信息论导论

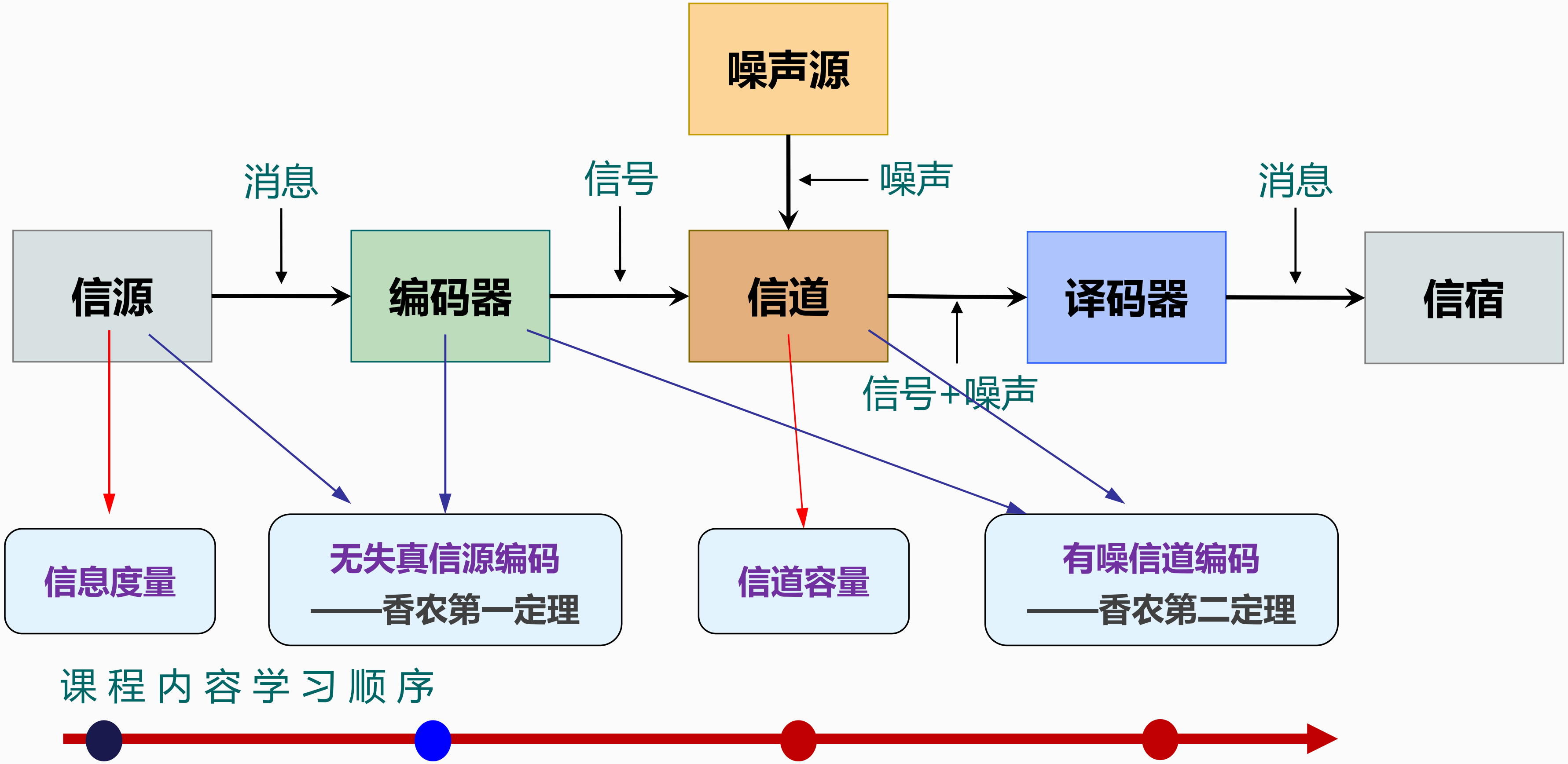
第3讲 信源编码的基本概念、最优信源编码的码长

[信息论教材中页码范围] 信源编码、非奇异码、唯一可译码、即时码：p103~p107，Kraft不等式：p107~p110 & p115~p118，最优编码码长、Shannon码：p110~p112 & p115

信息学部-信息科学与技术学院 吴绍华

hitwush@hit.edu.cn

课程内容进度安排



作为导论课，本课程只讨论“离散”信源、信道

信源编码的基本概念

- **信源编码**：是一个从消息到码符号串的映射，记为 $C : \mathscr{X} \to \mathscr{D}^+$
  - $\mathscr{X}$———消息集合
  - $\mathscr{D}^+$———码符号集合 $\mathscr{D}$（亦称作编码字母表 $\mathscr{D}$）上任意有限长码符号串所构成的集合
  - $\mathscr{D}$ 在数字通信系统中默认是二进制的，即 $\mathscr{D} = \{0, 1\}$
  - 举例：$\{\mathrm{E}, \mathrm{F}, \mathrm{G}\} \to \{0, 1\}^+$：$C(\mathrm{E}) = 0, C(\mathrm{F}) = 10, C(\mathrm{G}) = 11$

- **信源编码的扩展**：是一个从消息串到码符号串的映射，记为 $C^+ : \mathscr{X}^+ \to \mathscr{D}^+$
  - 具体操作：将消息串中各个消息 x_i 的编码结果 $C(x_i)$ **不间断的串联**起来，即得到**信源编码的扩展**
  - 举例：$C^+(\mathrm{EFEEGE}) = 01000110$

非奇异码、唯一可译码

- **非奇异码：** $x_1 \neq x_2 \Rightarrow C(x_1) \neq C(x_2)$
  - 非奇异码可以无歧义地表示任意单个消息符号的值

- **唯一可译码：** 若 C^+ 是非奇异码，则 C 是唯一可译码
  - 对 $C^+(x^+)$ 的译码不会产生歧义，即：对任意信源编码结果及其扩展的译码结果是唯一的

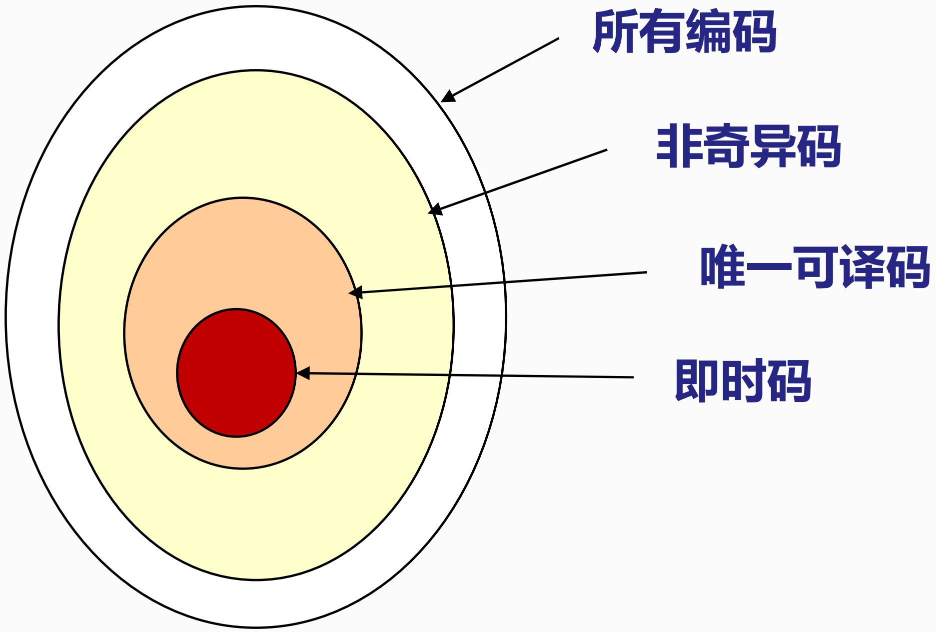
- 尽管“唯一可译”，然而，部分唯一可译码的译码效率可能存在问题：
  - 译码过程中，需“延迟确认”（举例说明）
  - 极端情形下，甚至需要**延迟**至整串编码结果尾部，才可**确认**译码结果中的第一个消息符号的值
  - 为提高译码效率，可在两个码字间**添加间断符号**，但显然这会导致编码效率降低
  - 更好的思路：设计出“**自间断码**”，以提高译码效率的同时不损失编码效率

即时码

- **即时码（又称前缀码）**
  - 任何码字都不是其他码字的前缀
  - 在某个码字出现后，无需延迟往后检索，可立刻译出
- **即时性 $\Rightarrow$ 唯一可译性 $\Rightarrow$ 非奇异性**

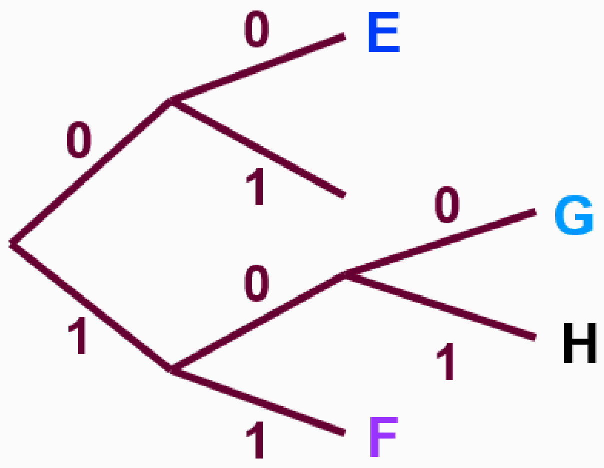
例

1.$C(\mathrm{E},\mathrm{F},\mathrm{G},\mathrm{H}) = (0, 11, 00, 11)$
2.$C(\mathrm{E},\mathrm{F}) = (0, 101)$
3.$C(\mathrm{E},\mathrm{F}) = (1, 101)$
4.$C(\mathrm{E},\mathrm{F},\mathrm{G},\mathrm{H}) = (00, 01, 10, 11)$
5.$C(\mathrm{E},\mathrm{F},\mathrm{G},\mathrm{H}) = (0, 01, 011, 111)$
答：2、3、4、5 是唯一可译码，其中 2、4 还是前缀码



码树

- 即时码：$C(\mathrm{E},\mathrm{F},\mathrm{G},\mathrm{H})=(00,11,100,101)$
- 码树的构建：D 元字母表（即码符号集合大小为 D，$D=|\mathscr{D}|$）上的编码，对应 D 叉树:
  - 从根节点开始，每一节点都可长出 D 个子节点，在对应的 D 个分支上分别标记出 D 个码符号
  - 即时码的码字，对应码树上的叶子节点
  - 码树上的中间节点不能用作码字：因为中间节点是其长出的叶子节点的前缀
  - 允许有部分叶子不被使用，即叶子节点可以不被用完。如果某个即时码对应的码树上，所有的叶子均用完，则有：$|\mathscr{X}|-1$ 是 $D-1$ 的整数倍



$$111011000000 \rightarrow \mathrm{FHGEE}$$

Kraft不等式（for即时码）

定理 5.2.1 Kraft 不等式

（正定理）对于 D 元字母表上的任意即时码（前缀码），码字长度 $l_1, l_2, ..., l_{|\mathscr{X}|}$ 必定满足不等式

$$\sum_{i=1}^{|\mathscr{X}|} D^{-l_i} \leq 1$$

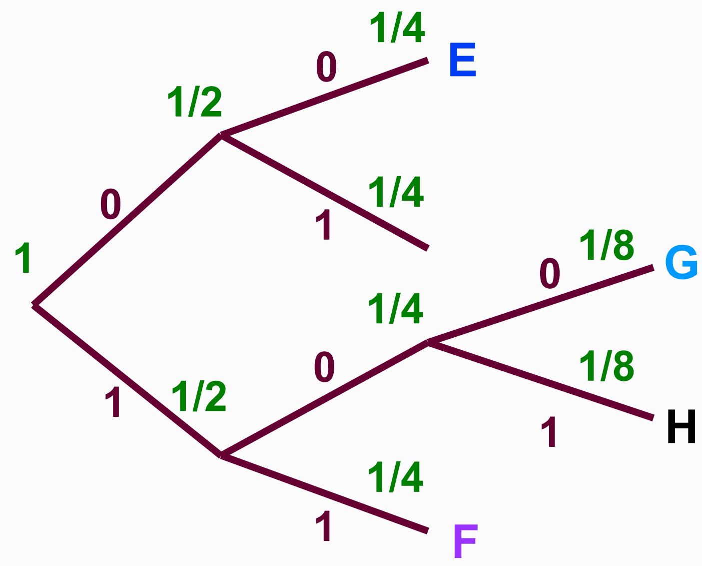
（逆定理）给定一组码字长度，若满足上述不等式，则存在对应此组码字长度的即时码。

对于即时码，由Kraft不等式可知，码字长度不可能全部都很短

Kraft不等式（for即时码）的证明

哈爾濱工業大學(深圳) HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

- 以 $D = 2$ 为例，可以构造一棵二叉树
- 深度为 l 的节点标为 2^{-l}
- 每个节点的值等于它所有叶子值的总和
- 显然 Kraft 不等式成立
- 并且当所有叶子节点均被利用时，等号成立
- 2^{-l} 可理解为编码预算。总的编码预算为 1:
  - 码字 00 使用了 $1/4$ 预算
  - 码字 100 使用了 $1/8$ 预算
- 对于 $D \neq 2$ 的情形，道理完全一样，结论同样成立
- **逆定理显然是成立的**



Kraft不等式（for唯一可译码）

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

定理 5.5.1 McMillan 不等式

（正定理）对于 D 元字母表上的任意唯一可译码，码字长度 $l_1, l_2, \ldots, l_{|\mathscr{X}|}$ 必定满足不等式

$$\sum_{i=1}^{|\mathscr{X}|} D^{-l_i} \leq 1$$

和即时码一样！

（逆定理）给定一组码字长度，若满足上述不等式，则存在对应此组码字长度的唯一可译码。

直接启示 —— 唯一可译码相比即时码并不能进一步减少码字长度

Kraft不等式（for唯一可译码）的证明

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

证明.

令 $S=\sum_{i=1}^{|\mathscr{X}|} D^{-l_i}, M=\max\{l_i\}, m=\min\{l_i\}$, 则对于任意 N,

$$S^N=\left(\sum_{i=1}^{|\mathscr{X}|} D^{-l_i}\right)^N=\sum_{i_1=1}^{|\mathscr{X}|}\sum_{i_2=1}^{|\mathscr{X}|}\cdots\sum_{i_N=1}^{|\mathscr{X}|} D^{-(l_{i_1}+l_{i_2}+\cdots+l_{i_N})}$$

$$=\sum_{x^+\in\mathscr{X}^N} D^{-\underline{\mathrm{len}}\{C^+(x^+)\}}=\sum_{l=Nm}^{NM} D^{-l}\left|\left\{x^+:\underline{\mathrm{len}}\left\{C^+\left(x^+\right)\right\}=l\right\}\right|$$

$$\le\sum_{l=Nm}^{NM} D^{-l}D^{l}=\sum_{l=Nm}^{NM} 1=N(M-m)$$

取码字长度

$S^N\le N(M-m)$ 对于任意 N 均成立，包括 $N\to\infty$，所以必然有 $S\le 1$ □

最优信源编码能有多短?

- (最优码的定义) 令 $l(x) = \mathrm{len}(C(x))$，当 $L = El(x)$ 最小时，认为 C 是最优的
- 我们可以建立优化问题对最优码长进行求解：目标是最小化 $\sum_{x\in\mathscr{X}} p(x)l(x)$，约束条件包括
  1. $\sum_{x\in\mathscr{X}} D^{-l(x)} \le 1$
  2. 所有 $l(x)$ 均为整数
- 将约束条件简化（松弛）:
  - 忽视条件 2，且假设条件 1 中取等号
  - 约束放宽后，求得的码长可能是比满足原优化问题的码长要短的，因此所求码长为实际码长的**下界**

松弛后的优化问题

Minimize $\sum_{i=1}^{|\mathscr{X}|} p(x_i)l_i$，subject to $\sum_{i=1}^{|\mathscr{X}|} D^{-l_i} = 1$

求解最优码长（约束条件简化后）

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

解答.

使用拉格朗日乘子法:

定义 $J=\sum_{i=1}^{|\mathscr{X}|}p(x_i)l_i+\lambda\sum_{i=1}^{|\mathscr{X}|}D^{-l_i}$，令偏导数 $\frac{\partial J}{\partial l_i}=0$

$$\frac{\partial J}{\partial l_i}=p(x_i)-\lambda\ln(D)D^{-l_i}=0\Rightarrow D^{-l_i}=\frac{p(x_i)}{\lambda\ln(D)}$$

由 $\sum_{i=1}^{|\mathscr{X}|}D^{-l_i}=1\Rightarrow\lambda=1/\ln(D)\Rightarrow\boxed{l_i=-\log_D(p(x_i))}$

此时期望码长

$$\boxed{El(x)=E(-\log_D(p(x)))=H_D(X)}=\frac{H(X)}{\log_2 D}$$

最优码长定理

定理 5.3.1

随机变量 X 的任意 D 元唯一可译码（或即时码）的期望码长大于或等于熵 $H_D(X)$，即

$$L \geq H_D(X)$$

当且仅当 $D^{-l_i} = p(x_i)$ 时，等号成立

证明.

$$L - H_D(X) = E\boxed{l(x)} + E\log_D p(x) = E\left(\boxed{-\log_D D^{-l(x)}} + \log_D p(x)\right)$$

令 $c = \sum_{i=1}^{|\mathscr{X}|} D^{-l_i} \leq 1$，$q(x) = \frac{D^{-l(x)}}{c}$，可以得到

D-adic（D进制）概率分布

$$L - H_D(X) = E\left(-\log_D q(x) + \log_D p(x) - \log_D c\right)$$
$$= E\left(\log_D \frac{p(x)}{q(x)}\right) - \log_D c = D(\mathbf{p}||\mathbf{q}) - \log_D c \geq 0$$

当且仅当 $c = 1$ 且 $\mathbf{p} = \mathbf{q}$ 时取等号，此时有 $D^{-l(x)} = p(x)$ □

由前述证明直接能想到的信源编码方法

- 设 $l_1^*, l_2^*, \cdots, l_m^*$ 是 D 元字母表上关于信源分布 $\mathbf{p}$ 的一组信源编码码长:
  - 如果 $\mathbf{p}$ 是整进制的，则 $l_i^* = -\log_D(p(x_i))$
  - 如果 $\mathbf{p}$ 不是整进制的，即 $-\log_D(p(x_i))$ 不是整数，我们总可以找到一个接近 $\mathbf{p}$ 的整进制分布，然后用其计算码长 $\{l_1^*, l_2^*, \cdots, l_m^*\}$

- 然而寻找最接近 $\mathbf{p}$ 的整进制分布并不简单。我们可以采用一种次优方法——Shannon 码。

Shannon码

由于 $-\log_D(p(x_i))$ 可能不是整数，我们可以直接通过向上取整的方式得到整数码长，即 $l_i = \lceil -\log_D(p(x_i)) \rceil$

Shannon编码

- 由下式可知 l_i 满足 Kraft 不等式

$$\sum D^{-l_i} = \sum D^{-\lceil -\log_D(p(x_i)) \rceil} \le \sum D^{\log_D(p(x_i))} = \sum p(x_i) = 1$$

- 进一步地，我们可以依据 Shannon 码长构造出相应即时码：可以使用码树，或参照第五章习题 25 中的方法来构造。
- 期望码长 L_s：由于 $-\log_D(p(x_i)) \le l_i \le -\log_D(p(x_i)) + 1$，两边同时求期望可得期望码长为

$$H_D(X) \le L_s < H_D(X) + 1$$

Shannon-Fano
编码

最优信源编码的码长范围/上下界（逐符号编码）

定理 5.4.1

设 L^* 是最优信源编码对应的期望码长 $L^* = \sum p(x_i)\, l_i^*$，则

$$H_D(X) \le L^* < H_D(X) + 1$$

证明.

- 由定理 5.3.1 可知，$L^* \ge H_D(X)$
- 令 L_s 表示 Shannon 码的期望码长，由于最优编码一定优于 Shannon 码，有 $L^* \le L_s$
- 总结可得 $H_D(X) \le L^* < H_D(X) + 1$

□

?

- 如何构造最优前缀码 (即时码)?
- 最优编码码长上界中的 1bit “额外开销”是否可以减少甚至消除?

阶段小结

哈爾濱工業大學(深圳)
HARBIN INSTITUTE OF TECHNOLOGY, SHENZHEN

- 关于 D 元码的 Kraft 不等式
  - 对于任意唯一可译码 C，有 $\sum_{i=1}^{|\mathscr{X}|} D^{l_i} \le 1$
  - 若 $\sum_{i=1}^{|\mathscr{X}|} D^{l_i} \le 1$，则可依此码长构造出即时码。
- 对于唯一可译码，有 $El(x) \ge H_D(X)$，当且仅当 $D^{-l(x)} = p(x)$ 时等号成立。
- Shannon 码：码长为 $l_i = \lceil -\log_D(p(x_i)) \rceil$，期望码长范围为 $H_D(X) \le El(x) < H_D(X) + 1$，这种编码方式是次优的
- 最优信源编码的期望码长上下界：$H_D(X) \le L^* < H_D(X) + 1$

结束