信源与信源无失真编码

2020年4月3日 21:59

0. 概述

我们可以通过信息论来讨论无失真压缩的极限。这也是香农信息论的第一大定律，信源无失真编码定理。基本思路则是在信息量不变的前提下，提高每个码字的信息量，从而降低码字长度。那么我们已知的一般是从非等概分布逐渐走向等概分布的，我们可以通过渐进等同分割性质和典型序列完成这一目标。下面来讨论一下。

1. 渐进等同分割性质

首先回顾一下概率论中的大数定律。伯努利在1713年首先给出了频率稳定性的数学表述，该结论是大数定律的第一个。

在一个"成功"概率为p的伯努利试验序列中

$$X_n=\begin{cases}1, & \text{第一次试验A发生.}\\ 0, & \text{第一次试验A不发生.}\end{cases}\qquad S_n=\sum_{i=1}^{n}X_i,\quad n=1,2,\cdots$$

对于任意正数 ε 和 δ，存在正整数N，当 $n>N$ 时，有

$$P\left(\left|\frac{S_n}{n}-p\right|\geq\varepsilon\right)<\delta \quad 或 \quad P\left(\left|\frac{S_n}{n}-p\right|<\varepsilon\right)\geq 1-\delta$$

通俗点讲就是频率等于概率，经典分布趋于统计分布。

那么渐进同等分割（Asymptotic Equipartition Property）就是对一段随机变量序列进行等同分割，让长度从一进行到无穷大。

由弱大数定律可知，当分割长度无限大时，任何事件出现的频率都将趋于概率，那么差不多所有的事件都将等概出现，对于数据压缩的意义也很清晰。(即每一个事件出现的个数 n_k，有 $\left|\frac{n_k}{n}-p(k)\right|<\varepsilon$)

- 等同分割 ⇒ 等概率 ⇒ 最大离散熵 ⇒ 定长编码定理。
- 暗示了一种定长编码的方法。

下面我们对其进行严格的数学推导：

如果 $X_1, X_2, \cdots$ 是独立同分布的离散随机变量，分布服从 $p(x)$，则 $-\frac{1}{n}\log p(X_1,X_2,\cdots,X_n)\xrightarrow{P}H(X)$

"$\xrightarrow{P}$"代表以概率P收敛于 $H(X)$。

对于这个数学表达可以这样理解。首先我们考虑的都是无记忆的稳恒信源，即

$$P(w_i, w_{i+1},\cdots,w_{i+N}=\vec{A}) = P(w_j, w_{j+1},\cdots,w_{j+N}=\vec{A})$$

其中A为固定的字符序列。这样就是离散平稳随机过程。那么对于某一随机序列的信息量即为 $-\log p(x_1,x_2,\cdots,x_n)$

当n趋于无穷大时，则有弱大数定律，成为等概的情况。$\frac{1}{n}-\log p(x_1,x_2,\cdots,x_n)$ 即为平均一个符号携带的信息量，那么即为信息熵。

偏离这个等概分布的概率很小。

由此我们就可以给出一个典型序列，我们给出数学上的严格定义。

相对于分布 $p(x)$ 和序列 $(x_1,x_2,\cdots,x_n)\in X_n$，典型序列集合 $A_\varepsilon^{(n)}$ 定义为满足下列不等式约束的所有序列集合

$$2^{-n(H(X)+\varepsilon)}\leq p(x)=p(x_1,x_2,\cdots,x_n)\leq 2^{-n(H(X)-\varepsilon)}$$

这个其中也是由弱大数定律直接得来。

$$P\left(\left|\frac{S_n}{n}-p\right|\geq\varepsilon\right)<\delta \;\longrightarrow\; \log p(x)=\sum_{k=1}^{K}n_k\log p(a_k)\Rightarrow n\cdot\sum_{k=1}^{K}\frac{n_k}{n}P(a_k)$$

我们再将正数替换，即得

$$2^{-n(H(X)+\varepsilon)}\leq p(x)=p(x_1,x_2,\cdots,x_n)\leq 2^{-n(H(X)-\varepsilon)}$$

由上述的证明过程和定义，我们发现典型序列满足以下性质

1. 若 $x\in A_\varepsilon^{(n)}$，则 $H(X)-\varepsilon\leq-\frac{1}{n}\log p(x)\leq H(X)+\varepsilon$.
2. 若n足够大，$\Pr\{A_\varepsilon^{(n)}\}>1-\varepsilon$. (这里我们给出一个证明方法，见附录)
3. $(1-\varepsilon)2^{n(H(X)-\varepsilon)}\leq|A_\varepsilon^{(n)}|\leq 2^{n(H(X)+\varepsilon)}$

2. 若n足够大，$P_r\{A_\varepsilon^{(n)}\} > 1-\varepsilon$。（这里我们给出一个证明方法，见附录）

3. $(1-\varepsilon)2^{n(H(X)-\varepsilon)} \leq |A_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$

性质1和性质2可以直接由定义得到。简单证明一下性质3

$$1=\sum_{x^n\in X^n} P(x^n) \geq \sum_{x^n\in A_\varepsilon^{(n)}} P(x^n) \geq \sum_{x^n\in A_\varepsilon^{(n)}} 2^{-n(H(X)+\varepsilon)} = 2^{-n(H(X)+\varepsilon)}|A_\varepsilon^{(n)}|$$

$$1-\varepsilon \leq P_r\{A_\varepsilon^{(n)}\} \leq \sum_{x^n\in A_\varepsilon^{(n)}} 2^{-n(H(X)-\varepsilon)} = 2^{-n(H(X)-\varepsilon)}|A_\varepsilon^{(n)}|. \Rightarrow A_\varepsilon^{(n)} \geq (1-\varepsilon)2^{n(H(X)-\varepsilon)}$$

由此性质3得证明完毕。
典型序列的个数大约有 $2^{n(H(X)+\varepsilon)}$ 个，而 $|X|^n$ 有很多，典型序列的集合只是所有可能序列的一个子集，但几乎所有序列几乎都会是典型序列。致此，我们完成了几乎没有失真，从非等概到等概的过程，我们可以基于这个达到无失真的定长编码定理。

2. 定长编码定理.

定长编码定理就是尽量把N取得很大，从而使序列几乎都是典型序列，从而在出错率较小的情况下，尽可能的压缩码长。下面我们给出定理

设 X^n 是由独立同分布离散随机变量 $x\sim p(x)$ 构成的序列。对于任意正数ε，总存在足够大的n，可以找到一个一一映射，将 X^n 映射成二进制序列，且满足

$$E\left[\frac{1}{n}l(X^n)\right] \leq H(X)+\varepsilon.$$

下面我们先简单的证明并给出一个例子，最后总结一下，

我们编码的平均码长为 $E[l(X^n)]$

$$E[l(X^n)] = \sum_{x^n\in X^n} p(x^n)l(x^n) = \sum_{x^n\in A_\varepsilon^{(n)}} p(x^n)l(x^n) + \sum_{x^n\in \overline{A_\varepsilon^{(n)}}} p(x^n)l(x^n)$$

$$\leq \sum_{x^n\in A_\varepsilon^{n}} p(x^n)\cdot[n(H(X)+\varepsilon)+1+1] + \sum_{x^n\in \overline{A_\varepsilon^{(n)}}} p(x^n)[n\log|X|+2]$$

$$= P_r\{A_\varepsilon^{n}\}[nH(X)+n\varepsilon+2] + P_r\{\overline{A_\varepsilon^{(n)}}\}[n\log|X|+2]$$

$$\leq n(H(X)+\varepsilon)+2+n\varepsilon\log|X|+2\varepsilon$$

$$= n(H(X)+\varepsilon'). \quad \text{其中 } \varepsilon' = \varepsilon+\varepsilon\log|X|+\frac{2\varepsilon}{n}$$

解释说明：

平均码长可分为典型序列和非典型序列之和。

套用典型序列的定义进行缩放，非典型序列直接缩放到最大值。

+1，+1是进行取整和标记是否为典型序列。

代入定义，继续缩放。

可见，$n\to\infty$ 时，$\varepsilon'\to 0$。于是得证。

取一个实际的例子。

设信源输出符号分布为

$$\begin{pmatrix} X \\ p(x) \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 & a_7 & a_8 \\ 0.40 & 0.18 & 0.10 & 0.10 & 0.07 & 0.06 & 0.05 & 0.04 \end{pmatrix}$$

要求编码错误概率小于 10^{-8}，且编码效率大于90%，求最小可行分组长度。

首先定义一下错误概率，也就是出现非典型序列的概率。

$$P_r(\overline{A_\varepsilon^{(n)}}) \leq \frac{D(I(x))}{N\varepsilon^2}.$$

这个公式目前不知道怎么得来的。
接着，我们定义编码效率。编码效率的极限自然是完全去掉冗余度，即使条件增多，信息熵也尽可能不会减小。那么极限自然是 $H_\infty(X)$ 即 $H(X)$。那么我们能达到的则是 $H(X)+\varepsilon'$，这在上面的证明中也给出了。编码效率则为

$$\frac{H(X)}{H(X)+\varepsilon}, \quad \frac{H(X)}{H(X)+\varepsilon} > 0.9 \Rightarrow \varepsilon < 0.28 \xrightarrow{P_r(\overline{A_\varepsilon^{(n)}})\leq 10^{-8}} N > 10^{8}.$$

关于定长编码定理的总结.

$\frac{}{H(X)+\varepsilon}$, $\frac{}{H(X)+\varepsilon} > 0.9 \Rightarrow \varepsilon < 0.28 \Longrightarrow N > 10^{8}$.

关于定长编码定理的总结：

首先我们需要知道为什么会有定长编码，为此我们需要有一个熵率的概念，这个概念描述的是随着分组码的增长，平均每个字母的熵 $H_N(N)$ 的极限值，数学上即 $H_\infty(U) = \frac{1}{N} H(U_1, U_2, \ldots, U_N)$

而对于离散平稳信源，有一条定理 $H_\infty(U) = \lim_{N\to\infty} H(U_N/U_1, U_2 \cdots U_{N-1})$，证明过程可以参见教材，较为核心见下。

根据离散平稳信源的定义

$$H(U_N | U_2, U_3, \cdots, U_{N-1}) = H(U_{N-1} | U_1, \cdots, U_{N-2})$$

这体现了时间的无关性，而根据条件减少熵，

$$H(U_N | U_2, U_3, \cdots, U_{N-1}) \geq H(U_N | U_1, \ldots, U_{N-1})$$

我们可以得到条件熵是随着N的增大而减小的。

$$NH_N(U) = H(U_1) + H(U_2|U_1) + \cdots + H(U_N | U_1 U_2, \cdots, U_{N-1})$$
$$\geq NH(U_N | U_1 U_2 \cdots U_{N-1})$$

我们可以知道，信源输出中前后字母之间的统计关系是使信息熵减少的重要原因。

又一方面，$NH_N(U) = H(U_N | U_1 U_2 \cdots U_{N-1}) + (N-1)H_{N-1}(U)$，将 $H_N(U) \geq H(U_N | U_1 U_2 \cdots U_{N-1})$

可以得到 $H_N(U) \leq H_{N-1}(U)$，下界存在。

我们可知 N 减少时，熵增大，由熵的极值性可知 $H_1(U) \leq \log K$，K为信源字母表总数。

因此，当信源的一维分布不均匀时，即使不存在依存关系，也不能最大限度的携带信息

那么反映信息有效程度的物理量就是冗余度和相对冗余度了。

冗余度：$\log K - H_\infty(U)$，相对冗余度：$1 - \frac{H_\infty(U)}{\log K}$ 我们进行定长编码就是使非独立变为独立，非等概变为等概的过程。

关于定长编码定理我们还有一个形式的定理：

设熵率为 $H_\infty(U)$ 的离散无记忆信源被分成 N 个一组序列组，并用长为 M 的码字组进行编码，码字母表大小为 J。

则对于任意的 $\varepsilon > 0$ 和 $\delta > 0$，只要 N 足够大，且满足不等式

$$\frac{M}{N} \log J > H_\infty(U) + \delta.$$

则源字母组没有自己特定码字的概率为 p_e 可以小于 ε

这个定理证明比较简单

$$N_G = 2^{N(H_\infty(U)+\delta)}$$

我们只编码典型序列，那么总事码的总数大于典型序列数。

$$J^M \geq 2^{N(H_\infty(U)+\delta)} \Rightarrow M\log J \geq N(H_\infty(U)+\delta) \Rightarrow \frac{M}{N}\log J \geq H_\infty(U)+\delta.$$

但是定长编码定理只是证明了理论上的可行性，实际上，正如例题所示，一个较好的定长编码需要 10^8 的分组长度，这无论是时间延迟，还是硬件成本都是无法接受的，我们也无法找到一个满意的构造方案，这也就说明应用范围不广的弊端。

3. 变长码

3.1 变长码的类型

研究无失真编码一定要保证是一一映射，那么我们称为非奇异码，

若一个码 C 可以将不同的 X 映射为不同的 D 序列，即

$$x \neq x' \Rightarrow C(x) \neq C(x')$$

则称该码是非奇异的。

但是虽然是一一映射也会有麻烦，例如，对于不等长的编码

$C(x_1) = 0$, $C(x_2) = 1$, $C(x_3) = 01$

01 码的解出现了歧义，那么我们为了改进则有唯一可译码，定义如下。

称码 C^+ 是码 C 的扩展，当 C^+ 是有限长 X 序列到有限长 D 进制序列的映射，且满足

$$C(x_1 x_2 \cdots x_n) = C(x_1)C(x_2)\cdots C(x_n)$$

那么当其扩展是非奇异的，我们则称为唯一可译码

这其实只是一个定义上的解答，并没有可建设性的方案、换言之，没有码字是码字的组合

例如 $C(x_1) = 0$, $C(x_2) = 01$, $C(x_3) = 11$，则序列 0011100 10 唯一可译为 $x_1 x_1 x_3 x_1 x_2 x_1$

但是性质上的计算我们可以分析，但是有一个连续的并且直接的快速确定同 …

这其实则是一个定义上的解答，并没有可建设性的方案。换言之，没有码字是码字的组合
例如 $C(x_1)=0$，$C(x_2)=01$，$C(x_3)=11$，则序列 0011100010 唯一可译为 $x_1x_1x_3x_1x_2x_1$
但是经过自己的计算我们可以发现，这是有一个译错并重译的过程，需要回退。
虽然唯一可译码可以解决译码的问题，但对于解码器的要求较高，于是我们需要一种速度更快的码，称为即时码，也称为前缀码

当没有码字是其他码字的前缀，那么则称为前缀码或即时码。

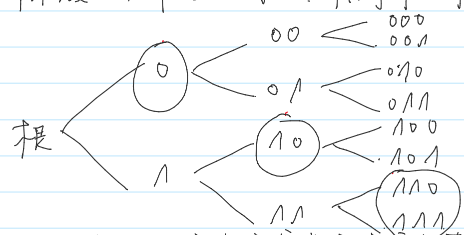
例如：$C(1)=0$，$C(2)=10$，$C(3)=110$，$C(4)=111$，则序列 0101111110110 直接可译为 1 2 4 3 2.
这是一种很好性质的码。
简单提一下各种码之间的关系：

即时码 $\subseteq$ 唯一可译码 $\subseteq$ 非奇异码 $\subseteq$ 所有码.

莫尔斯电码不是唯一可译码，需要先根据语法。
汉语：没有鸡鸭也可没有鱼肉也可青菜豆腐不可少。这是歧义句，没有汉语没有标点也不是唯一可译码。

3.2 Kraft不等式与香农第一定理

我们已经知道了前缀码具有较好的性能，我们从定义中也可以直接得到，前缀码不能使用最短的码字，否则前缀的条件无法满足，我们可以用二叉树来表示。



我们能够发现，无论是定长还是变长都对应树的叶子。例如，我们可以这样选取前缀码
那么我们仍需要回答如何构造前缀码？前缀码存在的条件是什么？
实际上通过树图，我们已经可以很直观的发现其中的规律，下面我们要做的是用数学语言来描述
由于前缀码只在树上，D进制则为D叉树，若被选为码字则后面的叶子会被删除，设其码长度为 l_i，最长为 l_{max}，
所以有
$$D^{-l_i} = D^{-l_{max}} \cdot D^{-l_i+l_{max}}$$
于是，设含有K个信源字母的信源要用J个字母的码字母表进行变长编码，则当且仅当各码字的长度 $l_1, l_2, \cdots, l_k$ 满足Kraft不等式
$$\sum_{k=1}^{K} J^{-l_k} \le 1.$$
我们已经证明了必要性了，关于充分性的证明其实差不多，我们总是要砍掉其他的叶子，所以这个过程一定是可逆的，所以这一定是个充要条件，但是这仅仅是从编码的构造考虑，与码是否最优无关，以后我们再考虑最优性。
关于码的长度，与定长编码一致，有以下的定理。

对于随机变量X进行D-进制前缀码编码，得到的码长满足 $L \ge H_D(X)$，等号当且仅当 $D^{-l_i}=p_i$ 时成立。

证明如下：
$$L - H_D(X) = \sum p_i l_i - \sum p_i \log_D \frac{1}{p_i}$$
$$= \sum p_i \log_D D^{l_i} - \sum p_i \log_D \frac{1}{p_i} = \sum p_i \log_D p_i - \sum p_i \log_D D^{-l_i}$$
$$= \sum p_i \log_D p_i - \sum p_i \log_D \frac{D^{-l_i}}{\Delta}\cdot\Delta = \sum p_i \log_D p_i - \sum p_i \log_D \frac{D^{-l_i}}{\Delta} - \log_D \Delta$$
其中 $\Delta = \sum_i D^{-l_i}$，那么上式则变为
$$L - H_D(X) = \sum p_i \log_D \frac{p_i}{D^{-l_i}/\Delta} + \log_D \frac{1}{\Delta}$$
根据鉴别信息或对数求和不等式，第一项大于0，由Kraft不等式，第二项也大于0，那么得证
于是我们找到了一个下界，我们发现当 $D^{-l_i}=p_i$ 时，可以直接取等。
这里稍微解释一下这个定理，对于一个有K个字母的信源，每个码出现的概率为 p_i，那么这个信源的熵则为他所携带的信息，而我们 [illegible]

于是我们找到了一个下界。我们发现当 $D^{-l_i} = p_i$ 时，可以直接取等。
这里稍微解释一下这个定理，对于一个有k个字母的信源，每个码出现的概率为 p_i，那么这个信源的熵即为他所携带的信息量。而我们的D进制编码实际上是一个D叉树，于是我们利用了这个D叉树实际也直接定义了一个新的概率分布。这个码的等概特性一定会比之前的来的更好，至少也是相同。所以我们就有了这样一个结论。
实际上鉴别信息也告诉我们了一定道理。要让这个平均长度也就是新的分布信息熵更小，我们应该让他的分布更可能的靠近原始分布，这样可以使信息熵的鉴别信息下降。
那么最优码长应为多少呢？有以下定理。这个定理也被称为香农第一定理。

对随机变量X进行D-进制前缀编码，得到的最优码长不等式满足下列不等式

$$H_D(X) \le L^* \le H_D(X) + 1$$

$H_D(X)$ 对于非很立的情形应严格的写为 $H_\infty(U)/\log J$
这个定理的证明也是非常巧妙的，香农也构想了一个香农码。使码长为 $\lceil \log \frac{1}{p_i} \rceil$，那么

$$\sum D^{-\lceil \log \frac{1}{p_i} \rceil} \le \sum D^{-\log \frac{1}{p_i}} = \sum p_i = 1$$

所以满足 kraft 不定义，一定是一个唯一可译码。

$$\log_D \frac{1}{p_i} \le \log_D \lceil \tfrac{1}{p_i} \rceil \le \log_D \frac{1}{p_i} + 1$$

$$\Rightarrow \quad H_D(X) \le \sum p_i l_i \le H_D(X) + 1$$

$$\Rightarrow \quad H_D(X) \le L^* \le H_D(X) + 1$$

最优码一定不能比这个码更差，所以巧妙的证明了结论。
那么定长编码告诉我们我们可以做到平均码长和之前相比只有ε的误差，而变长编码则有1比特。
我们还可以将两者结合做的更好。

对于信源X进行分组前缀编码，得到每消息符号数 L_n^* 满足不等式

$$\frac{H(X_1, X_2, \cdots X_n)}{n} \le L_n^* \le \frac{H(X_1, \cdots, X_n)}{n} + \frac{1}{n}$$

这个过程就是先分组，然后进行前缀编码。当信源稳恒，那么 $L_n^* \sim H(X)$。
如果我们对信源分布的估计出了偏差，那么性能一定不会变好。因为鉴别信息的非负性，定量的分析有如下定理

对于服从 $p(x)$ 信源X进行前缀编码，若码字长度取 $l(x) = \lceil \log \frac{1}{q(x)} \rceil$，则平均码长满足

$$H(p) + D(p\|q) \le E_p\, l(x) < H(p) + D(p\|q) + 1$$

因为 $H(q) = H(p) + D(p\|q)$，所以这个定理实际上并不需要证明。

3.3. 几种前缀变长码分析

3.3.1 香农码

香农码在一般情况下并非最优情况。
我们举一个简单例子。

$K=2$，$p(k_1) = 0.9999$，$P(k_2) = 0.0001$。那么 $l_1 = 1$，$l_2 = 14$。
我们可以计算一下

$$H(K) = P(k_1)\log\frac{1}{P(k_1)} + P(k_2)\log\frac{1}{P(k_2)} = 1.47 \times 10^{-3}$$

$E(l_{shannon}) = 0.9999 \times 1 + 0.0001 \times 14 = 1.0013$。但是我们显然可以用1比特搞定。且他离最优解差的很远。
其原因也很显而易见，当对小概率的 $\log \frac{1}{p_i}$ 向上取整时，误差可能会非常大，这时我们没有办法减小概率很小的码长。

3.3.2. Huffman 编码

Huffman 作为博士生在1952年提出，可以证明Huffman编码具有最优性，即在给定信源符号的概率分布和码字母表的前提下，没有其他码能获得比 Huffman 编码更短的平均码长。
先抛开Huffman编码不谈，我们对上面的例子进行更深一点的分析。

$$H(K) = P(k_1)\log\frac{1}{P(k_1)} + P(k_2)\log\frac{1}{P_{k_2}} = P(k_1)\cdot 1.44\times 10^{-3} + P(k_2)\cdot 13.2$$

$$E(l) = P(k_1)\, l_1 + P(k_2)\, l_2$$

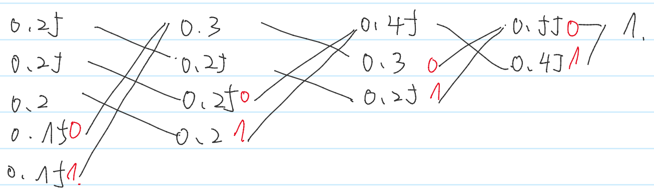
通过二叉树我们可以发现，让 k_2 的长度可以为1，最后得到长度为1的平均码长。
当然，这个过程通过了鉴别信息对于平均码长的理论指导。但是，我们经过人为的修正，可以更好的缩少误差。Huffman 就是这么做的，让概率大的信源的码长更可能短、当然前提是能短则短。

当然，这个过程仍存在监测信息对于平均码长的理论指导。但是，我们经过人为的修正，可以更好的缩少误差。Huffman 真正做到使概率大的信源的码长更可能短，当然前提是能短则短。

编码过程实际上是非常简单的。按概率由大到小分布，然后进行叉树排列，我们举一例来说明。

考虑 $\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 0.25 & 0.25 & 0.2 & 0.15 & 0.15 \end{pmatrix}$

Huffman 编码过程。



最后得到

$1 \longrightarrow 01$

$2 \longrightarrow 10$

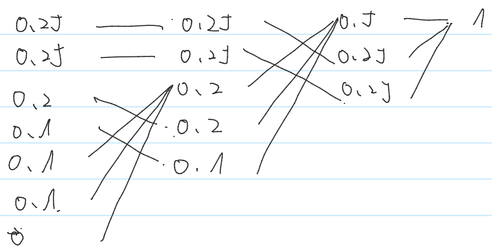
$3 \longrightarrow 11$

$4 \longrightarrow 000$

$5 \longrightarrow 001$

三进制有三个合并，如果出现了问题可以填一些假字母，如下所示情形，每次迭代缩减减少 $(D-1)$，由于最终有一定要求，所以要由 D 缩过为 1，填充后的总长一定为 $1+k(D-1)$

Codeword	X
1	1
2	2
01	3
02	4
000	5
001	6



Huffman 的精髓在于他最优性。我们接下来就从最简单的二元情形证明他，这个证明也比较容易推广到其他形式。

前面通过分析也发现了，对于二元情形，我们要让他尽可能短，一定是依次排列，使概率最小码字尽可能长，于是我们有两条定理

1. 对每一给定的离散无记忆信源，存在一个最优的二元前缀码，这个码中最不可能出现的两个码字必须具有相同的长度且码中相同长度的码字有两个或两个以上时，其中必有两个码字的差别只在最后一位。

2. 设 C' 是某信源经缩减后得到的缩减信源的最优前缀码，将 C' 中由原信源的最小概率的两个字母缩减得到的字母所对应的码字后各加 0 和 1，作为原信源的最小概率的两个码字，而其余码字不变，则这样得到的码 C 对原信源也是最优的。

第一个定理的证明可以采用反证法，第二个则可以用第一个定理得到。详细的证明过程可以参见《应用信息论基础》。

3.3.3 算术码

在之前的证明提到，当 $-\log_2 p(a_k)$ 不为整数时，我们要通过扩展信源的办法来改进压缩效果，即需要分组，这自然会使编译码更复杂。我们可以用树码来解决这一个问题。

对于离散无记忆信源树码的一种有效编码方法是由 P. Elias 最早在一篇未发表的研究报告中提到，后经 F. Jelinck、J. Rissanen 等人改进和发展进入实用，现在人们一般称之为算术码 (arithmetic coding)

3.3.3.1 过程描述

设信源字母表为 $\{a_1, a_2, \cdots, a_k\}$，字母 a_k 的概率是为 $p(a_k)$ $(k=1,2,\cdots,k)$，将字母按其顺序排序并记作 $a_1 > a_2 > \cdots > a_k$，于是，可以定义了字母 a_k 的累积概率

$$F(a_k) = \sum_{a_i \geq a_k} p(a_i)$$

再定义修正概率

$$\overline{F}(a_k) = \sum_{a_i > a_k} p(a_i) + \frac{1}{2} p(a_k)$$

那么我们会得到一个概率值 $F(a_k)$ 和 $\overline{F}(a_k)$，将这个值用相应的进制表示，则可以得到码字。

举一个栗子。

a_k	$p(a_k)$	$F(a_k)$	$\overline{F}(a_k)$	$\overline{F}(a_k)$ 二进制表示	码字	Huffman 编码
a_1	0.25	0.25	0.125	0.001	001	10
a_2	0.5	0.75	0.5	0.10	10	0
a_3	0.125	0.875	0.8125	0.1101	1101	110
a_4	0.125	1	0.9375	0.1111	1111	111

我们可以看到，对于信源序列直接进行编码效果不是很好，还不如Huffman编码，这是因为算术算是修正的$F(a_k)$代表的也是一个线段长度，所以是有冗余的，当有对信源做扩展，还是当我段的长度缩短，当信源长度趋于无限时，P所在在的区间成为一点，那么冗余则会随着长度自然的降低

举一个两项分布的例子 $X \sim B(0,6)$，

设信源在输出第$n-1$个符号后，P所在的区间为 $[A_{n-1}, B_{n-1})$，那么有

$$\begin{cases} A_n = A_{n-1} \\ B_n = A_{n-1} + 0.6(A_{n-1} - B_{n-1}) \end{cases} \quad u_n = a_0 \qquad \begin{cases} A_n = A_{n-1} + 0.6(B_{n-1} - A_{n-1}) \\ B_n = B_{n-1} \end{cases} \quad u_n = a_1$$

我们可以看出随着n增大区间越来越小，而最后趋于一个点，

当信源输出第N个信源序列时，码字输出的是此时A_n和B_n中某一致的部分，此时输出二进制数的位数。

二元表示码的长度应为m，

$$\frac{1}{2^m} > \prod_{n=1}^{N} p(u_N) > \frac{1}{2^{m+1}}$$

3.3.3.2 极限情况下的压缩效果

对于长的信源序列序列，同样可以加以排序。

例如有 $V_N = u_1 u_2 \cdots u_N$，$V_N' = u_1' u_2' \cdots u_N'$

若存在$t \le N$，使

$$u_n = u_n' \quad 当 n < t 时$$
$$u_t > u_t' \quad 当 n = t 时$$

则记 $V_N > V_N'$，我们对此信源序列序列进行排序，可定义V_N的累积概率为

$$F(V_N) = \sum_{V_N' > V_N} p(V_N')$$

假设序列V_N后接续 u_{N+1}，形成新序列 $V_{N+1} = V_N u_{N+1}$，则此序列的概率为

$$p(V_{N+1}) = p(V_N) p(u_{N+1})$$

此时的累积概率为

$$F(V_{N+1}) = \sum_{V_{N+1}' > V_{N+1}} p(V_{N+1}')$$

$$= F(V_N) + p V_N \sum_{u_{N+1}' > u_N} p(u_{N+1}')$$

$$= F(V_N) + p(V_N) F(u_{N+1})$$

由于累积概率随着序列的严格增加，所以只需将区间

$$[F(V_N),\ F(V_N) + p(V_N) F(u_{N+1})]$$

用有限精度的二进制数加以表示，并将此二进制数与V_N加以对应，就可以实现无失真编码，由于这种情况下存在的充分条件仍然是kraft不等式，所在在极限情况下，算术码可以实现对信源的理想压缩

3.4 离散马尔可夫信源的熵率

关于马尔可夫链的基本内容这里不在描述，参阅随机信号处理。

求熵率首先要求条件熵

设过去某一时刻的状态为$S_1 = i$，现在要求n时刻信源送出序列的熵

$$H(U_n \mid U_1 U_2 \cdots U_{n-1}, S_1 = i)$$

$$= -\sum_{u_1 u_2 \cdots u_n} p(u_1 u_2 \cdots u_n \mid S_1 = i) \log(U_n \mid U_1 U_2 \cdots U_{n-1}, S_1 = i)$$

由马尔可夫信源的定义可知，

1. 某时刻信源输出哪个符号只与此时信源所处的状态有关，而与以前的状态以及之前的输出符号均无关。
2. 信源某时刻所处的状态，由当前输出符号和前一时刻信源的状态可唯一确定，

那么，我们可以得到

$$p(u_n \mid u_1 u_2 \cdots u_{n-1}, S_1 = i) = p(u_n \mid S_n)$$

$$p(u_1, u_2, \cdots u_n \mid S_1 = i) = p(u_1, u_2, u_{n-1}, S_n \mid S_1 = i) \cdot p(u_n \mid S_n)$$

将这两个结果代入定义式中，

$$H(U_n \mid u_1 u_2, \cdots u_{n-1}, S_1 = i)$$

$$= -\sum_{S_n u_1 u_2 \cdots u_n} p(S_n, u_1, u_2, \cdots u_{n-1} \mid S_1 = i) \cdot p(u_n \mid S_n) \log (u_n \mid S_n)$$

$$= \sum_{j=1}^{S} p(S_n = j \mid S_1 = 1) \, H(U_n \mid S_n = j)$$

那么我们代入所有的 S_1 即可得到条件熵

$$H(U_n \mid U_1 U_2 \cdots U_{n-1}, S_1) = \sum_{j=1}^{S} p(S_n = j) H(U_n \mid S_n = j)$$

对于稳态的马尔可夫链来，$p(\ n=j)$ 与 S_1 的概率分布无关，为平稳分布 $p(j)$，且 $H(U_n \mid S_n = j)$ 与 n 无关，可以记做 $H(U \mid S = j)$，那么，

$$H(U_n \mid U_1 U_2 \cdots U_{n-1}, S_1) = \sum_{j=1}^{S} p(j) H(U \mid S = j)$$

有了条件熵之后，就可以根据可加性求得

$$\frac{1}{N} H(U_1 U_2 \cdots U_N \mid S_1) = \frac{1}{N} \sum_{n=1}^{N} H(U_n \mid U_1 U_2 \cdots U_{n-1}, S_1)$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{S} p(j) H(U \mid S = j)$$

$$= \sum_{j=1}^{S} p(j) H(U \mid S = j)$$

如果不是平稳变的，可定义，

$$P_{(1,N)}(j) = \frac{1}{N} \sum_{n=1}^{N} p(S_n = j)$$

代入可得，

$$H(U_1 U_2 \cdots U_N \mid S_1) = \sum_{j=1}^{S} P_{(1,N)}(j) H(U \mid S = j)$$

最后就可以求 $H_\infty(U)$ 了。

$$H(U_1 U_2 \cdots U_N) = I(S_1; U_1 \cdots U_N) + H(U_1 U_2 \cdots U_N \mid S_1)$$

其中，

$$I(S_1; U_1 \cdots U_N) = H(S_1) - H(S_1 \mid U_1 U_2 \cdots U_N) \leq H(S_1) \leq \log S.$$

代入 $H_\infty(U)$ 的定义式中，

$$H_\infty(U) = \lim_{N \to \infty} \frac{1}{N} H(U_1 U_2 \cdots U_N)$$

$$= \lim_{N \to \infty} \left[\frac{1}{N} I(S_1; U_1 U_2 \cdots U_N) + \frac{1}{N} H(U_1 U_2 \sim U_N \mid S_1) \right].$$

我们可知互信息 $I(S_1; U_1 \cdots U_N)$ 是有界的，所以 $\lim_{N \to \infty} \frac{1}{N} \log S = 0$.

$$H_\infty(U) = \lim_{N \to \infty} \frac{1}{N} H(U_1 U_2 \cdots U_N \mid S_1) = \sum_{j=1}^{S} p(j) H(U \mid S = j)$$

离散马尔可夫信源的熵率是由条件熵组成的，由条件减小熵可知马尔可夫信源的熵率相比之前提到的离散无记忆信源相比熵率更小。

马尔可夫链之所以这么重要，也是因为生活中很多情形下都可以用这个数学模型来表达，例如天气预报，语言文字等。若马尔可夫的输出字母概率与前m个字母输出字母有关，与更早的输出字母无关，则称之为m阶马尔可夫信源。若信源输出的字母总数为k，则一般情况下，m阶马尔可夫信源的状态数 $S = k^m$，显然m越高，对马尔可夫信源的数学模型就越复杂。

如何选择恰当阶数的马尔可夫信源来近似实际信源是工程上要考虑的重要问题，例如，英文字母通常认为前5个字母，虽然m是越大越好，但一般取5就够了。

3.5 离散马尔可夫信源的编码定理与最优编码

我们之前提到的一般的平稳遍历的信源定长编码定理可以实现理想压缩，这里也是成立的。

3.5 离散马尔可夫信源的编码定理与最佳编码.

我们之前提到了一般的平稳遍历的信源定长编码定理可以实现理想压缩.这里也是成立的.
对于变长编码,我们也可以简单的将离散无记忆信源的变长编码推广到离散马尔可夫信源.

设马尔可夫信源是离散,平稳,遍历的,其初始状态为 $S_1 = i$,信源输出的序列为 $u = u_1 \cdots u_N$. 每一序列可用变长编码的方法得到一个对应的码字,其码字长度 $L_i(u)$ 满足

$$J^{-L_i(u)} \le P(u_1 u_2 \cdots u_N / S_1 = i) \le J^{-L_i(u)+1}$$

此时有

$$\sum_{u} J^{-L_i(u)} \le \sum_{u} P(u_1 u_2 \cdots u_N / S_1 = i) = 1$$

那么所有码字满足 Kraft 不等式.我们应用香农第一定理.

$$\frac{H(U_1 U_2 \cdots U_N / S_1 = i)}{\log J} \le N \bar{L}_i \le \frac{H(U_1 U_2 \cdots U_N / S_1 = i)}{\log J} + 1$$

平均到每个字符则是

$$\frac{H(U_1 U_2 \cdots U_N / S_1)}{N \log J} \le \bar{L} \le \frac{H(U_1 U_2 \cdots U_N / S_1)}{N \log J} + \frac{1}{N}$$

让 N 趋于无穷,代入熵率.

$$\lim_{N \to \infty} \frac{1}{N} H(U_1 U_2 \cdots U_N / S_i) = \sum_{i=1}^{S} P(i) H(U = i)$$

即

$$H_\infty(U) = \sum_{i=1}^{\infty} P(i) H(U / S = i)$$

于是有

$$\frac{H_\infty(U)}{\log J} \le \bar{L} \le \frac{H_\infty(U)}{\log J} + \frac{1}{N}$$

就可以得到马尔可夫信源的变长编码定理.

当用 J 个字母的码字表对熵率为 $H_\infty(U)$ 的离散马尔可夫信源进行变长编码时,其平均码长满足

$$\frac{H_\infty(U)}{\log J} \le \bar{L} \le \frac{H_\infty(U)}{\log J} + \frac{1}{N}.$$

其中 N 是信源序列分组的长度
那么当 N 足够大时.$\bar{L}$ 是可以接近 $H_\infty(U) / \log J$,从而达到理想的压缩.

附录.

1.

$$P(X_n \in A_\varepsilon) > 1 - \varepsilon.$$

设 $Z_n = -\frac{1}{n} \log P(X_n) \quad -\frac{1}{n} \log \left[\prod_{i \in n} P(X_i)\right] = -\frac{1}{n} \sum_{i \in n} \log P(X_i)$

$\Rightarrow E(Z_n) = -\frac{1}{n} \cdot n \cdot H(X_k) = H(X)$

$\mathrm{var}(Z_n) = \frac{1}{n} \mathrm{var}(-\log_2 P(X))$

$P(X_n \in A_\varepsilon) = P\left(\left| -\frac{1}{n} \log_2 P(X_i) - H(X) \right| \le \varepsilon\right)$,由弱大数定律可得.上面我们也提到过.

$$= P(|Z_n - E(Z_n)| \le \varepsilon)$$

$= P(|Z_n - E(Z_n)| \le \varepsilon)$

应用马尔可夫不等式或者切比雪夫不等式

$$P(X_n \in A_\varepsilon) = 1 - P(|Z_n - E(Z_n)| \geqslant \varepsilon) \geqslant 1 - \frac{Var(Z_n)}{\varepsilon^2} = 1 - \frac{1}{n} \cdot \frac{Var(-\log_2 P(X))}{\varepsilon^2}$$

我们想要 $P(X_n \in A_\varepsilon) > 1 - \varepsilon$，则

$$\frac{1}{n} \cdot \frac{Var(-\log_2 P(X))}{\varepsilon^2} < \varepsilon. \Rightarrow n > Var(-\log_2 P(X))/\varepsilon^3.$$

则有 $P(X_n \in A_\varepsilon) > 1 - \varepsilon$.