第五章补充 无失真信源编码

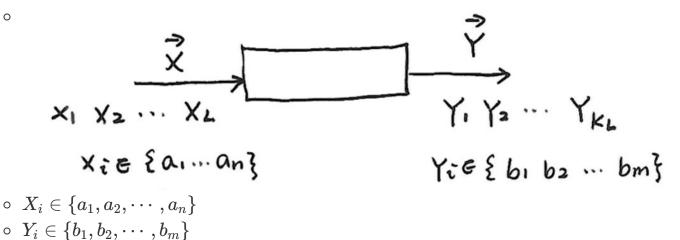
- 第五章补充 无失真信源编码
 - 。 新进均分性定理 (AEP)
 - 。 定长编码定理
 - 。 单符号变长编码定理
 - 。 离散平稳无记忆序列变长编码定理 (香农第一定理)

渐进均分性定理 (AEP)

• **定理**: $\vec{X} = (X_1 X_2 \cdots X_L)$, 为独立同分布 (i.i.d) 随机变量序列,具有渐近均分性质 (AEP, Asymptotic equipartition property) :

$$orall arepsilon > 0$$
,当 $L o \infty$ 时, $p(X_1, X_2, \cdots, X_L) o 2^{-LH(X)}$

• 信源编码模型:



- 证明1:
 - 互信息:

$$egin{aligned} I(ec{X}) &= -\log p(X_1, X_2, \cdots, X_L) \ &= -\log p(X_1) - \log p(X_2) - \cdots - \log p(X_L) \ &= \sum_{i=1}^L I(X_i) \end{aligned}$$

○ 均值:

$$E\{I(ec{X})\} = E\left\{\sum_{i=1}^{L}I(X_i)
ight\} = \sum_{i=1}^{L}E\{I(X_i)\} = \sum_{i=1}^{L}H(X_i) = LH(X)$$

○ 方差:

$$D\{I(\vec{X})\} = \sum_{i=1}^{L} D\{I(X_i)\} = LD(I(X))$$

其中

$$D(I(X)) = E\left[(I(X) - H(X))^2\right] = \sum_{i=1}^n p(x_i)(-\log p(x_i) - H(X))^2$$

。 根据切比雪夫不等式

$$P_r\{|X-\mu|\geq arepsilon\}\leq rac{\sigma^2}{arepsilon^2} \quad (arepsilon>0,\sigma>0)$$

有:

$$egin{aligned} P_r\left\{\left|I(ec{X})-LH(X)
ight|\geq Larepsilon
ight\} &\leq rac{LD(I(X))}{(Larepsilon)^2}\ P_r\left\{\left|rac{I(ec{X})}{L}-H(X)
ight|\geq arepsilon
ight\} &\leq rac{D(I(X))}{Larepsilon^2} \end{aligned}$$

当 $L \to \infty$ 时

$$egin{split} rac{D(I(X))}{Larepsilon^2} &
ightarrow 0, \ rac{-\log p(X_1,X_2,\cdots,X_L)}{L} &
ightarrow H(X), \ p(X_1,X_2,\cdots,X_L) &
ightarrow 2^{-LH(X)} \end{split}$$

- 证明2:
 - 。 AEP 也可由弱大数定律直接得到
 - \circ 当 X 为 i.i.d,n 很大时, $rac{1}{n}\sum_{i=1}^n x_i o E(X)$

$$egin{aligned} -rac{1}{L}\log p(X_1,X_2,\cdots,X_L) &= -rac{1}{L}\sum_{i=1}^L\log p(X_i) \ &\stackrel{a.s.}{
ightarrow} - E(\log p(X)) \ &= H(X) \end{aligned}$$

- 典型集A_ε
 - 定义

$$egin{aligned} A_arepsilon &= \left\{ec{x}: \left|rac{I(ec{x})}{L} - H(X)
ight| \leq arepsilon
ight\} \ A_arepsilon^c &= \left\{ec{x}: \left|rac{I(ec{x})}{L} - H(X)
ight| > arepsilon
ight\} \ A_arepsilon + A_arepsilon^c &= X^L = \{ec{x}_i\} \quad i = 1, 2, \cdots, n^L \end{aligned}$$

- 性质
 - a. 若 $\vec{x} \in A_{\varepsilon}$,则典型集中的元素几乎是等概率出现的:

$$2^{-L(H(X)+arepsilon)} \le p(\vec{x}) \le 2^{-L(H(X)-arepsilon)}$$

b. 当 L 充分大时,典型集的概率近似为1:

$$P_r(A_arepsilon) = \sum_{ec{x} \in A_arepsilon} p(ec{x}) > 1 - arepsilon$$

c. 典型集中元素的个数:

$$|A_arepsilon| \leq 2^{L(H(X)+arepsilon)}$$

■ 证明:

$$egin{aligned} 1 &= \sum_{ec{x} \in X^L} p(ec{x}) \geq \sum_{ec{x} \in A_arepsilon} p(ec{x}) \ &\geq \sum_{ec{x} \in A_arepsilon} 2^{-L(H(X) + arepsilon)} = |A_arepsilon| 2^{-L(H(X) + arepsilon)} \ \Rightarrow &|A_arepsilon| \leq 2^{L(H(X) + arepsilon)} \end{aligned}$$

d. 典型集中元素的个数:

$$|A_arepsilon| \geq (1-arepsilon) 2^{L(H(X)-arepsilon)}$$

当 L 充分大时, $P_r(A_{\varepsilon}) > 1 - \varepsilon$

■ 证明:

$$egin{aligned} 1-arepsilon & \leq P_r(A_arepsilon) = \sum_{ec{x} \in A_arepsilon} p(ec{x}) \ & \leq \sum_{ec{x} \in A_arepsilon} 2^{-L(H(X)-arepsilon)} = |A_arepsilon| 2^{-L(H(X)-arepsilon)} \ & \Rightarrow |A_arepsilon| \geq (1-arepsilon) 2^{L(H(X)-arepsilon)} \end{aligned}$$

定长编码定理

• **定理**: 对于由 L 个符号组成的,每个符号的熵为 $H_L(\vec{X})$ 的无记忆平稳符号序列 X_1,X_2,\cdots,X_L ,可用 K_L 个符号 Y_1,Y_2,\cdots,Y_{K_L} (每个符号有 m 种可能值,即m进制编码)进行定长编码。对于任意 $\varepsilon>0$, $\delta>0$,只要

$$rac{K_L}{L}\log m \geq H_L(ec{X}) + arepsilon$$

则当 L 足够大时,必可使译码差错小于 δ ; 反之,当

$$rac{K_L}{L}\log m \leq H_L(ec{X}) - 2arepsilon$$

时,译码差错一定是有限值,当 $L o\infty$ 时,译码几乎必定出错。

- 证明
 - \circ 对 A_{ε} 中的元素进行定长编码,为了唯一可译,码符号个数大于等于 $|A_{\varepsilon}|$,即

$$m^{K_L} \geq |A_arepsilon|$$

此时 $|A_{arepsilon}|$ 取上界,记平均符号熵 $H_L(ec{X})$ 为 H(X)

$$m^{K_L} \geq 2^{L(H(X)+arepsilon)} \ K_L \log m \geq L(H(X)+arepsilon) \ rac{K_L}{L} \log m \geq H(X)+arepsilon$$

此时,错误译码的概率为

$$P_e = P_r\{A_arepsilon^c\} \leq rac{D(I(X))}{Larepsilon^2}$$

$$rac{D(I(X))}{Larepsilon^2} \leq \delta \ L \geq rac{D(I(X))}{arepsilon^2 \delta}$$

。 反之, 若

$$egin{aligned} rac{K_L}{L} \log m & \leq H(X) - 2arepsilon \ \log m^{K_L} & \leq L(H(X) - 2arepsilon) \ m^{K_L} & \leq 2^{L(H(X) - 2arepsilon)} & < (1 - arepsilon) 2^{L(H(X) - arepsilon)} & = |A_arepsilon|_{min} \end{aligned}$$

即 $m^{K_L} < |A_{\varepsilon}|$,码字个数少于 A_{ε} 中的元素个数此时只能在 A_{ε} 中选取 m^{K_L} 个元素进行一对一编码,

$$egin{aligned} P_r(A_arepsilon egin{aligned} & P_r(A_arepsilon egin{aligned} & P_m(A_arepsilon eta_L ightarrow ar{\pi}_L ightarrow ar{\pi$$

译码错误概率为 $P_e=1-2^{-Larepsilon}$, $L o\infty$, $P_e o 1$

单符号变长编码定理

• **定理**: 对于离散无记忆信源的符号 X, 其熵为 H(X), 每个信源符号用 m 进制码元进行变长编码, 一定存在一种无失真编码方法, 其 (m 进制) 码元平均长度 \overline{k} 满足下列不等式:

$$\frac{H(X)}{\log m} \leq \overline{k} < \frac{H(X)}{\log m} + 1$$

其中, $\overline{k} = \sum_{i=1}^n p(a_i)l_i$, l_i 为符号 a_i 的码长。

- 证明:
 - \circ **定义**: l_i 为符号 a_i 的码长, $p(a_i)$ 为符号 a_i 的概率, $\overline{k} = \sum_{i=1}^n p(a_i) l_i$ 为码元平均长度。

$$\begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_n \\ p(\alpha_1) & p(\alpha_2) & \dots & p(\alpha_n) \end{bmatrix}$$

$$\frac{1}{3} \frac{1}{3} \frac{1}{3}$$

$$\bar{k} = \sum_{i=1}^{n} p(a_i) li$$

i. 证明必存在无失真编码,使得 $H(X) - (\log m)\overline{k} \leq 0$

$$egin{align} H(X) - (\log m)\overline{k} &= -\sum_{i=1}^n p(a_i)\log p(a_i) - \sum_{i=1}^n p(a_i)l_i\log m \ &= \sum_{i=1}^n p(a_i)\lograc{m^{-l_i}}{p(a_i)} \end{split}$$

根据Jensen不等式 $E[f(x)] \leq f(E(x))$,有

$$egin{align} H(X)-(\log m)\overline{k}&=\sum_{i=1}^n p(a_i)\lograc{m^{-l_i}}{p(a_i)}\ &\leq \log\sum_{i=1}^n p(a_i)rac{m^{-l_i}}{p(a_i)}=\log\sum_{i=1}^n m^{-l_i} \end{gathered}$$

唯一可译码存在(无失真),因此 $\sum_{i=1}^n m^{-l_i} \leq 1$

$$H(X) - (\log m)\overline{k} \le \log 1 = 0$$

 $\Rightarrow \overline{k} \ge \frac{H(X)}{\log m}$

若要等号成立

$$m^{-l_i} = p(a_i), \quad l_i = rac{-\log p(a_i)}{\log m}$$

ii. 证明 $\overline{k} < rac{H(X)}{\log m} + 1$

符号 a_i 的码长 l_i 需要是整数,因此取

$$rac{-\log p(a_i)}{\log m} \leq l_i < rac{-\log p(a_i)}{\log m} + 1$$

$$egin{split} rac{-\log p(a_i)}{\log m} & \leq l_i \ \Rightarrow -\log p(a_i) \leq \log m^{l_i} \ \Rightarrow p(a_i) \geq m^{-l_i} \ \Rightarrow \sum_{i=1}^n m^{-l_i} \leq \sum_{i=1}^n p(a_i) = 1 \end{split}$$

表明 l_i 满足唯一可译码存在的条件。

由

$$egin{aligned} l_i < rac{-\log p(a_i)}{\log m} + 1 \ &\sum_{i=1}^n l_i p(a_i) < \sum_{i=1}^n p(a_i) (rac{-\log p(a_i)}{\log m} + 1) \ &\overline{k} < rac{H(X)}{\log m} + 1 \end{aligned}$$

离散平稳无记忆序列变长编码定理 (香农第一定理)

• **定理**: 对于平均符号熵为 $H_L(\vec{X})$ 的离散平稳无记忆信源,必存在一种无失真编码方法,使平均信息率 \overline{K} 满足不等式:

$$H_L(ec{X}) \leq \overline{K} < H_L(ec{X}) + arepsilon$$

其中 arepsilon 为任意小正数, $\overline{K}=rac{\overline{k}_L}{L}\log m$, \overline{k}_L 为 L 个符号的平均码长。

• 证明

将 $ec{X}=(X_1X_2\cdots X_L)$ 作为一个整体,构成一个新的单符号信源。因此有:

$$egin{split} rac{LH_L(ec{X})}{\log m} & \leq \overline{k}_L < rac{LH_L(ec{X})}{\log m} + 1 \ H_L(ec{X}) & \leq rac{\overline{k}_L}{L} \log m < H_L(ec{X}) + rac{\log m}{L} \end{split}$$

令 $\overline{K} = rac{ar{k}_L}{L} \log m$,则有

$$H_L(ec{X}) \leq \overline{K} < H_L(ec{X}) + arepsilon \quad (L o \infty)$$

此时编码效率

$$\eta = rac{H_L(ec{X})}{\overline{k}_L} > rac{H_L(ec{X})}{H_L(ec{X}) + rac{\log m}{L}}$$