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

2024 年 5 月 28 日

第 1 题

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. Describe the rate-distortion function of S under $\tilde{d}$ at $D=0$.

解: 设 d 和 $\tilde{d}$ 是 $\mathcal{S}\times\hat{\mathcal{S}}$ 上的失真度量. 用 R 表示 S 在 $\tilde{d}$ 下的率失真函数. 如果随机变量 $\hat{S}$ 取值于 $\hat{\mathcal{S}}$ 则 $\mathbf{E}[\tilde{d}(S,\hat{S})]=P[d(S,\hat{S})>a]$. $R(0)$ 等于 $\mathbf{E}[\tilde{d}(S,\hat{S})]\le 0$ 的条件下 $I(S;\hat{S})$ 的最小值, 即 $P[d(S,\hat{S})\le a]=1$ 的条件下 $I(S;\hat{S})$ 的最小值. □

第 2 题

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

本题研究 D 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 $P_S(a_1)\ge P_S(a_2)\ge\dots$. 我们将每个符号编码为一个 D 进制码字 $W=f(S)$, 码本为

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

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 $\ell(a_i)=\lfloor\log_D(D-1)i\rfloor$.

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

$$\ell(a_i)=\lfloor\log_D(D-1)i\rfloor\le\log_D(D-1)i\le\log_D(D-1)-\log_D P_S(a_i),$$

与

$$\begin{aligned}\bar{\ell} &= \sum_{s\in\mathcal{S}} P_S(s)\ell(s)\\ &\le \log_D(D-1) - \sum_{s\in\mathcal{S}} P_S(s)\log_D P_S(s)\\ &= \log_D(D-1) - \sum_{s\in\mathcal{S}} P_S(s)\frac{\log_2 P_S(s)}{\log_2 D}\\ &= \log_D(D-1) + \frac{1}{\log_2 D}H(S).\end{aligned}$$

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

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

对于 $H(\ell(S))$, 满足

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

由此可得

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

即 $\bar{\ell}$ 满足下界

$$\bar{\ell} > \frac{1}{\log_2 D}H(S) - \log_D[e(\frac{1}{\log_2 D}H(S)+\log_D(D-1)+1)].$$

综上, 码长的期望值满足上下界

$$\frac{1}{\log_2 D}H(S) - \log_D[e(\frac{1}{\log_2 D}H(S)+\log_D(D-1)+1)] < \bar{\ell} \le \frac{1}{\log_2 D}H(S)+\log_D(D-1).$$

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

第 3 题

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

证明: 对于均匀分布的情况, 第 n 个码的码长为 $\lfloor \log_2(n) \rfloor$. 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：

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

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

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

根据 Section II, 我们对这种编码方式有上下界估计：

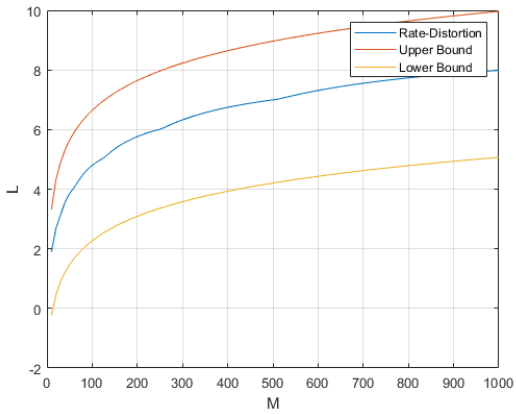
$$H(S) - \log_2[e(H(S)+1)] < \bar{\ell} \leq H(S).$$

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 $H_{geometric}(S) = \frac{h_2(\epsilon)}{\epsilon}$：

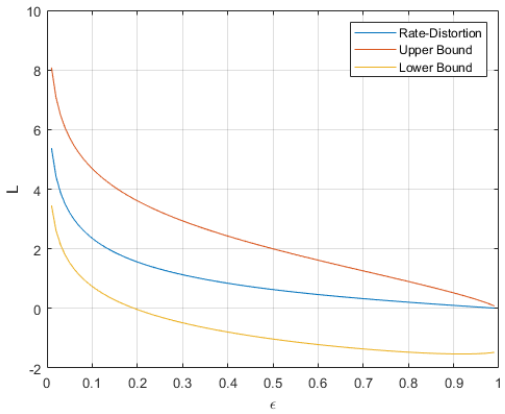
$$\log_2(M) - \log_2[e(\log_2(M)+1)] < \lfloor \log_2(M) \rfloor + \frac{\lfloor \log_2(M) \rfloor + 2 - 2^{\lfloor \log_2(M) \rfloor + 1}}{M} \leq \log_2(M).$$

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

上下界与真实的平均长度如图 3 所示. □



(a)



(b)

图 1: 左：均匀分布上界, 下界和平均码长与 M 的关系. 右：几何分布上界, 下界和平均码长与参数 ϵ 的关系.

第 4 题

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

证明: 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 $f(\underline{s}) \neq f(\underline{s}')$.

再假设对所有正整数 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

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

考虑可数个信源符号时, 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 Cover-Thomas Theorem 5.2.2 中的方法进行证明.

首先证明前缀码的码长满足 Kraft 不等式, 即

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

不妨考虑 D 进制的码本, 第 i 个码字的码长为 $y_1y_2\cdots y_{l_i}$. 令 $0.y_1y_2\cdots y_{l_i}$ 为 D 进制下的实数, 数值上等于

$$0.y_1y_2\cdots y_{l_i}=\sum_{j=1}^{l_i}y_jD^{-j}.$$

将该码字对应 $[0,1]$ 上的一个子区间

$$\left[0.y_1y_2\cdots y_{l_i},0.y_1y_2\cdots y_{l_i}+\frac{1}{D^{\ell_i}}.\right)$$

由于前缀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. 由于区间长度 $\frac{1}{D^{\ell_i}}$ 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 $l_1,l_2,\ldots$, 可以构造出具有相应码长的前缀码. 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 $l_1\le l_2\le\ldots$. 然后从 $[0,1]$ 区间的左端开始, 逐步分配长度为 $\frac{1}{D^{\ell_i}}$ 的区间, 由此可得前缀码的码字集合.

第 6 题

Describe 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: 我们先证明一个引理：对于全部均匀分布的信源, 在 Huffman 编码下不存在两个码字的编码长度之差大于 1.

证明：倘若存在 $l(s_i)-l(s_j)=m>1$, 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：

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

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一的码字.

那么我们假设有 x 个码字的长度为 ℓ, y 个码字的长度为为 $\ell+1$, 由 Kraft 不等式可以得到：

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

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

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

□

第 7 题

For a DMS S, we design a prefix-free code that minimizes the weighted expected codeword length $\bar{\ell}=\sum_{s\in\mathcal{S}}P_S(s)c(s)\ell(s)$, where $c(s)>0$ is the cost per codeword position when the source letter is 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.

a) Derive a lower bound on $\bar{\ell}$, and discuss when this lower bound can be achieved.
b) Generalize the Huffman algorithm 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

$$\bar{\ell}=\mathbf{E}[c(S)]\sum_{s\in\mathcal{S}}P_Y(s)\ell(s)=\mathbf{E}[c(S)]\mathbf{E}[\ell(Y)]. \tag{1}$$

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

b) 解: 根据 (1) 式, 我们只需要用 Huffman 算法找到 Y 平均码长最小的前缀码. 这个码就是最小化 $\bar{\ell}$ 的前缀码. □

第 8 题

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 last zero

probability, or including it. Find the relationship between the expected index string lengths of these two different Huffman codes.

先考虑 $D = 2$ 的情况. 在不考虑 a_{K+1} 的情况下构造霍夫曼编码, 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 $\ell_{\max}$. 当考虑 a_{K+1} 后, 此时在霍夫曼编码对应的树中, a_K 和 a_{K+1} 对应的码字为树的兄弟节点, 对应父节点的概率为 $P_S(a_K) + P_S(a_{K+1}) = P_S(a_K)$. 因此, 此时的树只是在将 a_K 节点扩充为 a_K 与 a_{K+1} 两个叶子, 即

$$\begin{aligned}
\bar{\ell} &= \sum_{s\in\mathcal{S}} P_S(s)\ell(s) \\
\bar{\ell}' &= \sum_{s\in\mathcal{S}} P_S(s)\ell'(s) \\
&= \sum_{s\in\mathcal{S}\setminus\{a_K\}} P_S(s)\ell'(s) + P_S(a_K)\ell'(a_K) + P_S(a_{K+1})\ell'(a_{K+1}) \\
&= \sum_{s\in\mathcal{S}\setminus\{a_K\}} P_S(s)\ell(s) + \left(P_S(a_{K+1}) + P_S(a_{K+1})\right)\left(\ell(a_K) + 1\right) \\
&= \sum_{s\in\mathcal{S}\setminus\{a_K\}} P_S(s)\ell(s) + P_S(a_K)\ell(a_K) + P_S(a_K) \\
&= \bar{\ell} + P_S(a_K),
\end{aligned}$$

其中 $\bar{\ell}$ 与 $\bar{\ell}'$ 分别为考虑 $P_S\left(a_{K+1}\right)$ 前后的霍夫曼平均码长. 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 $D > 2$ 时的扩充情况. 按照讲义中 (5.34) 的计算方式, 如果 $r = 0$, 即 $D - 1$ 整除 $(K - D)(D - 2)$, 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.

第 9 题

Consider independent DMSs S and T with finite alphabets. Denote their binary Huffman codes as f_S and f_T, respectively. Now view (S, T) as a single DMS, and use the concatenation $[f_S, f_T]$ as the code for (S, T); for example, if $f_S(s) = 001$ and $f_T(t) = 101$ for some (s, t), then $f_{ST}(s, t) = 001101$.

a) Show that f_{ST} is a prefix-free code.

b) Does the Kraft inequality for f_{ST} always hold equal?

解: a) 倘若存在码字 $f_{ST}(s,t)$ 是 $f_{ST}(s',t')$ 的前缀, 则 $f_{ST}(s',t')$ 前面 $\ell(f_{ST}(s,t))$ 位和 $f_{ST}(s,t)$ 完全相同.

若 $\ell(f_S(s)) > \ell(f_S(s'))$, 则 $f_S(s')$ 是 $f_S(s)$ 的前缀, 矛盾.

若 $\ell(f_S(s)) < \ell(f_S(s'))$, 则 $f_S(s)$ 是 $f_S(s')$ 的前缀, 矛盾.

若 $\ell(f_S(s)) = \ell(f_S(s'))$, 由于前缀特性, 亦即 $f_S(s) = f_S(s')$. 但此时 $f_T(t)$ 是 $f_T(t')$ 的前缀, 也会与 f_T 是非前缀码产生矛盾.

b) 仍然满足 Kraft 不等式. 由于 S, T 是独立的, 因此:

$$\begin{aligned}\sum_{s,t} 2^{-\ell(f_{ST}(s,t))} &= \sum_{s,t} 2^{-\ell(f_S(s)+f_T(t))} \\ &= \sum_{s} 2^{-\ell(f_S(s))} * \sum_{t} 2^{-\ell(f_T(t))} \\ &= 1 \times 1 = 1.\end{aligned}$$

□

第 10 题

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

a) Describe an algorithm for designing a prefix-free code that may outperforms the Shannon code, by selectively rounding down some non-integer values of $-\log_D P_S(s)$.

b) Find an example where your algorithm is outperformed by the Huffman algorithm.

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

$$-\log_D(P_S(s_F)) = \max_{s \in F} -\log_D(P_S(s)).$$

为 $\mathcal{S}$ 中除 s_F 之外的每个符号 s 分配码长 $\lceil -\log_D(P_S(s)) \rceil$. 如果

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

则为 s_F 分配码长 $\lfloor -\log_D(P_S(s_F)) \rfloor$. 否则为 s_F 分配码长 $\lceil -\log_D(P_S(s_F)) \rceil$.

这样设计的前缀码的平均码长不会超过 Shannon 码的平均码长, 有时小于 Shannon 码的平均码长. 设 S 服从 $\mathcal{S} = \{0, 1, 2\}$ 上的均匀分布, $D = 2$. 因为

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

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 0 的码长是 1, 1 和 2 的码长是 2. 0, 1 和 2 的码字可以分别是 0, 10 和 11. 由于 0, 1 和 2 Shannon 码字的长度都是 2, 这里设计的码的平均码长小于 Shannon 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 $\lfloor -\log_D(P_S(s))\rfloor$, 这时所有类似的算法都会失效. 设 $D=4$, $\mathcal{S}=\{0,1,\cdots,14\}$, $P_S(0)=1/8$, 对所有正整数 $s\le 14$ 有 $P_S(s)=1/16$. 这样

$$-\log_D(P_S(s))=\begin{cases}\frac{3}{2}, & s=0\\ 2, & s\in\{1,2,\cdots,14\}\end{cases}.$$

由于 $4^{-1}+\sum_{s=1}^{14}4^{-2}=9/8>1$, 我们不能为 0 分配码长 $\lfloor -\log_D(P_S(s))\rfloor=1$, 为 1, 2, $\cdots$, 14 分配码长 2. □

b) 解: 设 $D=2$, $\mathcal{S}=\{0,1\}$, $P_S(0)=1/5$, $P_S(1)=4/5$. 此时 $2<-\log_D(P_S(0))<3$, $0<-\log_D(P_X(1))<1$, $F=\{0,1\}$, $s_F=0$. 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 11 题

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 consider variable-length lossy source representation, by allowing W to be drawn from the set of all finite-length binary strings $\mathscr{C}^*=\{\emptyset,0,1,00,01,10,11,000,\cdots\}$. 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 $\mathbf{E}[d(\underline{S},\hat{\underline{S}})]\le D$，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

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

其中，

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

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：

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

故 $H(W) \leq nR + \log_2 e(nR+1)$，然后采用与讲义 4.3 中相同的步骤可以得到:

$$R + \frac{\log_2 e(nR+1)}{n} \geq R(D),$$

在 $n \to \infty$ 时，得到 $R \geq R(D)$. □