第四章无失真信源编码

第三节定长码

陈兴同

中国矿业大学 数学学院

2021 年 8 月

内容提要

内容提要

定长码含义：

定长码是码字长度全相同的编码方法。
设信源字符空间为 $\mathcal{X} = \{a_1, a_2, \cdots, a_N\}$，
D 进码字符空间为 $\mathcal{U} = \{0, 1, 2, \cdots, D-1\}$。
每个 n 长消息 $x^{(n)}$ 都对应一个码长为 k 的码字 $u^{(k)}$，总码字数 M 就是要被编码的消息数。如果采用 D 进编码，则 $M \leq D^k$。
对于定长码来说码率 $R = \frac{k}{n}$ 个 D 进字符，它表示平均每信源字符占用码字符数。定长码的优点是编码译码都十分容易。

编码方法：

无失真

对离散无记忆信源的 n 长消息进行 D 进等长编码，要想译码不出错即 $P_e = 0$，必须每个消息对应不同的码字，因此可选码字总数不能小于消息总数即 $D^k \geq N^n$，此时码率 $R \geq \log_D N$。反之，如果码率满足 $R \geq \log_D N$，就有 $D^k \geq N^n$，取最小的 k 作为码长进行定长编码，则可以实现无错编码。

例题 4.3.1：

设离散无记忆信源字符集的分布为

$$X \sim \begin{pmatrix} a_1 & a_2 \\ 4/15 & 11/15 \end{pmatrix},$$

（1）则熵率为 $H(X) = 0.9544$ bits。

0.8366

续例题 4.3.1：

（2）由于 3 长消息总共有 8 个，故要对所有 3 长消息进行编码必须至少有 8 个码字才能保证无错，所以码长为 3bits，编码方案是

信源字符	$a_1a_1a_1$	$a_1a_1a_2$	$a_1a_2a_1$	$a_2a_1a_1$	$a_2a_1a_2$	$a_1a_2a_2$	a_2
码字	000	001	010	100	101	011	1

续例题 4.3.1：

（3）当然对这 8 个消息也可以用码长为 4 进行定长编码，这时只要从 16 个可能的码字中挑选 8 个作为真正的码字即可实现无错编码，比如下面就是一种编码方法。

信源字符	$a_1a_1a_1$	$a_1a_1a_2$	$a_1a_2a_1$	$a_2a_1a_1$	$a_2a_1a_2$	$a_1a_2a_2$	a_2
码字	0000	0001	0010	0100	1000	0011	1

例题 4.3.2：

考虑由 26 个英文字母与空格组成的信源，如果只对单个字符进行定长编码，码长必须 $k \geq \log_2 27 = 4.75$bits ，码长至少为 5bits, 也即平均每个字符需要占用 5 个二进位来编码，但每个英文字母平均携带的信息量 $H(X)$ 可能大大小于 5bits（参考例题 ??），所以用 5 个二进位去编码实际上会造成较大浪费，编码效率低下；最理想的情况是每个字符携带多少位的信息量就用多少个二进位去编码，这就是压缩编码问题。

定理 4.3.1：定长码编码定理

如果一种编码方案的误差可以任意小即 $P_e \leq \varepsilon$，则称这种编码方案为**无失真编码或渐近无错编码**。利用渐近等分性可以建立定长码无失真编码的存在性定理。

要对离散无记忆信源的 n 长消息进行 D 进定长编码，

（1） *对 $\varepsilon > 0$，如果码率 $R \geq H(X) + \varepsilon$，只要 n 足够大，就存在定长码使译码的错误 $P_e \leq \varepsilon$。*

（2） *对 $\delta > 0$，如果码率 $R \leq H(X) - \delta$，不管 n 有多么大，无论怎样进行定长编码，译码的错误 P_e 接近 1。*

证明：

不妨就二进定长编码进行证明。第（1）条的证明。
对 $\forall\varepsilon>0$，总存在充分大的 n 使 $W_\varepsilon^{(n)}$ 是一个非空的弱典型序列集。由渐近等分性定理 4.2.3，弱典型序列数 M 满足 (4-9)。
另外选择定长码码长 k 使 $R=k/n\geq H(X)+\varepsilon$，就有 $2^k\geq 2^{n[H(X)+\varepsilon]}$，即可选码字总数大于弱典型序列总数，从而可以对每个弱典型序列进行 k 长二进编码，对非弱典型序列不进行编码或者编一个统一的码字，当译码器收到非弱典型序列时就会有译码错误，但这种错误不会超过非弱典型序列出现的概率即 $P_e=P\left\{X^{(n)}\notin W_\varepsilon^{(n)}\right\}\leq\varepsilon$。

续证明：

第（2）证明。对任意的自然数 n，选择码长 k 使码率 $R = k/n \leq H(X) - \delta$，则码字总数 $2^k \leq 2^{n(H(X)-\delta)}$，对于任何 $\varepsilon : 0 < \varepsilon < \delta$，有 $2^k \leq 2^{n(H(X)-\delta)} < 2^{n(H(X)-\varepsilon)}$ 。不论怎样编码，最多也只能对 2^k 个消息编码（比如：既对 ε 弱典型序列编码，也对非弱典型序列编码），因此能正确译码的概率就是被编码消息序列的概率之和

$$\begin{aligned} 1 - P_e &= \text{被编码典型序列概率之和} + \text{被编码的非典型序列概率之和} \\ &\leq 2^{n(H(X)-\delta)} 2^{-n(H(X)-\varepsilon)} + P\left\{X^{(n)} \notin W_\varepsilon^{(n)}\right\} < 2^{-n(\delta-\varepsilon)} + \varepsilon \end{aligned}$$

当 n 较大时有 $2^{-n[\delta-\varepsilon]} < \varepsilon$，从而

$$P_e > 1 - \varepsilon - 2^{-n(\delta-\varepsilon)} > 1 - 2\varepsilon,$$

即译错概率几乎为 1。

几点说明：

（1） 这个定理说明：若码率超过信源的熵，则存在无失真编码使译码的错误充分小；但码率小于信源的熵时，不存在无失真编码。

（2） 实践上不可能使用充分大的分组长度，常常考虑在分组长度 n 不太大时，寻求能使码率 R 尽可能接近理论最小值 $H(X)$ 的编码方案。

（3） 这个定理说明了渐近等分性质在编码过程中的作用：只需要对弱典型序列进行编码。