《2022 年本科生暑期数学前沿课程》作业

姓名：魏嵘　　学号：PB19151785　　教师姓名：孙雯

1 Background

1.1 Entropy and Information Theory

在物理中，熵往往用于衡量一个系统的“混乱程度”–越“混乱”，则熵越大；而“混乱”实则代表着一种不确定性和未知性。

对于随机事件来说，小概率事件发生、大偏差的出现则代表着”未知的、不确定的”事件发生，其相比于常规事件显然蕴含着更多的信息。

由此可见，可以自然地将“熵”的概念引入信息论中，用于衡量一个随机变量的取值结果的“不确定性”的平均水平，从而反应某一具体事件发生蕴含的信息量。直观理解就是随机实验中，某一随机事件的不确定性越大，所蕴含的信息量也就越大，也即“熵”越大。

因此，一个合适的信息函数 $I : P \to \mathcal{R}$ 应当满足：

1. $I(p)$ 关于 p 单调递减;
2. $I(1) = 0$，即必然事件所蕴含的信息量为 0;
3. 两个独立事件同时发生所提供的信息量应等于各自分别蕴含的信息量之和，即 $I(p_1 \cdot p_2) = I(p_1) + I(p_2)$;

而对数函数则可以很好地满足上述要求：

$$I(p) = \log(\frac{1}{p}) = -\log(p) \tag{1}$$

自然地，熵作为一种复杂系统中“平均性”水平的描述 $H : \mathcal{X} \to [0, 1]$：

$$H(X) = E[I(X)] = E[-\log(p(X))] \tag{2}$$

对于离散形式的随机变量，则可以写成

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$$

2 Shannon's source coding theorem

2.1 Settings

在模型中，假定我们的数据是来自给定的字母表 A 中的一列独立同分布的字母 $X_1, X_2, \ldots$。对字母表中的每个字母 $a \in A$，其出现的概率为常数 $\mathcal{P}(X_n = a) = p_a$。

所谓数据压缩则是采用二元序列 0-1 编码这组数据，将取自 A 的数据子集 S 中的每个元素 s，映射为不同的 0-1 的序列 $\phi(s)$。由于 S 中的信息 $s = (X_1, X_2, \ldots, X_n)$ 具有确定的概率分布，从而压缩后的数据 $\phi(s)$ 的长度 L 则也可以看作是随机变量, 而压缩的目标则是使 $E[L]$ 尽可能地小，但同时尽量不能损失原数据的信息。

一种直接的编码方式则是采用长度为 $\lceil \log_2(|A|) \rceil$ 的序列编码 A 中的字母，即 $A = \{a, b, c, d\}$，则编码 $a = 00, b = 01, c = 10, d = 11$，此时 $s = (X_1, X_2, \ldots, X_n)$ 则被编码后的长度即为 $n\lceil \log_2(|A|) \rceil$，即 $E[L] = n\lceil \log_2(|A|) \rceil$。那么是否能对这种编码方式进行改进? 最短能用多长的 0-1 序列去无损的编码原本的信息?

2.2 Main Theorem

1948 年，香农在其文章《A Mathematical Theory of Communication》中提出了"source coding theorem"，即指出 n 个蕴含着信息熵 $H(X)$ 的随机变量可被压缩成平均长度略超过 $nH(X)$ 的 0-1 序列，而超出部分的长度在 $n \to \infty$ 时可忽略。

定理 2.1 (Shannon's entropy Theorem)**.** *X 为值取自集合 A 的随机变量, 拥有概率分布 $P(X = a) = p_a$ $a \in A$，其信息熵记为 $H(X) = -\sum_{a \in A} p_a \log_2(p_a)$。$\phi$ 则是对 n 个独立同分布的随机变量组成的序列 $s = (X_1, X_2, \ldots, X_n)$ 的编码函数, 将 s 映射为一列 0-1 的序列 $\phi(s)$。则 $\phi(s)$ 的长度在平均意义下至少为 Hn, 即*

$$E(L) \geq Hn + o(n) \tag{3}$$

同时，也存在一种编码方式 ϕ 满足

$$E(L) \leq Hn + o(n) \tag{4}$$

Proof. 对长度为 n 的随机变量序列，如果在给定观测结果中 a_i 的个数为 n_i 的条件下，不同序列个数总共为 $\binom{n}{n_1, n_2, \ldots, n_k}$。

由二进制编码，k 个 0/1 的字节可以编码 2^k 个不同的数，从而可知而编码为 $\binom{n}{n_1, n_2, \ldots, n_k}$ 种序列所需要的字节数目至少为 $\log_2 \binom{n}{n_1, n_2, \ldots, n_k}$，对其采用 stirling 公式 $n! = \sqrt{2\pi n}(\frac{n}{e})^n$ 近似

$$\begin{aligned}\log_2 \binom{n}{n_1, n_2, \ldots, n_k} &= \log_2(n!) - \sum_{i=1}^{k} \log_2(n_i!) \\ &= n\log_2(n) - \sum_{i=1}^{k} n_i \log_2(n_i) + o(n) \\ &= -\sum_{i=1}^{k} n_i \log_2(\frac{n_i}{n}) + o(n) \\ &= nH(\frac{n_1}{n}, \ldots, \frac{n_k}{n}) + o(n)\end{aligned}$$

(1) 因此编码序列长度的条件下界为:

$$E[L|n_1, \ldots, n_k] \geq \log_2 \binom{n}{n_1, n_2, \ldots, n_k} = nH + o(n) \tag{5}$$

由大数定律 $\frac{n_i}{n} \underset{p}{\to} p_i$，从而对任意 $\epsilon > 0$

$$P(|n_i - np_i| \geq \epsilon n) \leq \frac{E[n_i - np_i]^2}{\epsilon^2 n^2} = \frac{p_i(1-p_i)}{\epsilon^2 n}$$

记 $\forall i \in \{1, 2, \ldots, k\}$，$|n_i - np_i| \leq \epsilon n$ 均成立为事件 Ω_ϵ, 其对立事件为 Ω_ϵ^c，从而有

$$P(\Omega_\epsilon^c) = P(\cup\{n_i : |n_i - np_i| > \epsilon n\}) \leq \sum_{i=1}^{k} \frac{p_i(1-p_i)}{\epsilon^2 n} \leq \frac{1}{\epsilon^2 n}(1 - \frac{1}{k}) < \frac{1}{\epsilon^2 n}$$

第一个等号在 $p_1 = \cdots = p_k = \frac{1}{k}$ 时取到;
即有

$$P(\Omega_\epsilon) \geq 1 - \frac{1}{\epsilon^2 n}$$

且 $(n_1, \ldots, n_k) \in \Omega_\epsilon^c$ 时，$np_i - \epsilon \leq n_i \leq np_i + \epsilon$

$$E[L|\Omega_\epsilon] \geq -n \sum i = 1k(p_i - \epsilon)\log_2(p_i + \epsilon) + o(n)$$

则采用条件期望公式:

$$\begin{aligned}E[L] &= E[L|\Omega_\epsilon]P(\Omega_\epsilon) + E[L|\Omega_\epsilon^c]P(\Omega_\epsilon^c) \\ &\geq E[L|\Omega_\epsilon]P(\Omega_\epsilon) + 0 \\ &\geq [-n\frac{i=1}{k}(p_i - \epsilon)\log_2(p_i + \epsilon) + o(n)](1 - \frac{1}{\epsilon^2 n})\end{aligned}$$

由 ϵ 的任意性，得到

$$E[L] \geq nH + o(n) \tag{6}$$

(2) 对于编码平均长度上界的估计，依然通过

$$E[L] = E[L|\Omega_\epsilon]P(\Omega_\epsilon) + E[L|\Omega_\epsilon^c]P(\Omega_\epsilon^c)$$

由 $P(\Omega_\epsilon^c) \leq \frac{1}{\epsilon^2 n}$, 因此当随机变量列 $(n_1, \ldots, n_k)$ 的取值在 Ω_ϵ^c 内时，不对原始随机变量列进行压缩，所需长度为 $log_2(k^n) = n\log_2(k)$，则 $E[L|\Omega_\epsilon^c]P(\Omega_\epsilon^c)$ 在 $n \to \infty$ 时可以视作常数项；

而对 $(n_1, \ldots, n_k) \in \Omega_\epsilon$，长度为 n 的随机变量序列的所有可能取值为为：

$$\begin{aligned}\sum_{i=1}^{k}\sum_{n_i\in\Omega_\epsilon}\binom{n}{n_1, n_2, \ldots, n_k} &\leq \sum_{i=1}^{k}\sum_{n_i\in\Omega_\epsilon}\binom{n}{n(p_1\pm\epsilon), n(p_2\pm\epsilon), \ldots, n(p_k\pm\epsilon)} \\ &\leq n^k\binom{n}{n(p_1\pm\epsilon), n(p_2\pm\epsilon), \ldots, n(p_k\pm\epsilon)}\end{aligned}$$

因此

$$\begin{aligned}E[L|\Omega_\epsilon]P(\Omega_\epsilon) &\leq E[L|\Omega_\epsilon] \\ &= \log_2\sum_{i=1}^{k}\sum_{n_i\in\Omega_\epsilon}\binom{n}{n_1, n_2, \ldots, n_k} \\ &\leq \log_2(n^k\binom{n}{n(p_1\pm\epsilon), n(p_2\pm\epsilon), \ldots, n(p_k\pm\epsilon)}) \\ &= k\log_2 n + nH + cn\epsilon\end{aligned}$$

由 $\log_2 n = o(n)$ 及 ϵ 的任意性, 得到

$$E[L] \leq nH + o(n) \tag{7}$$

□

2.3 Coding Method

上述定理给出了压缩编码长度的上下界估计，而 Huffman 则具体地给出了最优的编码算法，使得 $n \to \infty$ 时，编码的平均长度达到渐进下界。而 Huffman 编码的核心思想是选取出现次数最少的字母作为最底层的叶子，逐步构造树图；再从根节点开始左 0 右 1 沿着边依次往下编码。

References

MIT notes for Shannon's Entropy Theorem

wikipedia for shannon's source coding theorem

wikipedia for entropy

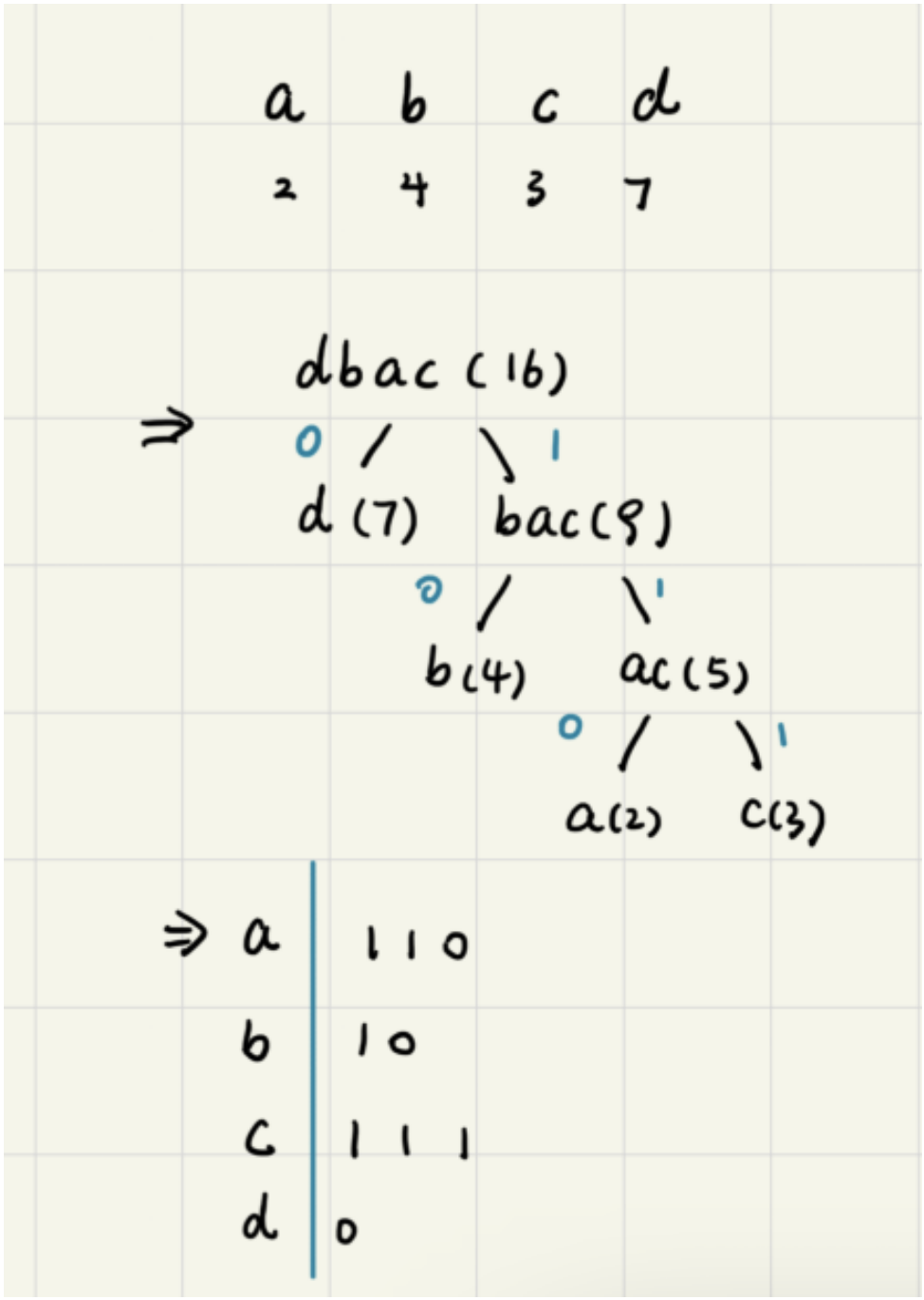

Figure 1: Huffman 编码示意图