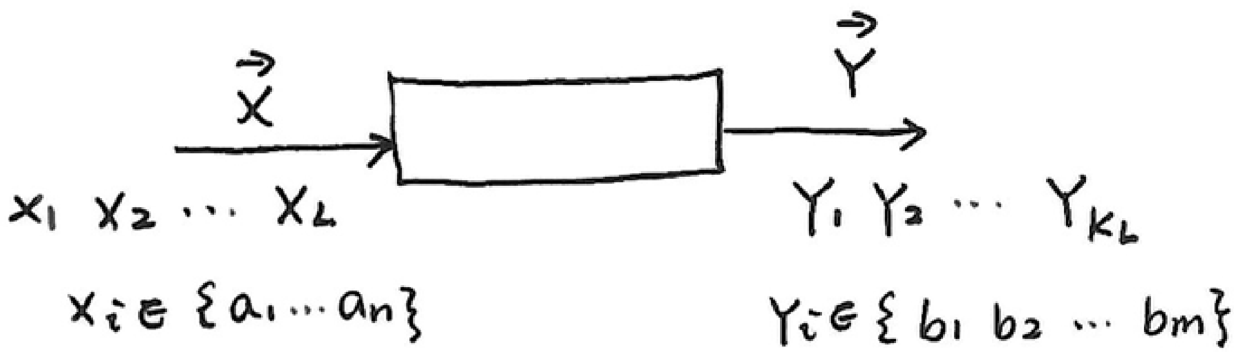
无失真信源编码

渐进均分性定理（AEP）

- **定理**：$\vec{X} = (X_1X_2\cdots X_L)$，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）：

$$\forall\varepsilon > 0\text{，当}L \rightarrow \infty\text{时，}p(X_1, X_2, \cdots, X_L) \rightarrow 2^{-LH(X)}$$

- **信源编码模型**：
  - 

  - $X_i \in \{a_1, a_2, \cdots, a_n\}$
  - $Y_i \in \{b_1, b_2, \cdots, b_m\}$
- **证明1**：
  - **互信息**：

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

- **均值**:

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

- **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

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

有:

$$P_r\left\{\left|I(\vec{X}) - LH(X)\right| \geq L\varepsilon\right\} \leq \frac{LD(I(X))}{(L\varepsilon)^2}$$
$$P_r\left\{\left|\frac{I(\vec{X})}{L} - H(X)\right| \geq \varepsilon\right\} \leq \frac{D(I(X))}{L\varepsilon^2}$$

当 $L \to \infty$ 时

$$\frac{D(I(X))}{L\varepsilon^2} \to 0,$$
$$\frac{-\log p(X_1, X_2, \cdots, X_L)}{L} \to H(X),$$
$$p(X_1, X_2, \cdots, X_L) \to 2^{-LH(X)}$$

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

$$
\begin{aligned}
-\frac{1}{L}\log p(X_1, X_2, \cdots, X_L) = -\frac{1}{L}\sum_{i=1}^{L}\log p(X_i) \\
\xrightarrow{a.s.} -E(\log p(X)) \\
= H(X)
\end{aligned}
$$

- **典型集A_ε**
  - **定义**

$$
\begin{aligned}
A_\varepsilon &= \left\{\vec{x} : \left|\frac{I(\vec{x})}{L} - H(X)\right| \le \varepsilon\right\} \\
A_\varepsilon^c &= \left\{\vec{x} : \left|\frac{I(\vec{x})}{L} - H(X)\right| > \varepsilon\right\} \\
A_\varepsilon + A_\varepsilon^c &= X^L = \{\vec{x}_i\} \quad i = 1, 2, \cdots, n^L
\end{aligned}
$$

  - **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：

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

    b. 当 L 充分大时，典型集的概率近似为1：

$$
P_r(A_\varepsilon) = \sum_{\vec{x}\in A_\varepsilon} p(\vec{x}) > 1 - \varepsilon
$$

    c. 典型集中元素的个数：

$$
|A_\varepsilon| \le 2^{L(H(X)+\varepsilon)}
$$

      - 证明：

$$
\begin{aligned}
1 = \sum_{\vec{x}\in X^L} p(\vec{x}) &\ge \sum_{\vec{x}\in A_\varepsilon} p(\vec{x}) \\
&\ge \sum_{\vec{x}\in A_\varepsilon} 2^{-L(H(X)+\varepsilon)} = |A_\varepsilon| 2^{-L(H(X)+\varepsilon)} \\
\Rightarrow & |A_\varepsilon| \le 2^{L(H(X)+\varepsilon)}
\end{aligned}
$$

    d. 典型集中元素的个数：

$$
|A_\varepsilon| \ge (1-\varepsilon) 2^{L(H(X)-\varepsilon)}
$$

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

- 证明：

$$\begin{aligned}1-\varepsilon \le P_r(A_\varepsilon) &= \sum_{\vec{x}\in A_\varepsilon} p(\vec{x}) \\ &\le \sum_{\vec{x}\in A_\varepsilon} 2^{-L(H(X)-\varepsilon)} = |A_\varepsilon| 2^{-L(H(X)-\varepsilon)} \\ \Rightarrow |A_\varepsilon| \ge (1-\varepsilon)2^{L(H(X)-\varepsilon)}\end{aligned}$$

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

$$\frac{K_L}{L}\log m \ge H_L(\vec{X}) + \varepsilon$$

则当 L 足够大时，必可使译码差错小于 δ；
反之，当

$$\frac{K_L}{L}\log m \le H_L(\vec{X}) - 2\varepsilon$$

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

- **证明**
  - 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

$$m^{K_L} \ge |A_\varepsilon|$$

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

$$\begin{aligned}m^{K_L} &\ge 2^{L(H(X)+\varepsilon)} \\ K_L \log m &\ge L(H(X)+\varepsilon) \\ \frac{K_L}{L}\log m &\ge H(X)+\varepsilon\end{aligned}$$

此时，错误译码的概率为

$$P_e = P_r\{A_\varepsilon^c\} \le \frac{D(I(X))}{L\varepsilon^2}$$

若需 $P_e \le \delta$

$$\frac{D(I(X))}{L\varepsilon^2} \le \delta$$
$$L \ge \frac{D(I(X))}{\varepsilon^2 \delta}$$

- 反之，若

$$\begin{aligned}
&\frac{K_L}{L}\log m \le H(X) - 2\varepsilon \\
&\log m^{K_L} \le L(H(X) - 2\varepsilon) \\
&m^{K_L} \le 2^{L(H(X)-2\varepsilon)} < (1-\varepsilon)2^{L(H(X)-\varepsilon)} = |A_\varepsilon|_{min}
\end{aligned}$$

即 $m^{K_L} < |A_\varepsilon|$，码字个数少于$A_\varepsilon$ 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，

$$\begin{aligned}
P_r(A_\varepsilon\text{中}m^{K_L}\text{个元素}) &\le m^{K_L} \cdot \max_{\vec{x}\in A_\varepsilon} p(\vec{x}) = m^{K_L} 2^{-L(H(X)-\varepsilon)} \\
&\le 2^{L(H(X)-2\varepsilon)-L(H(X)-\varepsilon)} \\
&= 2^{-L\varepsilon}
\end{aligned}$$

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

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：

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

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**：
  - **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。

-

$$\begin{array}{c} \begin{bmatrix} X \\ P \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \\ p(a_1) & p(a_2) & \cdots & p(a_n) \end{bmatrix} \\ \text{码长} \quad l_1 \quad l_2 \quad \cdots \quad l_n \end{array}$$

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

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

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

根据Jensen不等式 $E[f(x)] \le f(E(x))$，有

$$\begin{aligned} H(X) - (\log m)\overline{k} &= \sum_{i=1}^{n} p(a_i) \log \frac{m^{-l_i}}{p(a_i)} \\ &\le \log \sum_{i=1}^{n} p(a_i) \frac{m^{-l_i}}{p(a_i)} = \log \sum_{i=1}^{n} m^{-l_i} \end{aligned}$$

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

$$\begin{aligned} &H(X) - (\log m)\overline{k} \le \log 1 = 0 \\ &\Rightarrow \overline{k} \ge \frac{H(X)}{\log m} \end{aligned}$$

若要等号成立

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

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

符号 a_i 的码长 l_i 需要是整数，因此取

$$\frac{-\log p(a_i)}{\log m} \le l_i < \frac{-\log p(a_i)}{\log m} + 1$$

由

$$
\begin{aligned}
&\frac{-\log p(a_i)}{\log m} \leq l_i \\
&\Rightarrow -\log p(a_i) \leq \log m^{l_i} \\
&\Rightarrow p(a_i) \geq m^{-l_i} \\
&\Rightarrow \sum_{i=1}^{n} m^{-l_i} \leq \sum_{i=1}^{n} p(a_i) = 1
\end{aligned}
$$

表明 l_i 满足唯一可译码存在的条件。

由

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

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：

$$
H_L(\vec{X}) \leq \overline{K} < H_L(\vec{X}) + \varepsilon
$$

  其中 ε 为任意小正数，$\overline{K} = \frac{\overline{k}_L}{L} \log m$，$\overline{k}_L$ 为 L 个符号的平均码长。
- **证明**
  将 $\vec{X} = (X_1 X_2 \cdots X_L)$ 作为一个整体，构成一个新的单符号信源。因此有：

$$
\begin{aligned}
\frac{LH_L(\vec{X})}{\log m} \leq \overline{k}_L &< \frac{LH_L(\vec{X})}{\log m} + 1 \\
H_L(\vec{X}) \leq \frac{\overline{k}_L}{L} \log m &< H_L(\vec{X}) + \frac{\log m}{L}
\end{aligned}
$$

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

$$
H_L(\vec{X}) \leq \overline{K} < H_L(\vec{X}) + \varepsilon \quad (L \to \infty)
$$

此时编码效率

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