元符号来编码，这比由式

$$\frac{K}{L} \geq \log(n)$$

计算的需要5位二元符号减少了许多，从而提高了信息传输速率。

编码信息率

- 编码信息率：编码后平均每个信源符号能载荷的最大信息量。
- 若对长为L的信源符号序列进行定长编码，每个序列对应的码字长度为K，则

$$R' = \frac{K \times \log m \quad \mapsto K\text{长码字的最大信息量}}{L \quad \mapsto \text{信源符号序列长度}} = \overline{K} \times \log m$$

比特/信源符号

编码效率

编码效率： $=\dfrac{\text{要求平均每个信源符号携带的实际信息量}}{\text{编码后平均每个信源符号的最大可能载信量}}$

$=\dfrac{\text{最小可能码长}}{\text{编码后的实际码长}}$

对于等长编码 $\eta=\dfrac{H(X)}{R'}=\dfrac{H(X)}{\frac{K}{L}\log m}=\dfrac{H(X)}{H(X)+\varepsilon}$

说明：

编码效率是小于或等于1的数。对同一信源，平均码长越短，信息传输率就越高，编码效率也越接近1。

编码效率可以用来衡量各种编码方法在有效性方面的优劣。

编码效率分析

观察正定理： $\frac{K \cdot \log m}{L} \geq H(X) + \varepsilon$  $\frac{K_{min} \cdot \log m}{L} = H(X) + \varepsilon$

当 $\varepsilon = 0$ **时，** $\frac{K_{min} \cdot \log m}{L} = H(X)$

ε的物理含义：

说明：

折算后平均每个信源符号携带的最大可能信息量等于要求携带的实际信息量。

此时编码效率为100%**。**

ε越大，编码效率越低。

编码效率与扩展次数L的关系

问题：什么时候编码效率趋近1？

$$\frac{K \cdot \log m}{L} \geq H(X) + \varepsilon$$

- 回答：定长编码定理中，只有在L足够大的时候，必可使译码差错小于δ编码效率才能趋近于1
- 经计算，当允许错误概率P_E小于δ时，信源序列长度L必满足

$$L \geq \frac{\sigma^2(x)}{\varepsilon^2 \delta}$$

$$\sigma^2(x) = D[I(x_i)] = \sum_{i=1}^{n} p(x_i)[\log p\,(x_i)]^2 - [H(X)]^2$$

定理的应用—例1

例1　设有离散无记忆信源

$$\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 \\ 0.4 & 0.18 & 0.1 & 0.1 & 0.07 & 0.06 & 0.05 & 0.04 \end{bmatrix}$$

要求编码效率$\eta = 90\%$**，译码错误概率**$\delta \le 10^{-6}$**，求需要的信源序列长度** L **。**

解：　分析计算步骤　　$L \ge \dfrac{D[I(x_i)]}{\varepsilon^2 \delta}$　　$\eta = \dfrac{H(X)}{H(X)+\varepsilon}$

(1) 计算自信息量的数学期望 $H(X)$ **和方差** $D[I(x_i)]$

$$H(X) = -0.4\log 0.4 - \cdots - 0.04\log 0.04 \approx 2.55$$ **比特/信源符号**

$$D[I(x_i)] = \sum_{i=1}^{8} p(x_i)\cdot[I(x_i) - H(X)]^2 \approx 1.31$$ **比特²/信源符号²**

定理的应用—例1（续）

或者利用方差的简便计算公式：

$$D[I(x_i)] = E[I^2(x_i)] - E^2[I(x_i)] = E[I^2(x_i)] - H^2(X)$$

(2) 根据要求的编码效率 η 计算 ε

$$\eta = \frac{H(X)}{H(X)+\varepsilon} \Rightarrow \varepsilon = \frac{1-\eta}{\eta}\cdot H(X) = \frac{1-0.9}{0.9}\cdot 2.55 \approx 0.28$$ **比特/信源符号**

(3) 代入公式，计算信源符号序列长度 L 。

$$L \geq \frac{D[I(x_i)]}{\varepsilon^2\delta} = \frac{1.31}{0.28^2\cdot 1\times 10^{-6}} \approx 1.63\times 10^7$$

定理的应用—例2

- 例2 设离散无记忆信源 $\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ \frac{3}{4} & \frac{1}{4} \end{bmatrix}$，采取等长二元编码时，要求编码效率$\eta = 0.96$，允许错误概率$\delta \leq 10^{-5}$。

$$L \geq \frac{\sigma^2(x)}{\varepsilon^2 \delta}$$

- 求得：H(X)=0.811

由$\eta = \frac{H(X)}{H(X)+\varepsilon}$,得$\varepsilon = 0.034$

又可计算：$\sigma^2(x) = D[I(x_i)] = 0.47$

故，可得出：$L \geq \frac{\sigma^2(x)}{\varepsilon^2 \delta} \approx 4.13 \times 10^7$

分　析

- **在编码效率和译码差错概率都不十分苛刻的情况下，所需的符号长度也是非常惊人的。一般只有在信源接近等概率分布时，算出的符号序列长度才是实际可接受的。**

对定长编码定理应用范围的说明：

虽然定长编码定理的推导过程中要求信源是无记忆信源，但所得结论可推广到有记忆信源。

	无记忆信源	有记忆信源
正定理	$\frac{L\cdot\log m}{N}\geq H(X)+\varepsilon$	$\frac{\boldsymbol{L\cdot\log m}}{\boldsymbol{N}}\geq \boldsymbol{H_\infty(X)+\varepsilon}$
逆定理	$\frac{L\cdot\log m}{N}\leq H(X)-2\varepsilon$	$\frac{\boldsymbol{L\cdot\log m}}{\boldsymbol{N}}\leq \boldsymbol{H_\infty(X)-2\varepsilon}$

多符号信源： $H_\infty(X)\approx\frac{H(X_1X_2\cdots X_N)}{N}$

马尔可夫信源： $H_\infty(X)\approx H_{m+1}$

谢谢！

黑晓军
华中科技大学
电子信息与通信学院
Email: heixj@hust.edu.cn
网址： http://eic.hust.edu.cn/aprofessor/heixiaojun