信息论第五讲作业解答

中国科学技术大学《信息论 A》006125.01 班助教组

2025 年 4 月 29 日

第 1 题 对于在失真度量 $d(s, \hat{s})$ 下的离散无记忆信源 S，给定参数 $a \geq 0$，定义一个新的失真度量 $\tilde{d}(s, \hat{s})$：若 $d(s, \hat{s}) > a$，则 $\tilde{d}(s, \hat{s}) = 1$，否则 $\tilde{d}(s, \hat{s}) = 0$。描述在 $\tilde{d}$ 下，当 $D = 0$ 时 S 的率失真函数。

For a DMS S under distortion measure $d(s, \hat{s})$, define a new distortion measure as $\tilde{d}(s, \hat{s}) = 1$ if $d(s, \hat{s}) > a$ and 0 otherwise, given some parameter $a \geq 0$. Describe the rate-distortion function of S under $\tilde{d}$ at $D = 0$.

解: 依题意，d 和 $\tilde{d}$ 是 $\mathcal{S} \times \hat{\mathcal{S}}$ 上的失真度量,

$$\tilde{d}(s, \hat{s}) = \begin{cases} 0 & \text{if } d(s, \hat{s}) \leq a \\ 1 & \text{if } d(s, \hat{s}) > a \end{cases},$$

用 $\tilde{R}(D)$ 表示 S 在 $\tilde{d}$ 下的率失真函数, 即 $\tilde{R}(D) = \min_{\mathbf{E}[\tilde{d}(S,\hat{S})] \leq D} I(S, \hat{S})$. 当 $D = 0$ 时, 有

$$\mathbf{E}[\tilde{d}(S, \hat{S})] = P[d(S, \hat{S}) > a] \leq 0,$$

此时率失真函数为

$$\tilde{R}(0) = \min_{P_{\hat{S}|S}: P[d(S,\hat{S}) \leq a] = 1} I(S; \hat{S}). \tag{1}$$

□

讲义 Theorem 4.1 的成立基于讲义 (4.5) 式定义的 $d(\underline{s}, \underline{\hat{s}}) = \frac{1}{n} \sum_{i=1}^{n} d(s_i, \hat{s}_i)$. 现在我们将关于序列的失真定义为该序列中每对符号失真的最大值 (Peak distortion measures)，即

$$d(\underline{s}, \underline{\hat{s}}) = \max_{i \in \{1,\ldots,n\}} d(s_i, \hat{s}_i), \tag{2}$$

其余定义保证不变. 那么该如何得到此定义下的率失真函数呢？我们直接给出如下结论:

分别用 $R^P(\Delta)$ 和 $R_I^P(\Delta)$ 表示此时的率失真函数和信息率失真函数, 则有 [1, pp.117]

$$R^P(\Delta) = R_I^P(\Delta) = \tilde{R}(0) = \min_{P_{\hat{S}|S}: P[d(S,\hat{S}) \leq \Delta] = 1} I(S; \hat{S}), \tag{3}$$

其中 $\tilde{R}(0)$ 表示在 $\tilde{d}$ 下, $D=0$ 的率失真函数, 并且 $\tilde{d}(s,\hat{s})=0$ if $d(s,\hat{s})\leq\Delta$, and 1 otherwise. 也就是说 $R_I^P(\Delta)$ 其实就是 (1) 式 $a=\Delta$ 时的 $\tilde{R}(0)$. 如此的联系，基于下列分析，对 $\forall(\underline{s},\underline{\hat{s}})$

$$d(\underline{s},\underline{\hat{s}})\leq\Delta\Longleftrightarrow\Delta\geq\max_{i\in\{1,\dots,n\}}d(s_i,\hat{s}_i)\Longleftrightarrow d(s_i,\hat{s}_i)\leq\Delta,\forall i\Longleftrightarrow\tilde{d}(s_i,\hat{s}_i)=0,\forall i,$$

即此时的 $d(\underline{s},\underline{\hat{s}})\leq\Delta$ 等价于 $\frac{1}{n}\sum_{i=1}^n\tilde{d}(s_i,\hat{s}_i)=0$.

第 2 题 在第二节中，我们研究了索引字符串为二进制时的精确无损压缩。若索引字符串为 \$q\$ 进制（\$q \geq 2\$ ），通过推广第二节中的分析，推导期望索引字符串长度的上下界。

In Section II, we have studied exactly lossless compression when the index string is binary. If the index string is q-ary, $q\geq 2$, derive lower and upper bounds on the expected index string length, by generalizing the analysis in Section II.

解: 本题研究 q 进制下严格无损压缩期望码长 $\bar{\ell}=\sum_{s\in\mathcal{S}}P_S(s)\ell(s)$ 的上下界. **若不加说明，则以下不等号的成立原因均与讲义 Section II 中的二进制情况分析相同**.

首先对信源符号 $\mathcal{S}=\{a_1,a_2,\dots\}$ 进行重排, 使概率大小关系满足 $P_S(a_1)\geq P_S(a_2)\geq\dots$, 我们将每个符号编码为一个 q 进制码字 $W=f(S)\in\mathscr{W}^*$, 码本为

$$\mathscr{W}^*=\{\emptyset,0,1,\cdots,q-1,00,01,\cdots,0(q-1),10,11,\dots\}.$$

将更短的码字分配给发生概率更高的信源符号, 则码长满足 $\ell(a_i)\overset{(a)}{=}\lfloor\log_q(q-1)i\rfloor$, 见注 1.

首先考虑 $\bar{\ell}$ 的上界, 由于 $P_S(a_i)\leq 1/i$, 因此有

$$\ell(a_i)=\lfloor\log_q(q-1)i\rfloor\leq\log_q(q-1)i\leq\log_q(q-1)-\log_q P_S(a_i),$$

与

$$\begin{aligned}\bar{\ell}&=\sum_{s\in\mathcal{S}}P_S(s)\ell(s)\\&\leq\log_q(q-1)-\sum_{s\in\mathcal{S}}P_S(s)\log_q P_S(s)\\&=\log_q(q-1)-\sum_{s\in\mathcal{S}}P_S(s)\frac{\log_2 P_S(s)}{\log_2 q}\\&=\log_q(q-1)+\frac{1}{\log_2 q}H(S).\end{aligned}\tag{4}$$

接下来分析 $\bar{\ell}$ 的下界. 由于以下关系

$$\begin{aligned}H(S)&=H(S)+H(\ell(S)|S)\\&=H(S,\ell(S))\\&=H(S|\ell(S))+H(\ell(S)),\end{aligned}$$

分别分析 $H(S|\ell(S))$ 与 $H(\ell(S))$ 两项. 对于 $H(S|\ell(S))$, 满足

$$\begin{aligned} H(S|\ell(S)) &= \textstyle\sum_{\ell=0}^{\infty} P(\ell(S)=\ell) H(S|\ell(S)=\ell) \\ &\leq \textstyle\sum_{\ell=0}^{\infty} P(\ell(S)=\ell) \log_2 q^{\ell} \\ &= \mathbf{E}[\ell(S)] \log_2 q \\ &= \bar{\ell} \log_2 q. \end{aligned}$$

对于 $H(\ell(S))$, 满足

$$\begin{aligned} H(\ell(S)) &\leq (\bar{\ell}+1)\log_2(\bar{\ell}+1) - \bar{\ell}\log_2 \bar{\ell} \\ &< \log_2[e(\bar{\ell}+1)] \\ &\leq \log_2[e(\tfrac{1}{\log_2 q} H(S) + \log_q(q-1)+1)]. \end{aligned}$$

由此可得

$$\begin{aligned} &H(S) = H(S|\ell(S)) + H(\ell(S)) \\ &\qquad < \bar{\ell}\log_2 q + \log_2[e(\frac{1}{\log_2 q} H(S) + \log_q(q-1)+1)]. \end{aligned}$$

即 $\bar{\ell}$ 满足下界

$$\bar{\ell} > \frac{1}{\log_2 q} H(S) - \log_q[e(\frac{1}{\log_2 q} H(S) + \log_q(q-1)+1)]. \tag{5}$$

综上 (4)(5) 式, 码长的期望值满足上下界

$$\frac{1}{\log_2 q} H(S) - \log_q[e(\frac{1}{\log_2 q} H(S) + \log_q(q-1)+1)] < \bar{\ell} \leq \frac{1}{\log_2 q} H(S) + \log_q(q-1).$$

注 1. 为何码长满足 $\ell(a_i) = \lfloor \log_q(q-1)i \rfloor$: 根据变长编码的定义, 当从 k 位码长增加到 $k+1$ 位时, 可表示的码字数量增加 q^{k+1}. 因此, 码长 l 的变长编码可以容纳的码字数量为 $\sum_{k=0}^{l} q^k = \frac{1-q^{l+1}}{1-q}$. 因此, 给定第 i 个码字, 所需的位数为方程 $\frac{1-q^{l+1}}{1-q} = i$ 的解并上取整, 即 $\ell(a_i) = \lceil -1 + \log_q((q-1)i+1) \rceil = \lceil \log_q((q-1)i+1) \rceil - 1 \overset{(a)}{=} \lfloor \log_q((q-1)i) \rfloor$. 对于等式 (a) 而言, 由于 q 与 i 均为整数, 因此不存在整数 n 使得 $q^n \in ((q-1)i, (q-1)i+1)$, 因此 $\log_q((q-1)i)$ 与 $\log_q((q-1)i+1)$ 必然在两个连续的整数之间, 故等式 (a) 成立.

第 3 题

对于第二节中研究的精确无损压缩码，当 S 满足以下情况时，通过数值计算 $\bar{l}$ ：(a) 在 $\{1, 2, \ldots, M\}$ 上服从均匀分布；(b) 服从参数为 ϵ 的几何分布 。将这些情况下 $\bar{l}$ 的精确值与第二节中得到的上下界进行比较。

For the exactly lossless compression code studied in Section II, numerically evaluate $\bar{\ell}$ when S is (a) uniform over $\{1, 2, \ldots, M\}$, and (b) geometric with parameter ϵ. Compare the exact values of $\bar{\ell}$ under these cases with the upper and lower bounds obtained in Section II.

解: 对于均匀分布的情况, 第 i 个码的码长为 $\lfloor \log_2(i) \rfloor$. 我们令 $M = 2^k + s,\ k \in \mathbf{N}, 0 \leq s < 2^k$, 我们也可以得到 $k = \lfloor \log_2(M) \rfloor, s = M - 2^{\lfloor \log_2(M) \rfloor}$, 此时：

$$\begin{aligned}
\bar{\ell} &= \sum_{i=1}^{M} P_S(i)\ell(i) \\
&= \frac{1}{M}\left[0 \times 1 + 1 \times 2 + ... + (k-1) \times 2^{k-1} + k \times (s+1)\right] \\
&= \frac{1}{M}\left[(s+1)k + (k-2)2^k + 2\right] \\
&= \frac{1}{M}\left[Mk + k - 2^{k+1} + 2\right] \\
&= \lfloor \log_2(M) \rfloor + \frac{\lfloor \log_2(M) \rfloor + 2 - 2^{\lfloor \log_2(M) \rfloor + 1}}{M}.
\end{aligned} \tag{6}$$

对于几何分布的情况, 同样有第 i 个码的码长为 $\lfloor \log_2(i) \rfloor$, 我们计算其平均码长如下：

$$\begin{aligned}
\bar{\ell} &= \sum_{i=1}^{\infty} \epsilon(1-\epsilon)^{i-1} \lfloor \log_2(i) \rfloor \\
&= \sum_{k=1}^{\infty} k \sum_{i=2^k}^{2^{k+1}-1} \epsilon(1-\epsilon)^{i-1} \\
&= \sum_{k=1}^{\infty} k[(1-\epsilon)^{2^k-1} - (1-\epsilon)^{2^{k+1}-1}] \\
&= \sum_{k=1}^{\infty} (1-\epsilon)^{2^k-1}.
\end{aligned} \tag{7}$$

根据 Section II, 我们对这种编码方式有上下界估计：

$$H(S) - \log_2[e(H(S)+1)] < \bar{\ell} \leq H(S).$$

那么对于均匀分布和几何分布我们分别有 $H_{uniform}(S) = \log_2(M)$ 和 $H_{geometric}(S) = \frac{h_2(\epsilon)}{\epsilon}$, 所以有如下关系：

$$\log_2(M) - \log_2[e(\log_2(M)+1)] < \lfloor \log_2(M) \rfloor + \frac{\lfloor \log_2(M) \rfloor + 2 - 2^{\lfloor \log_2(M) \rfloor + 1}}{M} \leq \log_2(M).$$

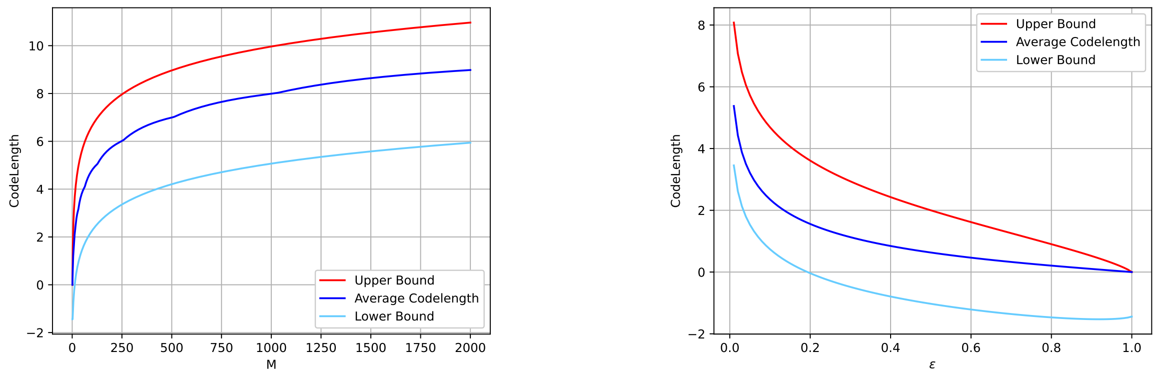
$$\frac{h_2(\epsilon)}{\epsilon} - \log_2[e(\frac{h_2(\epsilon)}{\epsilon} + 1)] < \sum_{k=1}^{\infty} (1-\epsilon)^{2^k-1} \leq \frac{h_2(\epsilon)}{\epsilon}.$$

上下界与真实的平均长度在图 1 中呈现. □

第 4 题

证明：一个码是唯一可译码，当且仅当对于任意整数 $n \geq 1$ ，以及任意 $\underline{s} \neq \underline{s}' \in \mathcal{S}^n$ ，有 $f(\underline{s}) \neq f(\underline{s}')$ 。

Prove that a code is uniquely decodable if and only if for any integer $n \geq 1$, and any $\underline{s} \neq \underline{s}' \in \mathcal{S}^n$, $f(\underline{s}) \neq f(\underline{s}')$.



(a) (b)

图 1: (a)：均匀分布平均码长及其上下界与 M 的关系. (b)：几何分布平均码长及其上下界与参数 ϵ 的关系.

证明:

必要性：如果 f 是惟一可译码, n 是正整数, $\underline{s}, \underline{s}' \in \mathcal{S}^n$, $\underline{s} \neq \underline{s}'$, 则 $f(\underline{s}) \neq f(\underline{s}')$.

充分性：对所有正整数 n 和 $\underline{s}, \underline{s}' \in \mathcal{S}^n$ 有 $f(\underline{s}) \neq f(\underline{s}')$. 对所有 $y_1, y_2, \cdots, y_m, z_1, z_2, \cdots, z_n \in \mathcal{S}$, 因为

$$f(y_1)f(y_2)\cdots f(y_m)f(z_1)f(z_2)\cdots f(z_n) \neq f(z_1)f(z_2)\cdots f(z_n)f(y_1)f(y_2)\cdots f(y_m),$$

所以 $f(y_1)f(y_2)\cdots f(y_m) \neq f(z_1)f(z_2)\cdots f(z_n)$. 因此 f 是惟一可译的. □

第 5 题

其中 \(A_{\ell}\) 表示长度为 \(\ell\) 的索引字符串的数量，而不是（5.21）式。让我们使用这种形式的克拉夫特不等式来证明定理 5.1 的逆定理部分，即对于给定的满足克拉夫特不等式的集合 \(\{A_{\ell} : \ell = 1, 2, \ldots\}\)，我们可以构造一个相应的前缀码。从根节点开始，完成以下归纳证明：a) 证明从根节点出发，在深度为 1 处至少有 \(A_1\) 个叶子节点，以容纳 \(A_1\) 个长度为 1 的索引字符串。b) 假设我们已经容纳了所有长度从 1 到 \(\ell - 1\) 的索引字符串。证明在深度为 \(\ell\) 处至少有 \(A_{\ell}\) 个未使用的叶子节点，以容纳 \(A_{\ell}\) 个长度为 \(\ell\) 的索引字符串。

Instead of (5.21) in Theorem 5.1, we may rewrite the Kraft inequality as

$$\sum_{\ell=1}^{\infty} A_\ell q^{-\ell} \leq 1,$$

where A_ℓ denotes the number of index strings of length ℓ. Let us use this form of the Kraft inequality to prove the converse part of Theroem 5.1; that is, for a given set of $\{A_\ell : \ell = 1, 2, \ldots\}$ satisfying the Kraft inequality, we can construct a corresponding prefix-free code. Starting with the root node, complete the following induction:

a) *Prove that from the root node, there are at least A_1 leaves at depth 1 to accommodate the A_1 length-1 index strings.*
b) *Suppose that we have already accommodated all index strings from length 1 to length $\ell - 1$. Prove that there are at least A_ℓ unused leaves at depth ℓ to accommodate the A_ℓ length-ℓ index strings.*

 6

证明: a) 由 Kraft 不等式：

$$A_1 q^{-1} + \sum_{\ell=2}^{\infty} A_\ell q^{-\ell} \leq 1,$$

可得 $A_1 q^{-1} \leq 1$, 即 $A_1 \leq q$, 所以至少有 A_1 个深度为 1 的叶子节点, 可以容纳 A_1 个索引字符串。

b) 假设已经容纳了长度为 1 到长度为 $\ell - 1$ 的所有索引字符串，且深度为 $\ell - 1$ 的叶子节点有 $A_{\ell-1}$ 个。由于 prefix-free 码的特性，先前选定为码字的节点的后续子树将不存在，所以在深度为 ℓ 时可以长出的叶子节点数目为：

$$q^\ell - \sum_{i=1}^{\ell-1} A_i q^{\ell-i}.$$

由 Kraft 不等式可知：

$$\sum_{i=1}^{\ell} A_i q^{-i} = \sum_{i=1}^{\ell-1} A_i q^{-i} + A_\ell q^{-\ell} \leq 1,$$

$$\text{即} \sum_{i=1}^{\ell-1} A_i q^{\ell-i} + A_\ell \leq q^\ell.$$

所以可以得到：$q^\ell - \sum_{i=1}^{\ell-1} A_i q^{\ell-i} \geq A_\ell$, 即至少有 A_ℓ 个深度为 ℓ 的叶子节点可以容纳 A_ℓ 个长度为 ℓ 的索引字符串。 □

第 6 题

如果对于任意 $s \neq s' \in \mathcal{S}$ ，$f(s)$ 都不是 $f(s')$ 的后缀，那么这个码被称为后缀无关码；如果一个码既是前缀无关码又是后缀无关码，那么它被称为固定无关码 。对于有限字母表 $|\mathcal{S}| < \infty$ 的离散无记忆信源 S ，当 $\sum_{s \in \mathcal{S}} q^{-\ell(s)} \leq 1/2$ 时，找到一种方法来构造一个码长为 $\{\ell(s) : s \in \mathcal{S}\}$ 的 q 进制固定无关码。

A code is called suffix-free if for any $s \neq s' \in \mathcal{S}, f(s)$ is not a suffix of $f(s')$, and is called fix-free if it is both prefix-free and suffix-free. For a DMS S with $|\mathcal{S}| < \infty$, when $\sum_{s\in\mathcal{S}} q^{-\ell(s)} \leq 1/2$, find a method to construct a q-ary fix-free code with lengths $\{\ell(s) : s \in \mathcal{S}\}$.

解: 根据题意可知，无缀码是指一个码既是无前缀的，也是无后缀的。

因为 $\sum_{s\in\mathcal{S}} q^{-\ell(s)} \leq 1/2 < 1$，即满足 Kraft 不等式，可知存在无前缀码。不妨考虑 $\ell(a_1) \leq \ell(a_2) \leq \cdots \leq \ell(a_{|\mathcal{S}|}) = \ell_{max}$，以及一个所有叶子节点深度均为 ℓ_{max} 的 q-叉树，我们先按照无前缀码的构造方式来构造：首先对于 a_1，在深度 $\ell(a_1)$ 为其分配一个节点即码字，记为 $f(a_1)$，并且删掉其所有的子节点，使其变成一个叶子节点；然后对于 a_2，不同的是，我们要在深度 $\ell(a_2)$ 的节点中找一个后缀不包含 $f(a_1)$ 的节点，记为 $f(a_2)$，然后删掉其所有的子节点变成一个叶子节点；以此类推，直到最后一个 $a_{|\mathcal{S}|}$，我们要在深度为 $\ell(a_{|\mathcal{S}|})$ 中的节点中找一个后缀不包含 $\{f(a_1), f(a_2), \ldots, f(a_{|\mathcal{S}|})\}$ 的节点，记为 $f(a_{|\mathcal{S}|})$.

下面证明我们这种构造方式是可以构造出来一个无缀码的：

利用数学归纳法，

- 当 $\ell_{max}=1$ 时，$\sum_{s\in\mathcal{S}} q^{-\ell(s)}=|\mathcal{S}|q^{-1}\le 1/2$，则 $|\mathcal{S}|\le 1/2\cdot q$，故显然可按照上述构造方法构造出 $|\mathcal{S}|$ 个无缀码；
- 假设当 $\ell_{max}=1,2,\ldots,\ell^*-1$ 成立，
  即考虑任意一个 $\mathcal{S}$ 满足 $\max_{s\in\mathcal{S}}\{\ell(s)\}=\ell^*$，此时有

$$\sum_{s\in\mathcal{S},\ell(s)<\ell^*} q^{-\ell(s)}+\sum_{s\in\mathcal{S},\ell(s)=\ell^*} q^{-\ell(s)}\le 1/2,$$

  由于 $\sum_{s\in\mathcal{S},\ell(s)<\ell^*} q^{-\ell(s)}\le 1/2$，则存在 $s\in\mathcal{S},\ell(s)<\ell^*$ 这些节点的无缀码。
  然后，我们考虑上述节点对 ℓ^* 层的影响，也就是说，ℓ^* 层中有多少个节点是以上述节点为前缀或者后缀的。对于前缀：我们根据讲义知道，对于长度为 $\ell(s)<\ell^*$ 的节点，在 ℓ^* 层中有 $q^{\ell^*-\ell(s)}$ 个节点是以它为前缀；同理，不难发现，对于后缀，对于长度为 $\ell(s)<\ell^*$ 的节点，在 ℓ^* 层中也有 $q^{\ell^*-\ell(s)}$ 个节点是以它为后缀的。但这两个集合可能是交叉重复的，因此 ℓ^* 层中以上述节点为前缀或者后缀的节点数量不会超过：

$$2\sum_{s\in\mathcal{S},\ell(s)<\ell^*} q^{\ell^*-\ell(s)},$$

  则 ℓ^* 层中无缀的节点数量至少为：

$$\begin{aligned}
q^{\ell^*}-2\sum_{s\in\mathcal{S},\ell(s)<\ell^*} q^{\ell^*-\ell(s)} &= 2q^{\ell^*}(1/2-\sum_{s\in\mathcal{S},\ell(s)<\ell^*} q^{-\ell(s)})\\
&\ge 2q^{\ell^*}\sum_{s\in\mathcal{S},\ell(s)=\ell^*} q^{-\ell(s)}\\
&= 2\sum_{s\in\mathcal{S},\ell(s)=\ell^*}\mathbf{1},
\end{aligned}$$

  其中 $\mathbf{1}$ 表示长度为 ℓ^* 的码字数量。又因为需要容纳的节点数为 $\sum_{s\in\mathcal{S},\ell(s)=\ell^*}\mathbf{1}$，且同一层（码长相等）节点不会为前缀或后缀，故 ℓ^* 层有足够的节点满足无缀码条件。

证明：对于字母表大小 $|S| = \infty$ 的离散无记忆信源 S，前缀码仍然满足克拉夫特不等式；反之，对于任何满足克拉夫特不等式的索引字符串长度，都存在一个相应的前缀码。

第 7 题

Prove that for a DMS S with $|S|=\infty$, a prefix-free code still satisfies the Kraft inequality, and conversely, for any index string lengths satisfying the Kraft inequality there exists a corresponding prefix-free code.

证明: 考虑无穷个信源符号时, 不能预先假设 $\ell_{\max}$, 因此讲义中的证明方法不再适用. 本题参考 [2, Theorem 5.2.2] 中的方法进行证明.

8

首先证明 prefix-free 码的码长满足 Kraft 不等式, 即

$$\sum_{i=1}^{\infty} q^{-\ell_i} \leq 1.$$

不妨考虑 q 进制的码本, 第 i 个码字为 $y_1 y_2 \cdots y_{\ell_i}$. 令 $0.y_1 y_2 \cdots y_{\ell_i}$ 为 q 进制下的实数, 数值上等于

$$0.y_1 y_2 \cdots y_{\ell_i} = \sum_{j=1}^{\ell_i} y_j q^{-j}.$$

将该码字对应 $[0,1]$ 上的一个子区间

$$\left[\sum_{j=1}^{\ell_i} y_j q^{-j}, \sum_{j=1}^{\ell_i} y_j q^{-j} + \frac{1}{q^{\ell_i}}\right),$$

这个区间的长度为 $q^{-\ell_i}$, 并且包含那些所有以 $0.y_1 y_2 \cdots y_{\ell_i} \cdots$ 表示的实数.

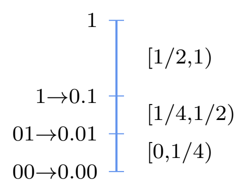


图 2: 一个例子：二进制下 $00, 01, 1$ 的构成的 prefix-free 码

由于 prefix-free 码中任意两个码字彼此互不为前缀, 因此任意两个码字对应的子区间不相交. 由于每个子区间长度 $q^{-\ell_i}$ 的总和不超过 1, Kraft 不等式得证.

接下来证明给定满足 Kraft 不等式的 $\ell_1, \ell_2, \ldots$, 可以构造出具有相应码长的 prefix-free 码. 依然沿用划分区间并与码字对应的思路, 先将码长重排列, 使之满足 $\ell_1 \leq \ell_2 \leq \ldots$, 然后从 $[0,1]$ 区间的左端开始, 逐步分配长度为 $q^{-\ell_i}$ 的不相交区间, Kraft 不等式的成立可以保证此分配能够完成, 由此可得 prefix-free 码的码字集合. $\square$

第 8 题 推导在 $\{1, 2, \ldots, 10000\}$ 上均匀分布的离散无记忆信源 S 的二进制霍夫曼码，并将得到的期望索引字符串长度与熵界 $\log_2{10000}$ 比特进行比较。

Derive the binary Huffman code for a DMS S uniformly distributed over $\{1, 2, \ldots, 10000\}$, and compare the resulting expected index string length with the entropy bound $\log_2 10000$ bits.

解: 可以 (见注 2) 假设有 x 个码字的长度为 ℓ, y 个码字长度为 $\ell + 1$, 由 Kraft 不等式得：

$$\begin{cases} x + y = 10000 \\ 2^{-\ell} x + 2^{-(\ell+1)} y = 1 \end{cases}.$$

由于 $2^{13}=8192<10000<16384=2^{14}$, 我们可以解出 $\ell=13, x=6384, y=3616$. 即：

$$\bar{\ell}=\frac{1}{10000}(6384\times 13+3616\times 14)\approx 13.3616>13.2877\approx \log_2(10000)=H(S) \text{ in bits.}$$

□

注 2. 引理: *对于均匀分布的信源, 在 Huffman 编码下不存在两个码字的码长之差大于 1.*

证明: 倘若存在 $\ell(s_i)-\ell(s_j)=m>1$, 我们考察 s_j 和具有 $\ell_{\max}$ 的 s_u, s_v, 我们将 s_u 和 s_j 均作为原先 s_j 的叶子节点, 此时 s_v 的编码长度也变为 $\ell_{\max}-1$, 那么我们操作后的编码方式和长度记为 ℓ', 便有：

$$\begin{aligned}
\bar{\ell} &= \sum_{s\in S} P_S(s)\ell(s)\\
&= \sum_{s\in S\setminus\{s_j,s_u,s_v\}} P_S(s)\ell(s)+P_S(s_j)\ell(s_j)+P_S(s_u)\ell(s_u)+P_S(s_v)\ell(s_v)\\
&= \sum_{s\in S\setminus\{s_j,s_u,s_v\}} P_S(s)\ell(s)+P_S(s_j)(\ell(s_j)+1)+P_S(s_u)(\ell(s_j)+1)+P_S(s_v)(\ell(s_v)-1)\\
&\quad -P_S(s_j)+P_S(s_u)(\ell(s_u)-\ell(s_j)-1)+P_S(s_v)\\
&= \sum_{s\in S} P_S(s)\ell'(s)+(-P_S(s_j)+P_S(s_u)(\ell(s_u)-\ell(s_j)-1)+P_S(s_v))\\
&= \sum_{s\in S} P_S(s)\ell'(s)+\frac{1}{n}(\ell(s_u)-\ell(s_j)-1)\\
&\geq \bar{\ell}'+\frac{1}{n}(\ell(s_i)-\ell(s_j)-1)>\bar{\ell}'.
\end{aligned}$$

从而我们这种构造方式可以将码长极差大于等于 *2* 的编码方式进行优化, 故最优 *Huffman* 码在均匀信源下将所有码字编为长度差至多为 *1* 的码字. □

第 9 题

对于离散无记忆信源 S，我们设计一个前缀码，使加权期望索引字符串长度 $\bar{l} = \sum_{s \in \mathcal{S}} P_S(s)c(s)\ell(s)$ 最小，其中 $c(s) > 0$ 是源消息 s 在每个索引位置的成本 。注意，当对于所有 $s \in \mathcal{S}$ ，$c(s) = 1$ 时，就回到了第四节研究的问题，该问题可通过那里的霍夫曼码解决。
a) 推导 $\bar{l}$ 的下界，并讨论何时能达到该下界。
b) 推广霍夫曼码，以得到使 $\bar{l}$ 最小的前缀码。

For a DMS S, we design a prefix-free code that minimizes the weighted expected index string length $\bar{\ell}=\sum_{s\in\mathcal{S}}P_S(s)c(s)\ell(s)$, where $c(s)>0$ is the cost per index position for source message s. Note that when $c(s)=1, \forall s\in\mathcal{S}$, we return to the problem studied in Section IV and it is solved by the Huffman code there.

a) *Derive a lower bound on $\bar{\ell}$, and discuss when this lower bound can be achieved.*
b) *Generalize the Huffman code to yield the prefix-free code that minimizes $\bar{\ell}$.*

a) 解: 设随机变量 Y 取值于 $\mathcal{S}$, 对每个 $s\in\mathcal{S}$ 有 $P_Y(s)=P_S(s)c(s)/\mathbf{E}[c(S)]$. 这样

$$\bar{\ell}=\mathbf{E}[c(S)]\sum_{s\in\mathcal{S}}P_Y(s)\ell(s)=\mathbf{E}[c(S)]\mathbf{E}[\ell(Y)]. \tag{8}$$

根据讲义第 IV 节, $\mathbf{E}[\ell(Y)]$ 大于等于 Y 以 q 为底的熵 $H_q(Y)$, 等号成立当且仅当对所有 $s \in \mathcal{S}$ 有 $\ell(s) = -\log_q(P_Y(s))$. 所以 $\bar{\ell} \geq \mathbf{E}[c(S)]H_q(Y)$, 等号成立当且仅当对所有 $s \in \mathcal{S}$ 有 $\ell(s) = -\log_q(P_Y(s))$. □

b) 解: 根据 (8) 式, 我们只需要用 Huffman 算法找到 Y 平均码长最小的 prefix-free 码. 这个码就是最小化 $\bar{\ell}$ 的 prefix-free 码. □

第 10 题

对于具有 K 个非零概率和一个零概率的离散无记忆信源 S，即 $P_S(a_1) \geq P_S(a_2) \geq \cdots \geq P_S(a_K) > P_S(a_{K + 1}) = 0$，我们既可以设计一个忽略零概率的霍夫曼码，也可以设计一个包含零概率的霍夫曼码 。找出这两种不同霍夫曼码的期望索引字符串长度之间的关系。

For a DMS S with K positive probabilities and one zero probability, i.e., $P_S(a_1) \geq P_S(a_2) \geq \ldots \geq P_S(a_K) > P_S(a_{K+1}) = 0$, we may either design a Huffman code omitting the zero probability, or including it. Find the relationship between the expected index string lengths of these two different Huffman codes.

解: 先考虑 $q = 2$ 的情况. 在不考虑 a_{K+1} 的情况下构造霍夫曼编码, 用 $\ell(s)$ 表示此时每个符号对应的码字长度, 则由构造规则可知 a_K 和 a_{K-1} 对应的码字为树的兄弟节点, 且对应的码长为 $\ell(a_K) = \ell(a_{K-1}) = \ell_{\max}$. 当考虑 a_{K+1} 后, 用 $\ell'(s)$ 表示此时每个符号对应的码字长度, 此时在霍夫曼编码对应的树中, a_K 和 a_{K+1} 对应的码字应为树的兄弟节点, 对应父节点的概率为 $P_S(a_K) + P_S(a_{K+1}) = P_S(a_K)$. 因此, 此时的树只是将第一种情况中的 a_K 节点扩充为 a_K 和 a_{K+1} 两片叶子，则有 $\ell'(a_K) = \ell'(a_{K+1}) = \ell_{\max} + 1$. 接下来分析平均码长，即

$$
\begin{aligned}
\bar{\ell} &= \sum_{s \in \{a_1,\ldots,a_K\}} P_S(s)\ell(s) \\
\bar{\ell}' &= \sum_{s \in \{a_1,\ldots,a_{K+1}\}} P_S(s)\ell'(s) \\
&= \sum_{s \in \{a_1,\ldots,a_{K-1}\}} P_S(s)\ell'(s) + P_S(a_K)\ell'(a_K) + P_S(a_{K+1})\ell'(a_{K+1}) \\
&= \sum_{s \in \{a_1,\ldots,a_{K-1}\}} P_S(s)\ell(s) + \left(P_S(a_K) + P_S(a_{K+1})\right)(\ell_{\max} + 1) \\
&= \sum_{s \in \{a_1,\ldots,a_{K-1}\}} P_S(s)\ell(s) + P_S(a_K)\ell_{\max} + P_S(a_K) \\
&= \bar{\ell} + P_S(a_K),
\end{aligned}
$$

其中 $\bar{\ell}$ 与 $\bar{\ell}'$ 分别为不考虑 $P_S(a_{K+1})$ 和考虑 $P_S(a_{K+1})$ 的霍夫曼平均码长. 因此考虑 a_{K+1} 后的霍夫曼码平均码长将增大 $P_S(a_K)$.

同理考虑 $q > 2$ 时的扩充情况. 按照讲义中 (5.35) 的计算方式, 如果 $r = 0$, 即 $q - 1$ 整除 $(K - q)(q - 2)$, 此时无未使用的叶节点, 扩充零概率节点会增加 $P_S(a_K)$ 的码长的期望值；如果 $r \neq 0$, 则无需扩充新的节点, 此时两种情况下码长的期望值相同. □

考虑具有（不一定相同的）有限字母表的独立离散无记忆信源 S_1 和 S_2 。分别将它们的二进制霍夫曼码记为 f_{S_1} 和 f_{S_2} 。现在将 (S_1, S_2) 视为一个单一的离散无记忆信源，并使用串联码 $[f_{S_1}, f_{S_2}]$ 作为 (S_1, S_2) 的码；例如，对于某些 (s_1, s_2) ，如果 $f_{S_1}(s_1) = 001$ 且 $f_{S_2}(s_2) = 101$ ，那么 $f_{S_1,S_2}(s_1, s_2) = 001101$ 。

第 11 题

a) 证明 f_{S_1,S_2} 是一个前缀码。
b) f_{S_1,S_2} 的克拉夫特不等式是否总是取等号？

Consider independent DMSs S_1 and S_2 with (not necessarily identical) finite alphabets. Denote their binary Huffman codes as f_{S_1} and f_{S_2}, respectively. Now view (S_1, S_2) as a single DMS, and use the concatenation $[f_{S_1}, f_{S_2}]$ as the code for (S_1, S_2); for example, if $f_{S_1}(s_1) = 001$ and $f_{S_2}(s_2) = 101$ for some (s_1, s_2), then $f_{S_1,S_2}(s_1, s_2) = 001101$.

a) Show that f_{S_1,S_2} is a prefix-free code.
b) Does the Kraft inequality for f_{S_1,S_2} always hold equal?

解: a) 对任意的 $(s_1, s_2) \neq (s_1', s_2')$, 假设存在码字 $f_{S_1,S_2}(s_1, s_2)$ 是 $f_{S_1,S_2}(s_1', s_2')$ 的前缀, 则 $f_{S_1,S_2}(s_1', s_2')$ 前面 $\ell(f_{S_1,S_2}(s_1, s_2))$ 位和 $f_{S_1,S_2}(s_1, s_2)$ 完全相同.

- 若 $\ell(f_{S_1}(s_1)) > \ell(f_{S_1}(s_1'))$, 则 $f_{S_1}(s_1')$ 是 $f_{S_1}(s_1)$ 的前缀, 与 f_{S_1} 是 prefix-free 码矛盾.
- 若 $\ell(f_{S_1}(s_1)) < \ell(f_{S_1}(s_1'))$, 则 $f_{S_1}(s_1)$ 是 $f_{S_1}(s_1')$ 的前缀, 与 f_{S_1} 是 prefix-free 码矛盾.
- 若 $\ell(f_{S_1}(s_1)) = \ell(f_{S_1}(s_1'))$, 由于前缀特性, 亦即 $f_{S_1}(s_1) = f_{S_1}(s_1')$. 但此时会有 $f_{S_2}(s_2)$ 是 $f_{S_2}(s_2')$ 的前缀, 也会与 f_{S_2} 是 prefix-free 码产生矛盾.

综上，所以 f_{S_1,S_2} 是 prefix-free 码.

b) 仍然满足 Kraft 不等式. 由于 S_1, S_2 是独立的, 因此：

$$\begin{aligned}
\sum_{s_1,s_2} 2^{-\ell(f_{S_1,S_2}(s_1,s_2))} &= \sum_{s_1,s_2} 2^{-\ell(f_{S_1}(s_1)+f_{S_2}(s_2))} \\
&= \sum_{s_1} 2^{-\ell(f_{S_1}(s_1))} * \sum_{s_2} 2^{-\ell(f_{S_2}(s_2))} \\
&= 1 \times 1 = 1.
\end{aligned}$$

香农 - 范诺码采用一种保守策略，将 $-\log_q P_S(s)$ 的所有非整数值向上取整 。或许可以明智地将 $-\log_q P_S(s)$ 的某些非整数值向下取整，从而得到一个期望索引字符串长度更短的前缀码。

第 12 题

a) 提出一种设计前缀码的算法，通过有选择地将 $-\log_q P_S(s)$ 的某些非整数值向下取整，使其性能可能优于香农 - 范诺码。
b) 找出一个例子，其中你设计的码明显比霍夫曼码差。

The Shannon-Fano code adopts a conservative philosophy by rounding up all non-integer values of $-\log_q P_S(s)$. It may be possible to judiciously round down some non-integer values of $-\log_q P_S(s)$, so as to obtain a prefix-free code with a smaller expected index string length.

a) Propose an algorithm for designing a prefix-free code that may outperforms the Shannon-Fano code, by selectively rounding down some non-integer values of $-\log_q P_S(s)$.
b) Find an example where your designed code is strictly worse than the Huffman code

a) 解: 记 $F = \{s \in \mathcal{S} | -\log_q(P_S(s))$ 不是整数$\}$. 任取 $s_F \in F$ 使

$$-\log_q\left(P_S\left(s_F\right)\right) = \max_{s \in F} -\log_q\left(P_S\left(s\right)\right). \tag{9}$$

我们为 $\mathcal{S}$ 中除 s_F 之外的每个符号 s 分配码长 $\lceil -\log_q\left(P_S\left(s\right)\right)\rceil$. 如果

$$q^{-\lfloor -\log_q(P_S(s))\rfloor} + \sum_{s \in \mathcal{S}, s \neq s_F} q^{-\lceil -\log_q(P_S(s))\rceil} \leq 1 \tag{10}$$

则为 s_F 分配码长 $\lfloor -\log_q\left(P_S\left(s\right)\right)\rfloor$. 否则为 s_F 分配码长 $\lceil -\log_q\left(P_S\left(s\right)\right)\rceil$.

按 (9) (10) 设计的 prefix-free 码的平均码长不会超过 Shannon-Fano 码的平均码长, 有时小于 Shannon-Fano 码的平均码长. 见下面两个例子:

① 设 S 服从 $\mathcal{S} = \{0, 1, 2\}$ 上的均匀分布, $q = 2$. 因为

$$1 < -\log_q(P_S(0)) = -\log_q(P_S(1)) = -\log_q(P_S(2)) < 2,$$

所以 $F = \{0, 1, 2\}$, 我们可以取 $s_F = 0$. 因为 $2^{-1} + 2 \times 2^{-2} = 1$, 所以此时 0, 1, 2 对应的码字长度为 1, 2, 2, 码字可以分别是 0, 10 和 11. 而 0, 1 和 2 的 Shannon-Fano 码字的长度都是 2, 所以这里设计的码的平均码长小于 Shannon-Fano 码的平均码长.

类似的算法还有很多. 有时我们不能简单地取码长为 $\lfloor -\log_q\left(P_S\left(s\right)\right)\rfloor$, 这时所有类似的算法都会失效.

②设 $q = 4$, $\mathcal{S} = \{0, 1, \cdots, 14\}$, $P_S(0) = 1/8$, 对所有正整数 $1 \leq s \leq 14$ 有 $P_S(s) = 1/16$. 这样

$$-\log_q\left(P_S\left(s\right)\right) = \begin{cases} \dfrac{3}{2}, & s = 0 \\ 2, & s \in \{1, 2, \cdots, 14\} \end{cases}.$$

由于 $4^{-1} + \sum_{s=1}^{14} 4^{-2} = 9/8 > 1$, 我们不能为 0 分配码长 $\lfloor -\log_q\left(P_S\left(s\right)\right)\rfloor = 1$, 所以此时平均码长等于 Shannon-Fano 码的平均码长. □

b) 解: 设 $q = 2$, $\mathcal{S} = \{0, 1\}$, $P_S(0) = 1/5$, $P_S(1) = 4/5$. 此时 $2 < -\log_q(P_S(0)) < 3$, $0 < -\log_q(P_S(1)) < 1$, $F = \{0, 1\}$, $s_F = 0$. 因为 $2^{-2} + 2^{-1} = 3/4 < 1$, 所以 0 的码长是 2, 1 的码长是 1. 由于 0 和 1 的 Huffman 码长都是 1, 这里设计的码的平均码长大于 Huffman 码的平均码长.

第 13 题

在第4讲有损信源表示的问题设定中，编码索引 $W \in \{1, 2, \cdots, M_n\}$ 也可看作是固定长度为 $\lceil \log_2 M_n \rceil$ 的二进制字符串 。现在，如果我们允许 W 为可变长度，从所有有限长度二进制字符串的集合 $\mathcal{W}^* = \{\emptyset, 0, 1, 00, 01, 10, 11, 000, \cdots\}$ 中选取。定义码率为 $R = \mathbb{E}[\ell(\underline{S})]/n$ ，其中 $\ell(\underline{S})$ 是对 $\underline{S}$ 进行编码的 W 的长度，n 是 $\underline{S}$ 的长度。修改第4讲中逆定理部分的证明，以表明可变长度编码仍无法超越率失真函数。

In the problem formulation of lossy source representation in Lecture 4, the encoded index $W \in \{1, 2, \cdots, M_n\}$ may also be viewed as a binary string of a fixed length $\lceil \log_2 M_n \rceil$. Now, if we allow W to be of variable length, drawn from the set of all finite-length binary strings

$\mathscr{W}^* = \{\emptyset, 0, 1, 00, 01, 10, 11, 000, \cdots\}$. *Define the rate of a code by* $R = \mathbf{E}[\ell(\underline{S})]/n$, *where* $\ell(\underline{S})$ *is the length of* W *encoding* $\underline{S}$ *and* n *is the length of* $\underline{S}$. *Modify the proof of the converse part in Lecture 4, to show that variable-length coding still cannot outperform the rate-distortion function.*

解:

证明逆定理，需要假设任意一对编译码器 $f_n^{(s)}, g_n^{(s)}$，满足失真约束 $\mathbf{E}[d(\underline{S}, \underline{\hat{S}})] \le D$，对于本题，$\ell(\underline{S}) = T(W) = \{0, 1, 2\cdots\}$, $nR = \mathbf{E}[\ell(\underline{S})] = \mathbf{E}[T(W)]$.

由于

$$\begin{aligned} H(W) &= H(W) + H(T(W)|W) \\ &= H(W, T(W)) \\ &= H(W|T(W)) + H(T(W)) \end{aligned} \tag{11}$$

其中，

$$\begin{aligned} H(W|T(W)) &= \sum_{t=0}^{\infty} H(W|T(W)=t)P(T(W)=t) \\ &\le \sum_{t=0}^{\infty} tP(T(W)=t) \\ &= \mathbf{E}[T(W)] \\ &= nR \end{aligned} \tag{12}$$

对于另一项 $H(T(W))$，我们知道在均值一定时，几何分布熵最大，由于 $T(W)$ 取值从 0 开始，我们对其进行平移操作，即：

$$\begin{aligned} H(T(W)) &= H(T(W)+1) \\ &\le (nR+1)\log_2(nR+1) - nR\log_2(nR) \\ &= nR\log_2(1+\frac{1}{nR}) + \log_2(nR+1) \\ &\le nR\cdot\frac{1}{nR}\log_2 e + \log_2(nR+1) \\ &= \log_2 e(nR+1) \end{aligned} \tag{13}$$

其中 $n \to \infty$ 时，最后一个不等号成立，结合 (11), (12), (13) 式，可得：

$$nR + \log_2 e(nR+1) \ge H(W) \tag{14}$$

然后采用与讲义 4.3 节中相同的步骤 ((4.31)-(4.36) 式) 可以得到:

$$\frac{1}{n}H(W) \ge R_I(D), \tag{15}$$

结合 (14) 和 (15) 式，所以在 $n \to \infty$ 时，得到 $R \ge R_I(D)$. □

参考文献

[1] I. Csiszár and J. Körner, *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011.

[2] T. M. Cover and J. A. Thomas, *Elements of information theory, 2nd ed.* John Wiley & Sons, 2006.