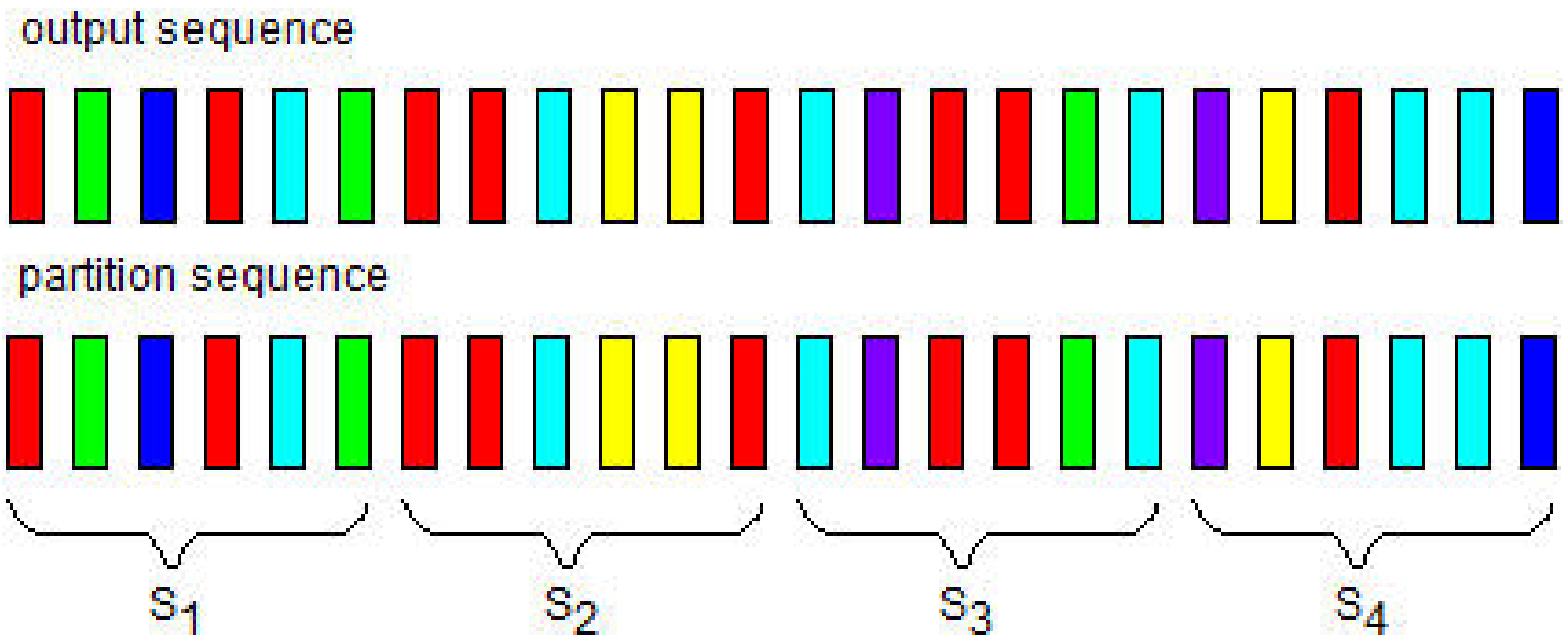
基础信息论

变长编码定理

华中科技大学电信学院

定长码存在的问题



- 非均匀分布的离散信源，收敛速度慢
- 对长子串进行定长编码，需要更大内存
- 对长子串进行定长编码，需要更大的编解码时延

→ 变长码

学习目标

- 分析变长编码比定长编码的优势
- 写出变长码唯一可译码的条件
- 定义紧致码
- 记忆单符号信源的变长编码定理
- 应用变长编码定理，给定编码效率要求，计算变长码的最小长度

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

变长编码的必要性及付出代价

变长码唯一可译码的条件

变长信源编码定理

变长编码的必要性

定长编码在理论上可以达到很高的编码效率，但是从定长码设计案例，在编码效率、错误概率要求较高的情况下，扩展次数L（**定长编码需要的符号数**）需要非常大，这在实际工程中是无法实现的。

当L有限时，高传输效率的等长码往往要**引入一定的失真和错误**，它不像变长码那样可以实现无失真编码。

➡ 在实际过程中，普遍使用**变长编码**。

变长编码付出的代价

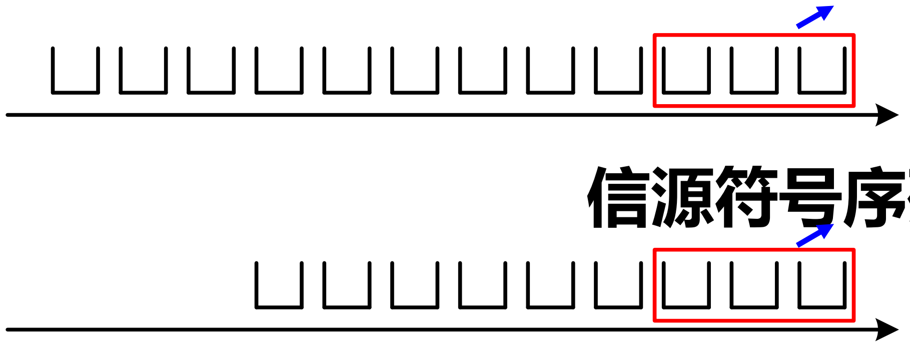
定长码

（例：$L=3$ ）

译码为

(1) 译码时需要同步。

信源符号序列1



变长码通常需要专门的同步通道。

⋮ 不需要同步通道

(2) 可能遇到译码延迟。

延迟问题：可利用编成即时码解决。

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

变长编码的必要性及付出代价

变长码唯一可译码的条件

变长信源编码定理

变长码唯一可译码的条件

唯一可译码的条件: 变长码必须是非奇异码，而且任意有限长L次扩展码也应该是非奇异码。

为了能够即时译码，变长码必须是即时码

即时码的判定问题

问题1: 如何判断一组由给定码字构成的码是否为即时码?

回答: 要求任何一个码字都不是其它码字的前缀。

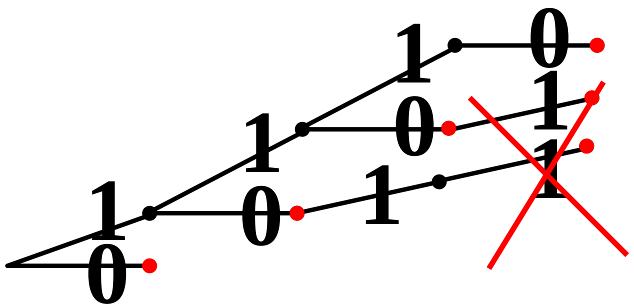
即时码的判定

判断任何一个码字是不是其它码字的前缀的方法?

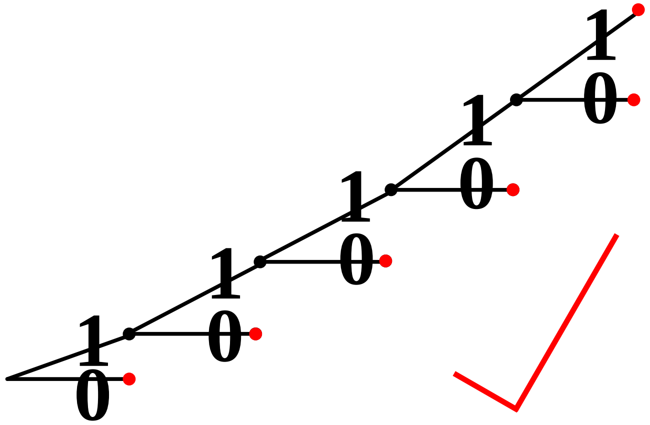
将给定的一组码字转成码树图的形式：
若所有码字均对应从根节点到终节点的联枝,则为即时码;
若某个码字位于另一码字对应节点之后，则为非即时码。

实例：判断以下码是否为即时码?

0
10
110
1110
1011
1101



0
10
110
1110
11110
11111



即时码的条件

问题2: 当未给定码字，如何判断能否构造出一种即时码？

定理 设信源符号集$X=(x_1,x_2,\ldots x_n)$，码符号集为$Y=(y_1,y_2,\ldots y_m)$，对信源进行编码，相应的码字为$W=(W_1,W_2,\ldots W_n)$，其分别对应的码长为 $k_1,k_2,\ldots k_n$，则即时码存在的**充要条件**是：

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

Kraft （克拉夫特）不等式

描述了信源符号数和码字长度之间满足什么条件才能构成**即时码**。

说明：给定码字个数，码符号集中的符号个数, 各码字的码元长度，满足**Kraft（克拉夫特）不等式**的构成一类码，这类码中一定至少有一个是即时码，但并不是该类码中的每个码都是即时码。

即时码的条件

➡ 即时码一定满足Kraft不等式，但反过来则不一定。

实例：仅举一反例即可。有如下码字：

$$\omega_1 = 1, \omega_2 = 01, \omega_3 = 011, \omega_4 = 0001$$

由于 ω_2 是 ω_3 的前缀，为非即时码。

但代入$n = 4, m = 2, k_1 = 1, k_2 = 2, k_3 = 3, k_4 = 4$，得：

$$\sum_{i=1}^{n} m^{-k_i} = 0.9375 < 1$$，满足Kraft不等式。

另外：　Kraft不等式同样适用于唯一可译码的判别。

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

变长编码的必要性及付出代价

变长码唯一可译码的条件

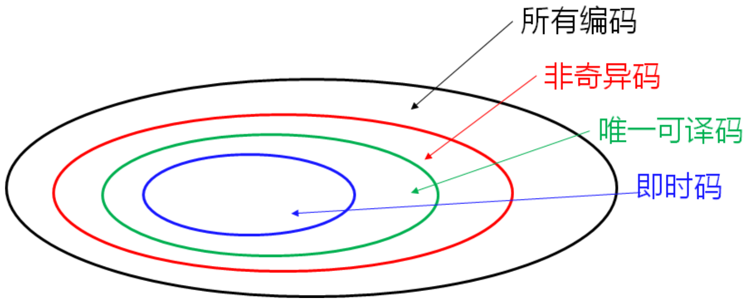
变长信源编码定理

紧致码的定义

对同一信源，用同一码符号集编成的即时码或唯一可译码可能有很多种。从高效传输信息的角度出发，当然是希望寻找平均码长最短的编码。

$$\begin{bmatrix} X \\ P(X) \\ W \\ K \end{bmatrix} = \begin{bmatrix} x_1 & \cdots & x_i & \cdots & x_n \\ p(x_1) & \cdots & p(x_i) & \cdots & p(x_n) \\ \omega_1 & \cdots & \omega_i & \cdots & \omega_n \\ k_1 & \cdots & k_i & \cdots & k_n \end{bmatrix}$$

因为是唯一可译码，信源符号x_i和码字W_i是一一对应的



平均码长： $\bar{K} = \sum_{i=1}^{n} p(x_i) \cdot k_i$ 码符号/信源符号

紧致码: 对于给定的信源和码符号集，若存在一个唯一可译码，其平均码长小于所有其它唯一可译码的平均码长，则称为紧致码。（**最佳码**）

信息传输率和信息传输速率

信息传输率　经信源编码后，平均每个**码符号**所携带的信息量。

单位：比特/码符号

如何计算?

$$\frac{\text{比特/信源符号}}{\text{码符号/信源符号}} = \frac{H(X)}{\bar{K}} = R$$

信息传输速率：单位时间传输的信息量。（设:传输一个码符号平均需要t秒）

$$R_t = \frac{H(X)}{\bar{K}\,t} \quad \text{比特/秒}$$

R_t越大，信息传输率就越高。信源熵H(X)是确定的，因此，提高信息传输率的方法是使平均码长最短。

单符号信源的变长编码定理

若单符号信源的熵为$H(X)$，码符号集中的符号个数为 m，则总可以找到一种无失真编码方法，构成唯一可译码，使其平均码长满足：

$$\boxed{\frac{H(X)}{\log m} \le \bar{K} < \frac{H(X)}{\log m} + 1}$$

即：$\bar{K}_{紧致码} \in \left[\frac{H(X)}{\log m}, \frac{H(X)}{\log m} + 1\right)$

问题：何时能取到最小的 $\bar{K} = \dfrac{H(X)}{\log m}$ （即上式左边取等号）

回答：要求 $p(x_i) = m^{-k_i}$。 $\boxed{k_i \text{ 必须为整数}}$

实例：对二元码，要求所有消息的概率必须是$\frac{1}{2}, \frac{1}{4}, \frac{1}{8} \cdots$

只有在上述条件下，唯一可译码的平均码长$\bar{L}$才能达到下限值 $\dfrac{H(X)}{\log m}$，而且可以保证所编得的码一定为**紧致码**。

单符号信源的变长编码示例

例 对 $\begin{bmatrix} X \\ P(X) \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{bmatrix}$ 编二元紧致码。

解： 算出各码字的码长，再利用码树图法进行构造。

(1) 计算各码字码长

$$k_i = \log_m^{\frac{1}{p(x_i)}}$$

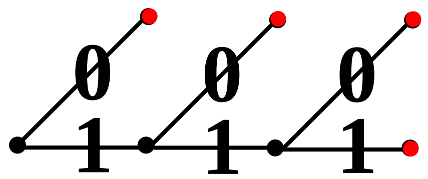
$$k_1 = \log_2^2 = 1$$

$$k_2 = \log_2^4 = 2$$

$$k_3 = \log_2^8 = 3$$

$$k_4 = \log_2^8 = 3$$

(2) 利用码树图法构造



$$\omega_1 = 0, \omega_2 = 10,$$

$$\omega_3 = 110, \omega_4 = 111$$

验证上述码为紧致码：

唯一可译码的码长下限： $\dfrac{H(X)}{\log m} = 1.75$

平均码长: $\bar{K} = 1.75$

无记忆信源 L 次扩展信源的变长编码定理

若 L 次扩展信源的熵为 $H(X^L)$，码符号集中的符号个数为 m，对信源 X^L 进行编码，总可以找到一种无失真编码方法，构成唯一可译码，使其平均码长满足：

$$\boxed{\frac{H(X)}{\log m} \le \frac{\bar{K}_L}{L} < \frac{H(X)}{\log m} + \frac{1}{L}}$$

单符号信源

$$\begin{array}{ccc} x_1 & \omega_1 & l_1 \\ x_2 & \omega_2 & l_2 \\ \vdots & \vdots & \vdots \\ x_n & \omega_n & l_n \end{array}$$

$$\bar{L} = \sum_{i=1}^{n} p(x_i) \cdot l_i$$

L 次扩展信源

$$\begin{array}{ccc} a_1 = x_1 x_1 \ldots x_1 & \omega_1 & k_1 \\ a_2 = x_1 x_1 \ldots x_2 & \omega_2 & k_2 \\ \vdots & \vdots & \vdots \\ a_q = x_n x_n \ldots x_n & \omega_q & k_q \end{array}$$

$q = n^L$

$$\bar{K}_L = \sum_{i=1}^{q} p(a_i) \cdot k_i$$

是每个 α_i 所对应的平均码长

无记忆信源L次扩展信源的变长编码定理

$$\boxed{\frac{H(X)}{\log m} \le \frac{\bar{K}_L}{L} < \frac{H(X)}{\log m} + \frac{1}{L}}$$

$\frac{\bar{K}_L}{L}$：离散无记忆信源X中每个信源符号x_i所对应的平均码长

与单符号信源进行比较：$\frac{H(X)}{\log m} \le \bar{K} < \frac{H(X)}{\log m} + 1$

$\frac{\bar{K}_L}{L}$趋于更短

特别地，当$L \to \infty$ 时：$\frac{\bar{K}_L}{L} \to \frac{H(X)}{\log m}$ 所得码一定是极致码。

变长无失真信源编码定理说明，只要编码后的码符号序列所能携带的信息量不小于信源本身的信息量，就可以实现**唯一可译码**。

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

单符号信源

$$\eta = \frac{H(X)}{\bar{K} \cdot \log m}$$

扩展信源

$$\eta = \frac{H(X)}{\dfrac{\bar{K}_L}{L} \cdot \log m}$$ *

【分母变大，结果变小】

$$\eta > \frac{H(X)}{\left[\dfrac{H(X)}{\log m} + \dfrac{1}{L}\right] \cdot \log m} = \frac{H(X)}{H(X) + \dfrac{\log m}{L}}$$

可用来估算，为达到效率η，所需要的信源符号序列长度L。

例

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

要求编码效率 $\eta = 90\%$，对二元变长编码，求需要的信源序列长度。

解： $\eta > \dfrac{H(X)}{H(X) + \dfrac{\log m}{L}}$

代入： $\eta = 90\% \quad m = 2$

$H(X)=2.55$ 比特/符号

计算得： $L \approx 4$

比较：定长编码 $L \approx 1.63 \times 10^7 \quad \delta \le 10^{-6}$

问题：为什么变长编码可做到完全无失真编码?

回答：因为变长编码可保证信源符号序列和码符号序列一一对应。

对变长编码定理应用范围的说明：

虽然变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。

无记忆信源的扩展信源

$$\frac{H(X)}{\log m} \le \frac{\bar{K}_L}{L} < \frac{H(X)}{\log m} + \frac{1}{L}$$

有记忆信源

$$\frac{H_\infty(X)}{\log m} \le \frac{\bar{K}_L}{L} < \frac{H_\infty(X)}{\log m} + \frac{1}{L}$$

多符号信源： $H_\infty(X) \approx \dfrac{H(X_1 X_2 \cdots X_N)}{L}$

马尔可夫信源： $H_\infty(X) \approx H_{m+1}$

此外，定长编码是变长编码的一个特例，定长编码定理也可以统一到香农第一定理。

变长编码的编码信息率R'

- 定义变长编码的编码信息率为：

$$R' \triangleq \frac{\overline{K_L}}{L}\log m$$

- 它表示编码后平均每个信源符号能载荷的最大信息量。
- 香农第一定理又可以表述为：
  - 若 $H(X) \le R' < H(X) + \varepsilon$，就存在唯一可译的变长编码。
  - 若 $R' < H(X)$，则不存在唯一可译的变长编码，不能实现无失真的信源编码。

信息传输率R

- 从信道角度看，信道的信息传输率

$$R=\frac{H(X)}{\overline{K}}\left(\frac{\text{比特}/\text{信源符号}}{\text{码符号}/\text{信源符号}}\right)=\frac{H(X)}{\overline{K}}(\text{比特}/\text{码符号})$$

因为$\overline{K}=\frac{\overline{K_L}}{L}\geq\frac{H(X)}{\log m}$ 所以$R\leq\log m$

当平均码长$\overline{K}$达到极限值$\frac{H(X)}{\log m}$时，

编码后的信道传输率$R=\log m(\text{比特}/\text{码符号})$

此时，R等于无噪无损信道的信道容量C，信息传输率最高。

编码效率和剩余度

编码效率定义为

$$\eta = \frac{H(X)}{R'} = \frac{H(X)}{\overline{K} \log m}$$

说明：

编码效率 η 一定是小于或等于1的数，平均码长越短，越接近它的极限值 $\frac{H(X)}{\log m}$，那么编码效率就越高。

定义码的**剩余度**为

$$\gamma = 1 - \eta = 1 - \frac{H_m(X)}{\overline{L}}$$

变长编码举例

设有一离散无记忆信源：$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \\ p_1 = 3/4 & p_2 = 1/4 \end{bmatrix}$

其熵为$H(X) = \frac{3}{4}\log\frac{4}{3} + \frac{1}{4}\log 4 = 0.811$ 比特/信源符号

现在我们用二元符号(0,1)来构造一个即时码

$$x_1 \to 0, x_2 \to 1$$

这时，平均码长$\overline{K} = 1$ 二元码符号/信源符号

编码的效率为$\eta = \frac{H(X)}{\overline{K}} = 0.811$

得信道的信息传输率为$R = 0.811$ 比特/二元码符号

变长编码举例—续

进一步，我们对信源 X 的二次扩展信源 X^2 进行编码。

其二次扩展信源 X^2 和即时码如下表所列

α_i	$p(\alpha_i)$	即时码
x_1x_1	9/16	0
x_1x_2	3/16	10
x_2x_1	3/16	110
x_2x_2	1/16	111

变长编码举例—续

这个码的平均长度

$$\overline{K_2}=\frac{9}{16}\times1+\frac{3}{16}\times2+\frac{3}{16}\times3+\frac{1}{16}\times3=\frac{27}{16}$$ **二元码符号/二个信源符号**

信源 X 中每一单个符号的平均码长为：

$$\overline{K}=\frac{\overline{K_2}}{2}=\frac{27}{32}$$ **二元码符号/信源符号**

其编码效率： $\eta_2=\dfrac{H(X)}{\overline{K}}=\dfrac{32\times0.811}{27}=0.961$

得 R_2=0.961 比特/二元码符号

编码虽然复杂了一些，但信息传输效率有了较大提高。

变长编码举例—续

- 用同样方法可进一步对信源X的三次和四次扩展信源进行编码，并求出其编码效率为：
  - η_1=0.811比特/二元码符号
  - η_2=0.961比特/二元码符号
  - η_3=0.985比特/二元码符号
  - η_4=0.991比特/二元码符号
- 对于同一信源，要求编码效率都达到96%，比较
  - 变长码只需对二次扩展信源（L = 2）进行编码；
  - 而等长码则要求L大于$4.13*10^7$.
- 很明显，用变长码编码时，L不需很大就可以达到相当高的编码效率，而且可实现无失真编码。

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