

第五章 数据压缩

➤ **定义** 关于随机变量 X 的**信源编码** C 是从 X 的取值空间 \mathcal{X} 到 \mathcal{D}^* 的一个映射，其中 \mathcal{D}^* 表示 D 元字母表 \mathcal{D} 上有限长度的字符串所构成的集合。用 $C(x)$ 表示 x 的码字， $l(x)$ 表示 $C(x)$ 的长度。

➤ **定义** 设随机变量 X 的概率密度函数为 $p(x)$ ，定义信源编码 $C(x)$ 的**期望长度** $L(C)$ 为

$$L(C) = \sum_{x \in \mathcal{X}} p(x)l(x)$$

信源编码的例子

例

$\Pr(X = 1) = 1/2,$	codeword	$C(1) = 0$
$\Pr(X = 2) = 1/4,$	codeword	$C(2) = 10$
$\Pr(X = 3) = 1/8,$	codeword	$C(3) = 110$
$\Pr(X = 4) = 1/8,$	codeword	$C(4) = 111$

$$H(X) = \frac{7}{4} \text{ 比特}$$

$$L(C) = El(X) = \frac{7}{4}$$

信源编码的例子

➤ 例 中文电报的编码方式

- 中文电报的基本编码方法是将每一个汉字或字符用4位十进制数来表示，每一个十进制数再用5位二进制数来表示。
- 例如，“信息论”三个字的电码分别是(0207)，(1873)，(6158)。以“信”为例，首先将它编成4位十进制的码0207，再将它们变换成20位二进制的码:01101 11001 01101 11100，

$$2^{20} = 1048576$$

常用汉字表+次常用汉字表：大约是2500到7000之间
毛泽东所有的著作仅含3136个汉字。孙中山'三民主义'用了2134个汉字。'骆驼祥子'用了2413个汉字。

信源编码的例子

- **例** 摩尔斯电码：使用四个字符的字母表（点，划，字母间隔和单词间隔）

摩 尔 斯 电 码 表

字符	电码符号	字符	电码符号	字符	电码符号
A	• —	N	— •	1	• — — — —
B	— • • •	O	— — —	2	• • — — —
C	— • — •	P	• — — •	3	• • • — —
D	— • •	Q	— — • —	4	• • • • —
E	•	R	• — •	5	• • • • •
F	• • — •	S	• • •	6	— • • • •
G	— — •	T	—	7	— — • • •
H	• • • •	U	• • —	8	— — — • •
I	• •	V	• • • —	9	— — — — •
J	• — — —	W	• — —	0	— — — — —
K	— • —	X	— • • —	?	• • — — • •
L	• — • •	Y	— • — —	/	— • • — •
M	— —	Z	— — • •	()	— • — — • —
				—	— • • • • —
				•	• — • — • —

编码的类型

- 非奇异（nonsingular）编码：

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

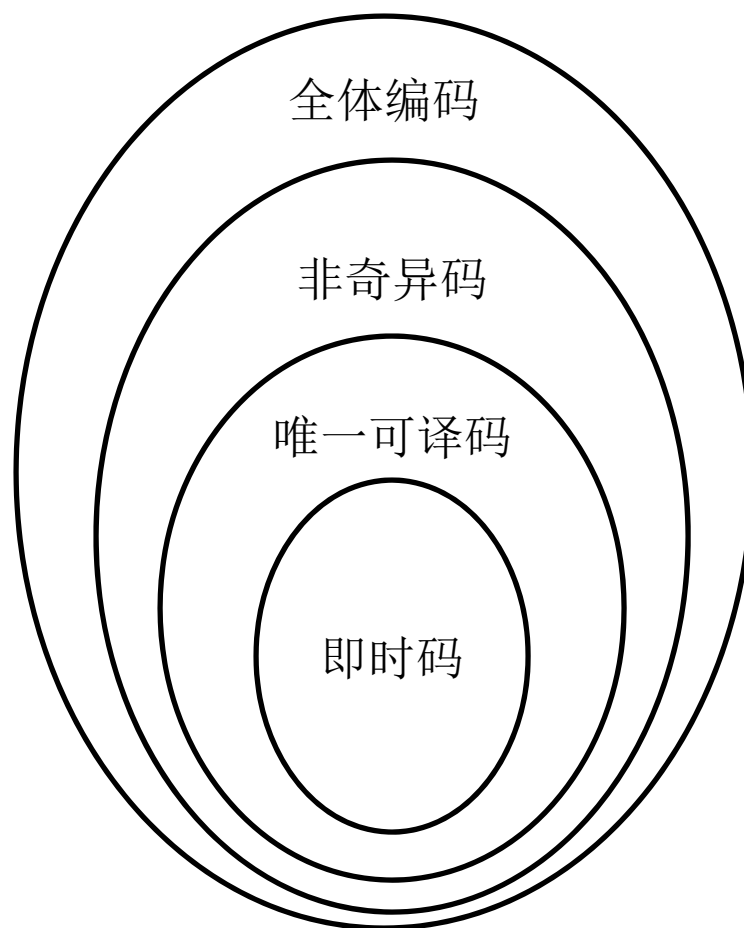
- 扩展（extension）编码：

$$C(x_1x_2 \cdots x_n) = C(x_1)C(x_2) \cdots C(x_n)$$

- 唯一可译（uniquely decodable）编码：扩展编码是非奇异的

- 前缀码（prefix code）或即时码（instantaneous code）：码中无任何码字是其他码字的前缀。

编码的类型



Kraft不等式

- 信源编码的目标：构造期望长度最小的即时码
- **Kraft不等式**：对于D元字母表上的即时码，码字长度 l_1, l_2, \dots, l_m 必须满足不等式

$$\sum_i D^{-l_i} \leq 1$$

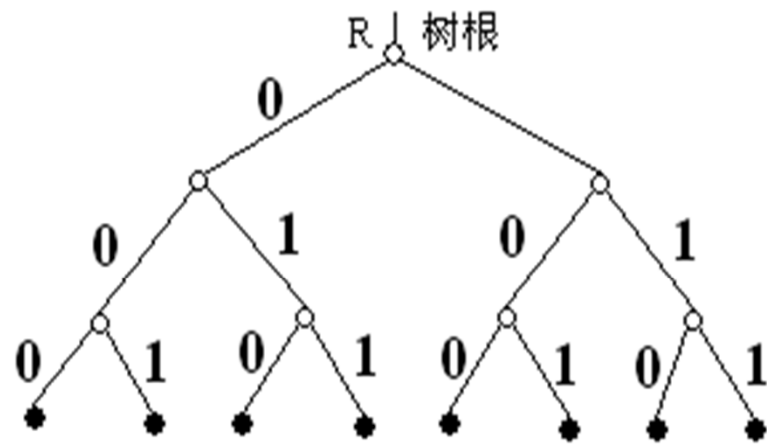
反之，若给定满足以上不等式的一组码字长度，则存在一个相应的即时码，其码字长度就是给定的长度。

- 推广的**Kraft不等式**：
$$\sum_{i=1}^{\infty} D^{-l_i} \leq 1$$

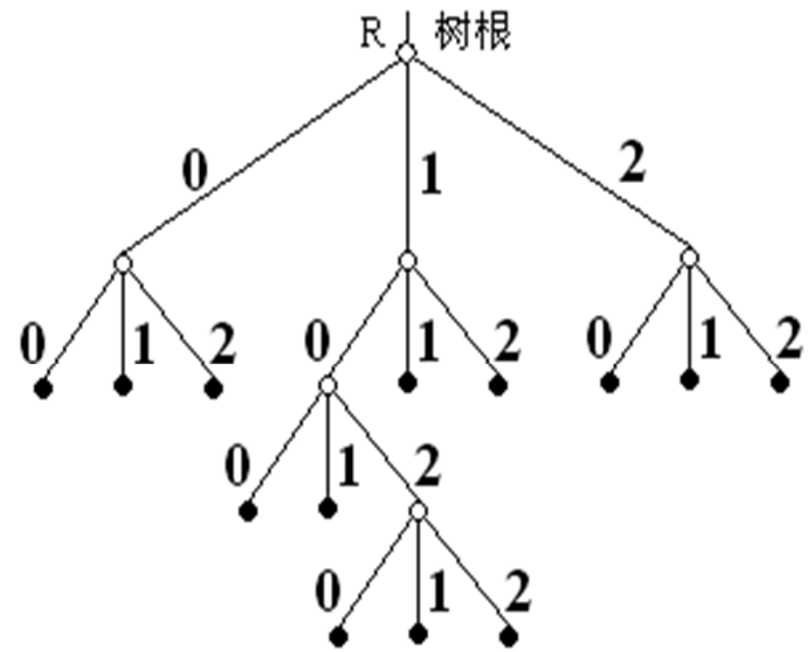
码树

- 对于给定码字的全体集合，可以用码树来描述。
 - 对于 r 进制的码树，如下页图所示，其中图(a)为二元码树，图(b)为三元码树。在码树中R点是树根，从树根伸出个树枝，构成 r 元码树。树枝的尽头是节点，一般中间节点会伸出树枝，不伸出树枝的节点为终端节点，编码时应尽量在终端节点安排码字。
-

码树

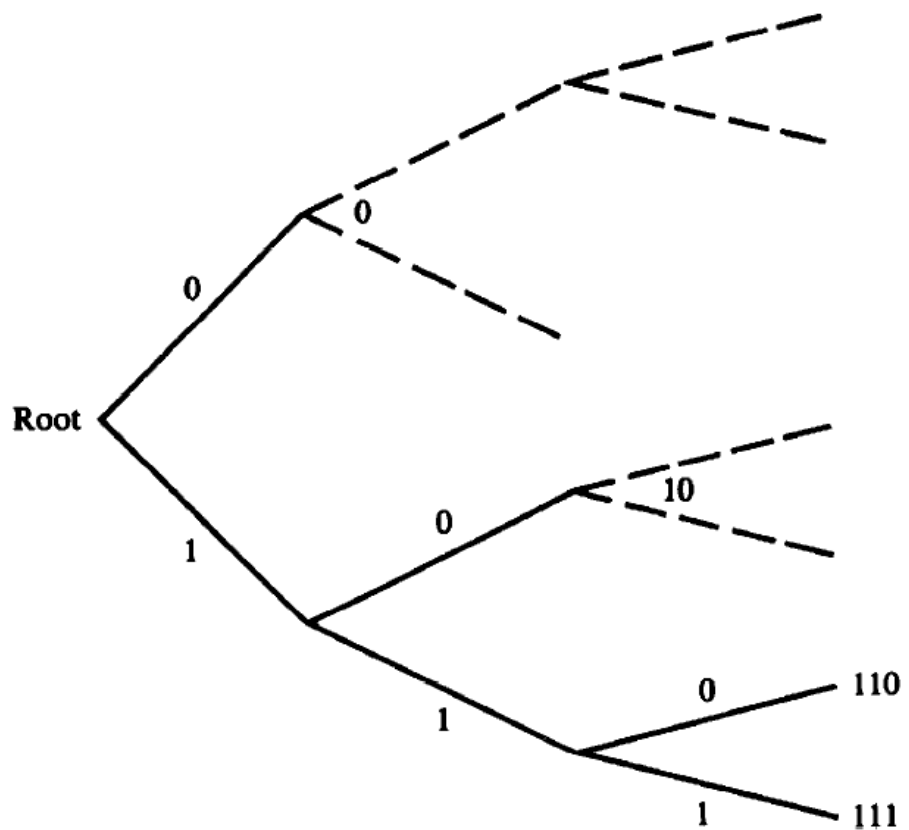


(a) 二进制码树



(b) 三元码树

Kraft不等式的证明



Kraft不等式的充分必要性

- Kraft不等式是即时码**存在**的充要条件，其必要性表现在如果码是即时码，则必定满足Kraft不等式；充分性表现在如果满足Kraft不等式，则这种码长的即时码一定**存在**，但并不表示所有满足Kraft不等式的码一定是即时码。
 - 因此，Kraft不等式是即时码**存在**的充要条件，而不是即时码的充要条件。
-

唯一可译码的Kraft不等式

- 对于D元字母表上的**唯一可译码**，码字长度 l_1, l_2, \dots, l_m 必须满足**Kraft不等式**

$$\sum_i D^{-l_i} \leq 1$$

反之，若给定满足以上不等式的一组码字长度，则可以构造出具有同样码字长度的**唯一可译码**。

最优码

➤ 最优化问题：在所有满足 $\sum_{i=1}^{\infty} D^{-l_i} \leq 1$ 整数 l_1, l_2, \dots, l_m 中，最小化 $L = \sum p_i l_i$

- ✓ 取消 l_i 必须是整数的限制
- ✓ 约束条件中的等号成立
- ✓ 拉格朗日乘子法（Lagrange Multiplier）

➤ **定理** 随机变量 X 的任一 D 元即时码的期望长度

$$L \geq H_D(X)$$

当且仅当 $D^{-l_i} = p_i$ ，等号成立

寻找最优码

- **定义** **D**进制的 (D-adic) 概率分布：每一个概率值均等于 D^{-n}
- 寻找最优码的方法
 - ✓ 找到与X的分布最接近的D进制分布（在相对熵意义下）
 - ✓ 该D进制概率分布可提供一组码字长度
 - ✓ 构造码字

最优码长的界

- **定理** 设 $l_1^*, l_2^*, \dots, l_m^*$ 是关于信源分布 \mathbf{p} 和一个 \mathbf{D} 元字母表的一组最优码长, L^* 为最优码的相应期望长度, 则

$$H_D(X) \leq L^* < H_D(X) + 1$$

- 多字符分组编码:

$$L_n = \frac{1}{n} \sum p(x_1, x_2, \dots, x_n) l(x_1, x_2, \dots, x_n) = \frac{1}{n} El(X_1, X_2, \dots, X_n)$$

$$\frac{H_D(X_1, X_2, \dots, X_n)}{n} \leq L_n < \frac{H_D(X_1, X_2, \dots, X_n)}{n} + \frac{1}{n}$$

$$X_1, X_2, \dots, X_n \text{ 独立同分布: } H_D(X) \leq L_n < H_D(X) + \frac{1}{n}$$

香农第一定理

➤ **定理** 香农第一定理（可变长无失真信源编码定理）

设 X_1, X_2, \dots, X_n 为离散无记忆信源 X 的 n 次扩展，对 n 次扩展信源进行编码，平均每字符期望码长为 L_n ，则对任意给定的 $\varepsilon > 0$ ，当 n 足够大时，总可以找到一种无失真惟一可译编码，满足

$$H_D(X) \leq L_n < H_D(X) + \varepsilon$$

反之，若 $L_n < H_D(X)$ ，则信源编码不可能无失真。

不正确分布引起的误差

➤ **定理** 码字长度分配 $l(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil$ 关于 $p(x)$ 的期望码长满足

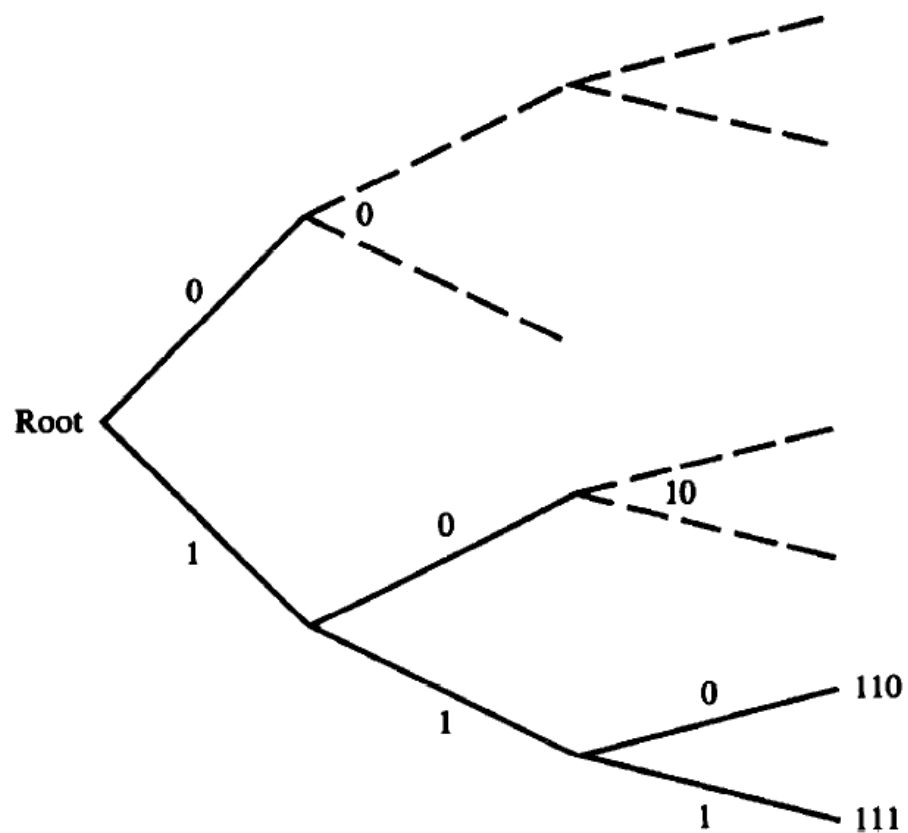
$$H(p) + D(p\|q) \leq E_p l(X) < H(p) + D(p\|q) + 1$$

例

$$\left[\begin{array}{c} X \\ p(X) \end{array} \right] = \left\{ \begin{array}{cccc} 1 & 2 & 3 & 4 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{array} \right\} \quad L_p(X) = 1.75 \text{ 比特}; L_q(X) = 2 \text{ 比特}$$

$$\left[\begin{array}{c} X \\ q(Y) \end{array} \right] = \left\{ \begin{array}{cccc} 1 & 2 & 3 & 4 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{array} \right\} \quad D(p\|q) = \sum p(x) \log \frac{p(x)}{q(x)} = 0.25 \text{ 比特}$$

构造最优即时码



$$l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil$$

赫夫曼码

例

$$D = 2$$

Codeword length	Codeword	X	Probability
2	01	1	0.25
2	10	2	0.25
2	11	3	0.2
3	000	4	0.15
3	001	5	0.15

$$L = 2.3 \text{ 比特}$$

赫夫曼码

例 $D = 3$

码字	X	概率	概率	概率
12	1	0.25	0.3	0.7
11	2	0.25	0.25	0.3
10	3	0.2	0.25	
02	4	0.1	0.2	
01	5	0.1		
00	6	0.1		

$L = 2$ 铁特 (ternary digit)

赫夫曼码

例

$D = 3$

$$|\mathcal{X}| = 1 + k(D - 1)$$

码字	X	概率	概率	概率
12	1	0.25	0.3	0.7
11	2	0.25	0.25	0.3
10	3	0.2	0.25	
02	4	0.1	0.2	
01	5	0.1		
00	6	0.1		

$L = 2$ 铁特 (ternary digit)

赫夫曼码

例

$$D = 3$$

$$|\mathcal{X}| = 1 + k(D - 1)$$

Codeword	X	Probability
1	1	0.25
2	2	0.25
01	3	0.2
02	4	0.1
000	5	0.1
001	6	0.1
002	Dummy	0.0

$$L = 1.7 \text{ 比特 (ternary digit)}$$

赫夫曼码的讨论

- 加权码字的赫夫曼编码：赫夫曼算法最小化的是码长加权和 $\sum p_i l_i$

<i>X</i>	Codeword	Weights
1	00	5
2	01	5
3	10	4
4	11	4

The diagram illustrates the merging process of weights. It shows four initial weights: 5, 5, 4, and 4. The two 5s are merged into 10, and the two 4s are merged into 8. Then, 10 and 8 are merged into 18. Lines connect the initial weights to their intermediate and final merged values.

赫夫曼码的讨论

- 赫夫曼编码和香农码（码字长度 $l(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil$ ）
 - ✓ 从平均意义上讲，赫夫曼编码具有更短的期望长度，但两者的差别不超过**1**比特
 - ✓ 对于单个字符来说，香农码可能具有比赫夫曼码更短的码字长度

码字长度	码字	x	概率
1	1	1	1/3
2	01	2	1/3
3	001	3	1/4
3	000	4	1/12

$$L_{Huffman} = 2 \text{ 比特}$$

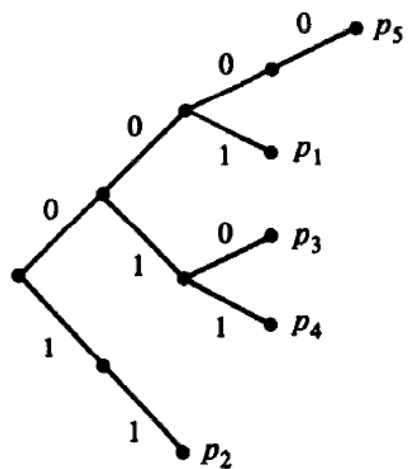
$$L_{Shannon} = 2.17 \text{ 比特}$$

赫夫曼码的最优性

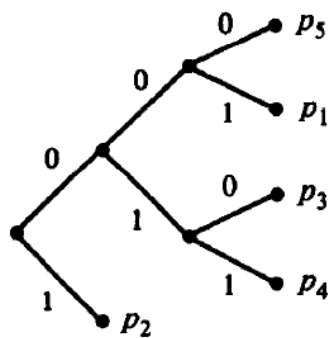
- **定理** 对任意一个分布，必然存在满足如下性质的一个最优即时码：
 1. 其长度序列与按概率分布排列的次序相反
 2. 最长的两个码字具有相同长度
 3. 最长的两个码字仅在最后一位上有所差别，且对应与两个最小可能发生的字符
- 满足以上定理得最优码称为**典则码**
- **定理** 赫夫曼码是最优的，它提供了最短的期望码长

$$L(C^*) \leq L(C')$$

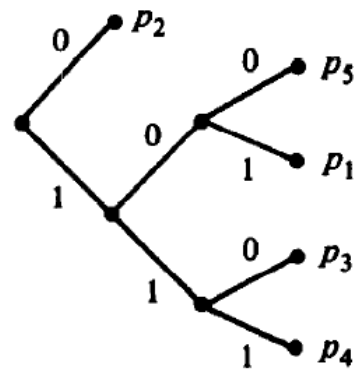
赫夫曼码的最优性



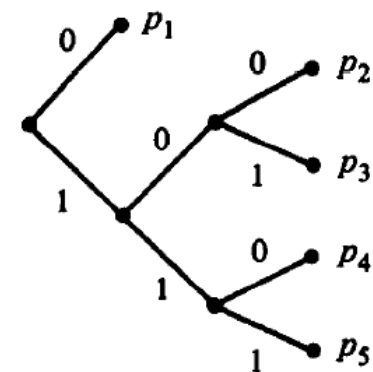
(a)



(b)

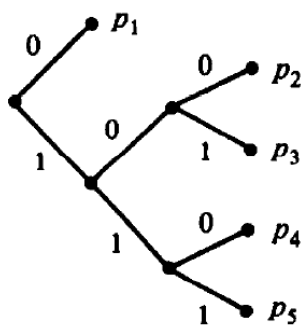


(c)

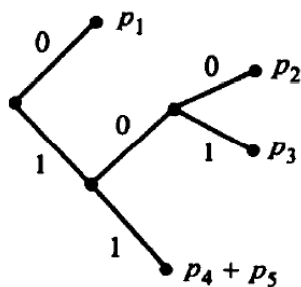


(d)

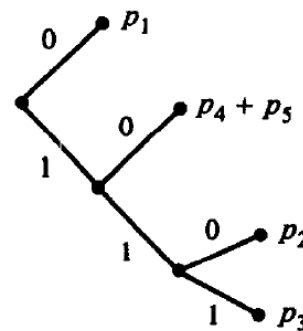
赫夫曼码的最优性



(a)



(b)



(c)

	$C_{m-1}^*(\mathbf{p}')$		$C_m(\mathbf{p})$	
p_1	w'_1	l'_1	$w_1 = w'_1$	$l_1 = l'_1$
p_2	w'_2	l'_2	$w_2 = w'_2$	$l_2 = l'_2$
\vdots	\vdots	\vdots	\vdots	\vdots
p_{m-2}	w'_{m-2}	l'_{m-2}	$w_{m-2} = w'_{m-2}$	$l_{m-2} = l'_{m-2}$
$p_{m-1} + p_m$	w'_{m-1}	l'_{m-1}	$w_{m-1} = w'_{m-1}0$	$l_{m-1} = l'_{m-1} + 1$
			$w_m = w'_{m-1}1$	$l_m = l'_{m-1} + 1$

Shannon-Fano-Elias编码

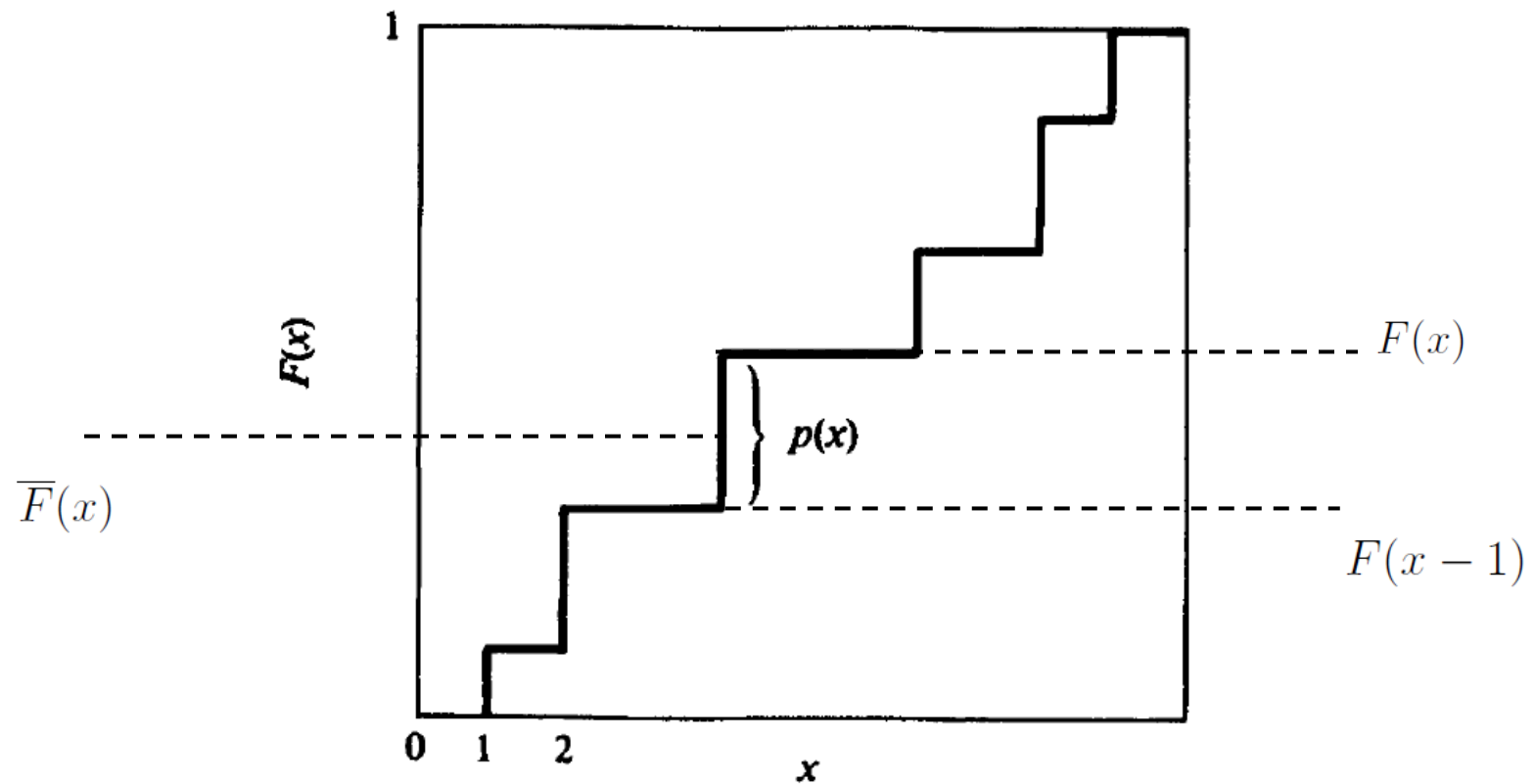
- 累计分布函数：假定取 $\mathcal{X} = \{1, 2, \dots, m\}$ ，并对所有 x ，有 $p(x) > 0$ 。定义累计分布函数

$$F(x) = \sum_{a \leq x} p(a)$$

- 修正的累计分布函数

$$\bar{F}(x) = \sum_{a < x} p(a) + \frac{1}{2}p(x)$$

Shannon-Fano-Elias 编码



Shannon-Fano-Elias 编码

- $\bar{F}(x)$ 唯一确定 x , 可作为 x 的编码
- 一般情况下, $\bar{F}(x)$ 需用无限多比特表示
- 取 $\bar{F}(x)$ 的前 $l(x)$ 位作为 x 的编码: $[\bar{F}(x)]_{l(x)}$

$$l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil + 1$$

- 此编码是前缀码
- 编码的期望长度:

$$L < H(X) + 2$$

Shannon-Fano-Elias编码的例子

x	$p(x)$	$F(x)$	$\bar{F}(x)$	$\bar{F}(x)$ in binary	$l(x) = \lceil \log \frac{1}{p(x)} \rceil + 1$	Codeword
1	0.25	0.25	0.125	0.001	3	001
2	0.5	0.75	0.5	0.10	2	10
3	0.125	0.875	0.8125	0.1101	4	1101
4	0.125	1.0	0.9375	0.1111	4	1111

$$L = 2.75 \text{ 比特}, \quad H(X) = 1.75 \text{ 比特}$$

x	$p(x)$	$F(x)$	$\bar{F}(x)$	$\bar{F}(x)$ in binary	$l(x) = \left\lceil \log \frac{1}{p(x)} \right\rceil + 1$	Codeword
1	0.25	0.25	0.125	0.001	3	001
2	0.25	0.5	0.375	0.011	3	011
3	0.2	0.7	0.6	0.10011	4	1001
4	0.15	0.85	0.775	0.1100011	4	1100
5	0.15	1.0	0.925	0.1110110	4	1110