信息论第三单元复习答案：信源编码

基本概念

1. **答案**：随机变量 X 的信源编码 C 是从 X 的取值空间 $\mathcal{X}$ 到字母表 D 上字符串集合的映射：

$$C : \mathcal{X} \to D^*$$

其中 D^* 表示 D 上所有有限长度字符串的集合。

2. **答案**：期望长度定义为：

$$L(C) = \mathbb{E}[l(C(X))] = \sum_{x \in \mathcal{X}} p(x) l(c(x))$$

其中 $l(c(x))$ 是码字 $c(x)$ 的长度。

答案：

3. (a) **奇异码**：不同符号可能映射到相同码字（$\exists x \neq x'$ 但 $c(x) = c(x')$）

   (b) **非奇异码**：不同符号映射到不同码字（$x \neq x' \Rightarrow c(x) \neq c(x')$）

   (c) **唯一可译码**：任意符号序列的编码结果唯一可解码（扩展编码也非奇异）

   (d) **前缀码**：无任何码字是另一个码字的前缀（$c(x)$ 不是 $c(x')$ 的前缀 $\forall x \neq x'$）

Kraft 不等式与 McMillan 定理

4. **答案：**对于字母表 D（$|D| = d$）上的前缀码，码长 $\{l_1, \ldots, l_n\}$ 满足：

$$\sum_{i=1}^{n} d^{-l_i} \leq 1$$

5. **证明：**考虑码树结构：

   - 设最大码长为 $l_{\max}$
   - 每个码字对应叶子节点，深度为 l_i
   - 每个深度 l_i 的节点有 $d^{l_{\max}-l_i}$ 个子孙节点
   - 所有码字的子孙节点互斥且总数不超过 $d^{l_{\max}}$

$$\sum_{i=1}^{n} d^{l_{\max}-l_i} \leq d^{l_{\max}} \Rightarrow \sum d^{-l_i} \leq 1 \quad \square$$

6. **证明**：构造性证明（按码长递增分配码字）：

- 设 $l_1 \leq l_2 \leq \cdots \leq l_n$
- 选择码字 c_i 满足：长度 l_i 且不是任何已分配码字的前缀
- 可用码字数：$d^{l_i} - \sum_{j=1}^{i-1} d^{l_i - l_j}$
- 由 Kraft 不等式：$d^{l_i} - \sum_{j=1}^{i-1} d^{l_i - l_j} \geq d^{l_i}(1 - \sum_{j=1}^{i-1} d^{-l_j}) > 0 \quad \square$

7. **答案**：对无限字母表 D，前缀码存在当且仅当：

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

构造方法：将码字映射到 $[0,1)$ 区间，码字 c_i 对应区间：

$$\left[\sum_{j=1}^{i-1} |D|^{-l_j}, \sum_{j=1}^{i} |D|^{-l_j}\right)$$

8.

定理：唯一可译码的码长必满足 Kraft 不等式。

证明：考虑扩展编码 C^k：

$$\begin{aligned}\left(\sum_{i=1}^{n} d^{-l_i}\right)^k &= \sum_{m=kl_{\min}}^{kl_{\max}} N_m d^{-m} \\ &\le \sum_{m=kl_{\min}}^{kl_{\max}} d^m d^{-m} \\ &= kl_{\max} - kl_{\min} + 1\end{aligned}$$

其中 N_m 是长度为 m 的码字数。由唯一可译性 $N_m \le d^m$，故：

$$\left(\sum d^{-l_i}\right)^k \le k(l_{\max} - l_{\min}) + 1$$

当 $k \to \infty$ 时，左边指数增长，右边线性增长，故 $\sum d^{-l_i} \le 1$。□

最优码长界与香农码

9.

证明：由相对熵非负性：

$$\begin{aligned}D(p \,\|\, q) &= \sum p_i \log_d \frac{p_i}{q_i} \ge 0 \\ \text{令} q_i &= \frac{d^{-l_i}}{\sum_j d^{-l_j}} \\ \Rightarrow \sum p_i \log_d p_i &\ge \sum p_i \log_d q_i \\ &= -\sum p_i l_i - \log_d\left(\sum d^{-l_j}\right) \\ &\ge -\sum p_i l_i \quad \text{(由 Kraft 不等式)} \\ \Rightarrow H_d(X) &\le L \quad \square\end{aligned}$$

10. **证明**：当 $p_i = d^{-l_i}$ 时：

$$\begin{aligned} L &= \sum p_i l_i = \sum p_i \log_d \frac{1}{p_i} \\ &= H_d(X) \end{aligned}$$

且此时 Kraft 不等式取等号：$\sum d^{-l_i} = \sum p_i = 1$。□

11. **定义**：香农码取 $l_i = \lceil \log_d \frac{1}{p_i} \rceil$

证明：

- 下界：$L \geq H_d(X)$（已证）
- 上界：

$$\begin{aligned} l_i &< \log_d \frac{1}{p_i} + 1 \\ \Rightarrow L &< \sum p_i \left(\log_d \frac{1}{p_i} + 1 \right) \\ &= H_d(X) + 1 \end{aligned}$$

- 验证 Kraft 不等式：

$$\sum d^{-l_i} \leq \sum d^{-\log_d(1/p_i)} = \sum p_i = 1 \quad \square$$

哈夫曼编码

算法：

12. (a) 将符号按概率递减排序

    (b) 合并概率最小的两个符号，赋予 0/1 标签

    (c) 将合并后的节点视为新符号（概率为和）

    (d) 重复直到只剩一个节点

    (e) 从根回溯得到码字

证明：

13. (a) 若存在 $p_j > p_k$ 但 $l_j > l_k$，交换码字可减小 L，矛盾

    (b) 若最长码字唯一，可缩短其长度而不破坏前缀性

    (c) 最长码字对应概率最小的两个符号，在最后一步合并

证明：（归纳法）

14.
- **基础：** $n = 2$ 时显然最优
- **归纳：**假设对 $n-1$ 个符号最优
- 设 x_1, x_2 是概率最小的两个符号，合并为 y
- 设 C' 是 $S' = (S \setminus \{x_1, x_2\}) \cup \{y\}$ 的最优码
- 对 S 构造码：$C(x_i) = C'(y) \cup \{i\}$，$i = 1, 2$
- 期望长度关系：$L(C) = L(C') + p(x_1) + p(x_2)$
- 若存在更优码 C^*，通过交换可构造 S' 的更优码，矛盾 $\square$

比较：

15.

- **香农码**：构造简单，$L < H_d(X) + 1$，但实际冗余常大于 1
- **哈夫曼码**：最优前缀码，L 最小，但构造复杂度 $O(n \log n)$
- 对均匀分布：两者性能接近
- 对非均匀分布：哈夫曼码显著优于香农码