信息论

杨勐 副教授、博导
人工智能学院
http://gr.xjtu.edu.cn/web/mengyang

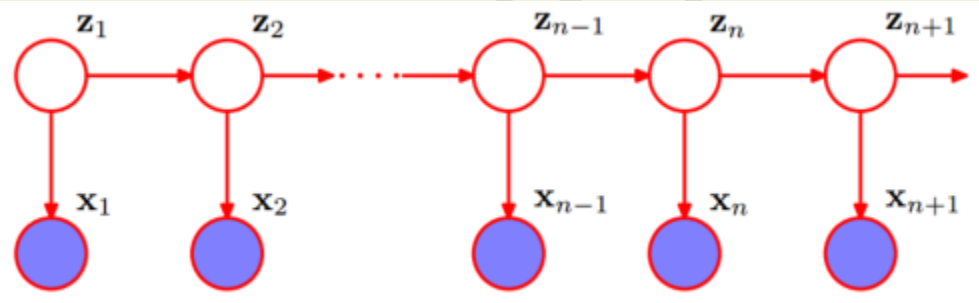
IAIR2023

为什么学《信息论》？

- **熵**——度量物理世界不确定性的基本概念
- **深度学习：**交叉熵——度量概率分布p(x)和q(x)间的距离

$$H(p,q) = -\sum_{x} p(x)\log q(x)$$

- **自然语言处理：**基于卷积码的维特比算法——一种广泛应用于机器学习的动态规划算法，比如求解HMM的隐状态



- **脑认知：**可能是未来若干年的研究热点，比如脑信号编码-解码

课程内容

- 熵、相对熵、互信息
- 渐进均分性
- 随机过程的熵率
- 数据压缩
- 信道容量
- 微分熵
- 高斯信道
- 率失真理论
- 博弈与数据压缩
- 信息论与统计学

IAIR2023

信息论（后半部分）的教学安排

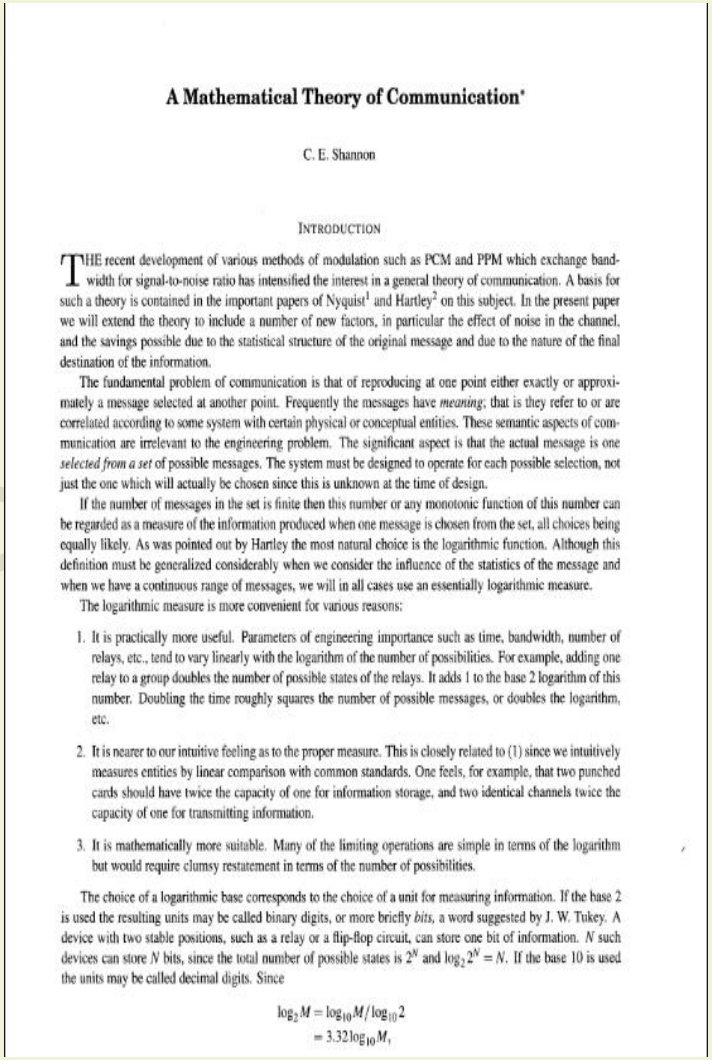
- **课时** 16学时课堂（含习题课、复习）
- **教材** 《信息论基础》（Thomas Cover著）
- **答疑** 周五下午2:30-5:30
- **联系** 手机、微信：13468675342

  邮箱：mengyang@mail.xjtu.edu.cn
- **网站** http://gr.xjtu.edu.cn/web/mengyang/7（课件、作业及参考答案等）

  (减少课程内容；重复回顾知识点；课后看书)
- **考核构成** **期末考试70%+作业20%+考勤10%**

IAIR2023

香农，1916年—2001年



A Mathematical Theory of Communication*

C. E. Shannon

INTRODUCTION

THE recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal-to-noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities. For example, adding one relay to a group doubles the number of possible states of the relays. It adds 1 to the base 2 logarithm of this number. Doubling the time roughly squares the number of possible messages, or doubles the logarithm, etc.

2. It is nearer to our intuitive feeling as to the proper measure. This is closely related to (1) since we intuitively measures entities by linear comparison with common standards. One feels, for example, that two punched cards should have twice the capacity of one for information storage, and two identical channels twice the capacity of one for transmitting information.

3. It is mathematically more suitable. Many of the limiting operations are simple in terms of the logarithm but would require clumsy restatement in terms of the number of possibilities.

The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly *bits*, a word suggested by J. W. Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of information. N such devices can store N bits, since the total number of possible states is 2^N and $\log_2 2^N = N$. If the base 10 is used the units may be called decimal digits. Since

$$\log_2 M = \log_{10} M / \log_{10} 2 = 3.32 \log_{10} M,$$

《通讯的数学原理》1948

第5章 数据压缩

IAIR2023

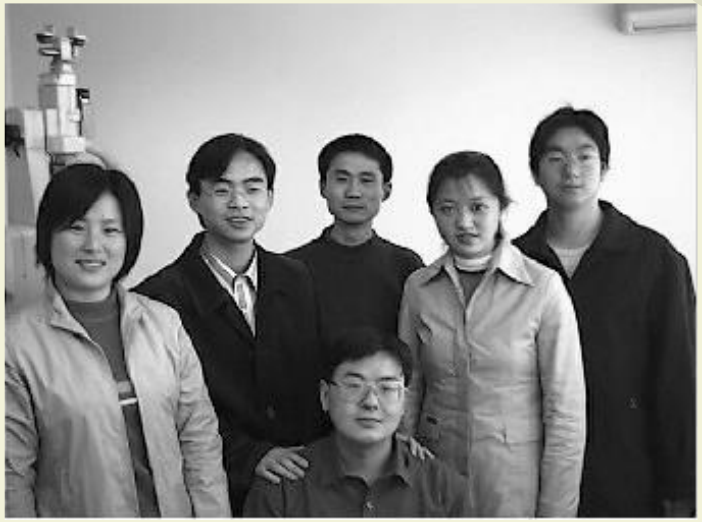
杨勐 主讲

什么是数据压缩?

- **举例：2D图像压缩**
- **目标：**减少数据量、提高数据存储、传输、处理效率
- **分类：**无损压缩（信源编码，消除冗余数据）——**熵**

  有损压缩（信道编码，降低数据质量）——**率失真理论**
- 其它：PCA降维、傅里叶变换、DCT变换等

本章数据压缩特指无损压缩



原图

失真图（PSNR=23.73dB）

什么是数据压缩?

- **定义：**关于随机变量X的信源编码C是从X的取值空间χ到D^*的一个映射，其中D^*表示D元字母表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，I(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

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

信源编码的例子

- 设随机变量X的分布及其码字分配如下

$$\Pr(X=1)=1/2,\ \text{codeword}\quad C(1)=0$$ 二进制

$$\Pr(X=2)=1/4,\ \text{codeword}\quad C(2)=10$$

$$\Pr(X=3)=1/8,\ \text{codeword}\quad C(3)=110$$

$$\Pr(X=4)=1/8,\ \text{codeword}\quad C(4)=111$$

$$H(X)=\frac{7}{4}\text{比特}\qquad L(C)=El(X)=\frac{7}{4}$$

IAIR 2023

信源编码的例子

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。

$$2^{20} = 1048576$$

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位十进制码0207，再将它们变换成20位二进制的码：01101 11001 01101 11100。

常用汉字表+次常用汉字表：大约2500到7000之间，毛泽东所有的著作仅包含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

- **摩斯电码**：使用两种字符的字母表（点“滴”、划“答”）
  - 根据字符使用概率确定电码长度
  - 如何保证电码不产生歧义？

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：

$$\boxed{x \neq x' \Rightarrow C(x) \neq C(x')}$$

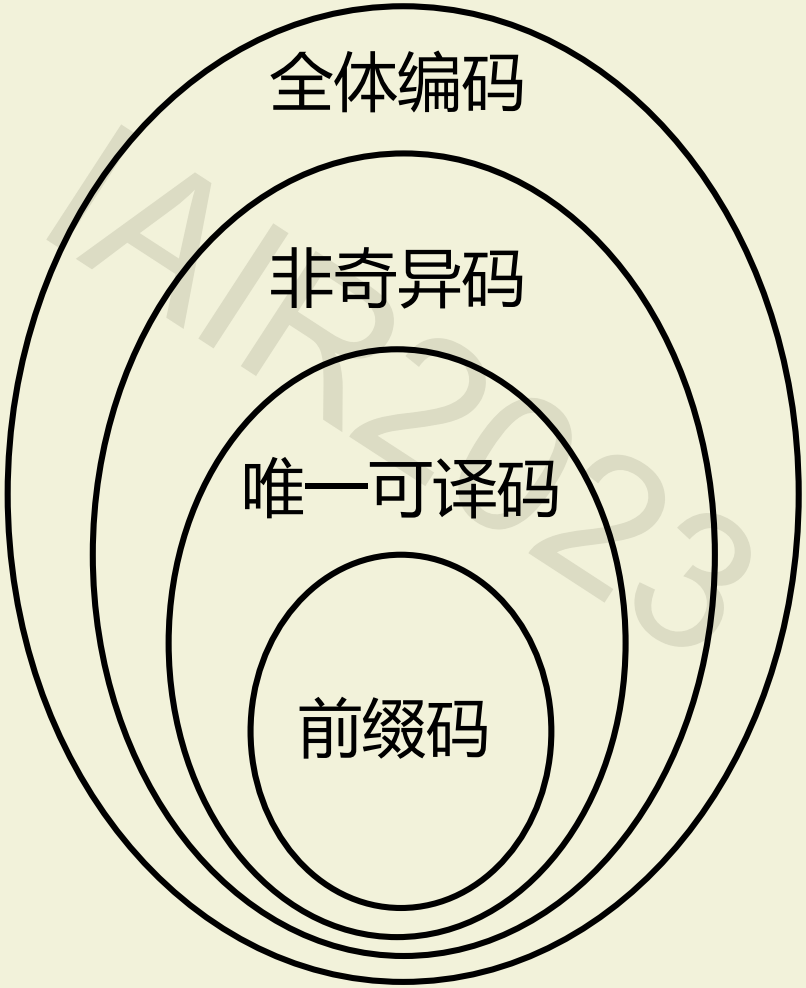
- 扩展（extension)编码：

$$\boxed{C(x_1 x_2 \cdots x_n) = C(x_1) C(x_2) \cdots C(x_n)}$$

- **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

IAIR-2023

编码的类型



IAIR2023

编码的类型

表 5-1　码的几种类型

X	奇异的	非奇异，但不是惟一可译的	惟一可译，但不是即时的	即时的
1	0	0	10	0
2	0	010	00	10
3	0	01	11	110
4	0	10	110	111

非奇异码010可能对应三种信源：2、14、31

Kraft不等式

- 信源编码的目标：构造**期望长度最小的前缀码**
- **Kraft不等式**：对于D元（进制）字母表上的即时码，码字长度 $l_1, l_2, \cdots, l_m$必须满足不等式

$$\boxed{\sum_i D^{-l_i} \le 1}$$

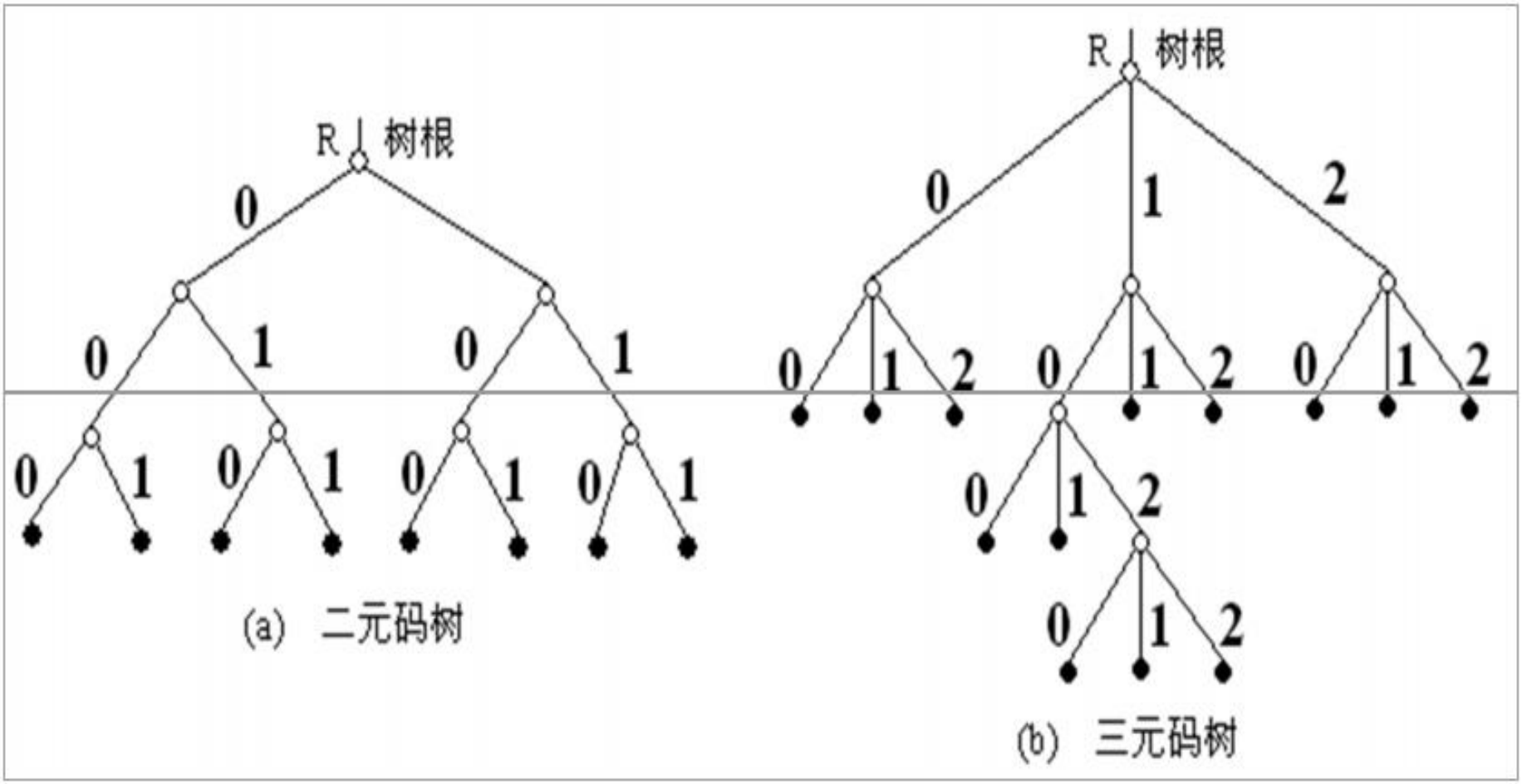
- 反之，若给定满足以上不等式的一组码字长度，则存在一个相应的前缀码，其码字长度就是给定的长度。
- 推广的Kraft不等式： $\sum_{i=1}^{\infty} D^{-l_i} \le 1$

IAIR2023

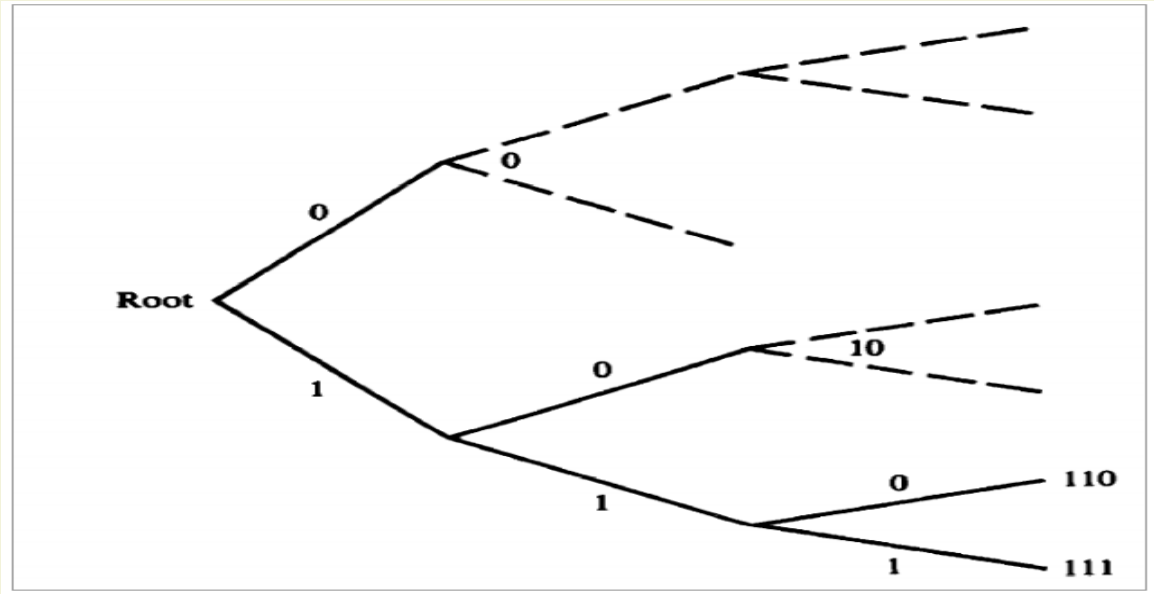
码树

- 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



Kraft不等式的证明



令$l_{\max}$为码字集最长码字的长度，对一个在l_i层的码字，其在$l_{\max}$层有$D^{l_{\max}-l_i}$个后代，所有这些后代集合都不相交；$l_{\max}$层共有$D^{l_{\max}}$个节点。对所有码字求和，即得

$$\boxed{\sum D^{l_{\max}-l_i} \le D^{l_{\max}} \quad \Rightarrow \quad \sum D^{-l_i} \le 1}$$

Kraft不等式的充分必要性（了解）

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

惟一可译码的Kraft不等式（了解）

- 对于D元字母表上的惟一可译码，码字长度$l_1, l_2, \cdots, l_m$必须满足Kraft不等式

$$\boxed{\sum_i D^{-l_i} \le 1}$$

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惟一可译码。

IAIR2023

寻找最优码

- **最优化问题**：在所有满足约束条件$\sum_{i=1}^{\infty} D^{-l_i} \leq 1$的整数解$l_1, l_2, \cdots, l_m$中，最小化期望长度 $L = \sum p_i l_i$
  - 取消 l_i 必须是整数的限制
  - 约束条件中的等号成立
  - 采用拉格朗日乘子法（Lagrange Multiplier）（推导）
- **定理**：随机变量X的任一D元前缀码的期望长度

$$\boxed{L \geq H_D(X)}$$

当且仅当$D^{-l_i} = p_i$时等号成立。

（求解的l_i可能是非整数）

寻找最优码

- **定义：**对于某个n值，如果码字概率分布的每一个概率值均等于D^{-n}，则称这个概率分布是D进制的
- 寻找最优码的方法
  - 找到与X的分布最接近的D进制分布
  - 该D进制概率分布可提供一组码字长度
  - 构造码字

不易求解

IAIR2025

最优码长的界（了解）

- **定理**：设 $l_1^*, l_2^*, \cdots, l_m^*$ 是关于信源分布p和一个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

$$\boxed{H_D(X) \le L^* < H_D(X) + 1}$$

- 多字符分组编码：

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

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

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

IAIR2023

香农第一定理

■ 香农第一定理（可变长无失真信源编码定理）

设$X_1, X_2, \cdots, X_n$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

$$\boxed{H_D(X) \le L_n < H_D(X) + \varepsilon}$$

反之，若$L_n < H_D(X)$，则信源编码不可能无失真。

IAIR2023

不正确分布引起的误差（了解）

- 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

$$\boxed{H(p)+D(p\|q)\le E_p l(X)<H(p)+D(p\|q)+1}$$

（证明见P66）

举例：

$$\begin{bmatrix} X \\ p(X) \end{bmatrix}=\begin{Bmatrix} 1 & 2 & 3 & 4 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{8} & \frac{1}{8} \end{Bmatrix}$$

$$\begin{bmatrix} X \\ q(Y) \end{bmatrix}=\begin{Bmatrix} 1 & 2 & 3 & 4 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{Bmatrix}$$

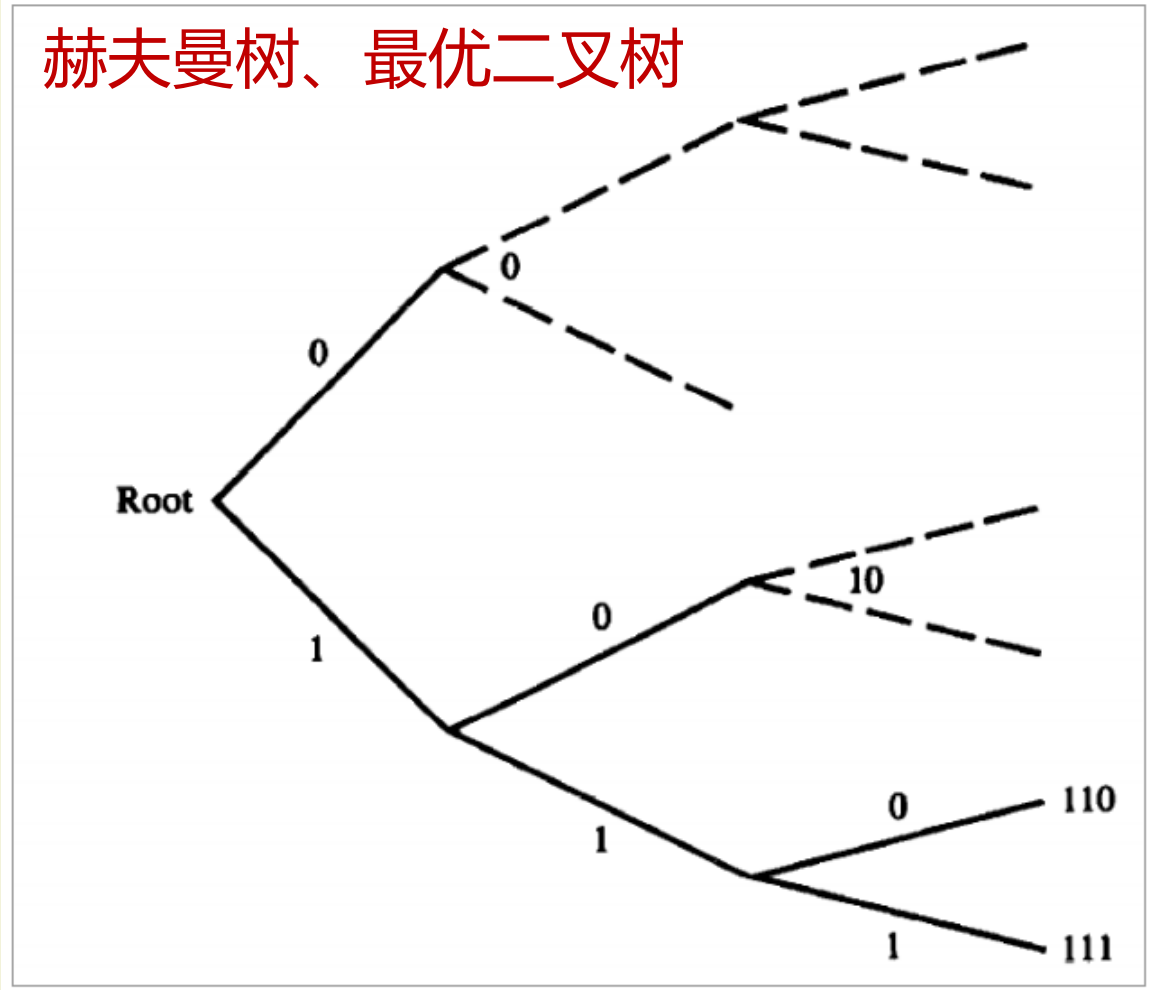
$L_p(X)=1.75$比特；$L_q(X)=2$比特

$$D(p\|q)=\sum p(x)\log\frac{p(x)}{q(x)}=0.25\text{比特}$$

根据不正确分布q设计的码字，期望码长增加$D(p\|q)$

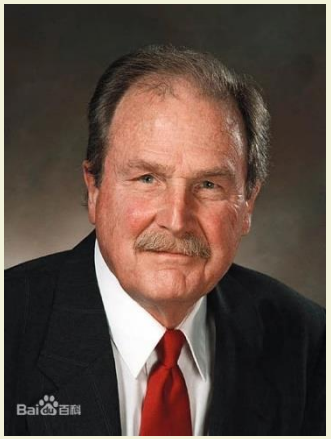
IAIR2023

构造最优前缀码（最短期望长度码）



如何设计长度为$l(x)$的码字?

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



David Albert Huffman
(1925~1999)

赫夫曼码

- **举例：** 随机变量 X={1, 2, 3, 4, 5}，概率为0.25, 0.25, 0.2, 0.15, 0.15， D=2

Codeword length	Codeword	X	Probability				
2	01	1	0.25	0.3	0.45	0.55	1
2	10	2	0.25	0.25	0.3	0.45	
2	11	3	0.2	0.25	0.25		
3	000	4	0.15	0.2			
3	001	5	0.15				

$$\boxed{\mathrm{H} = 1.5842, \quad L = 2.3bit}$$

赫夫曼码

- Four symbols: “a”,”b”,”c”,”d”
- pmf:

$$\mathbf{p}^T = [0.5000, 0.2143, 0.1703, 0.1154]$$

- 1st order conditional pmf:

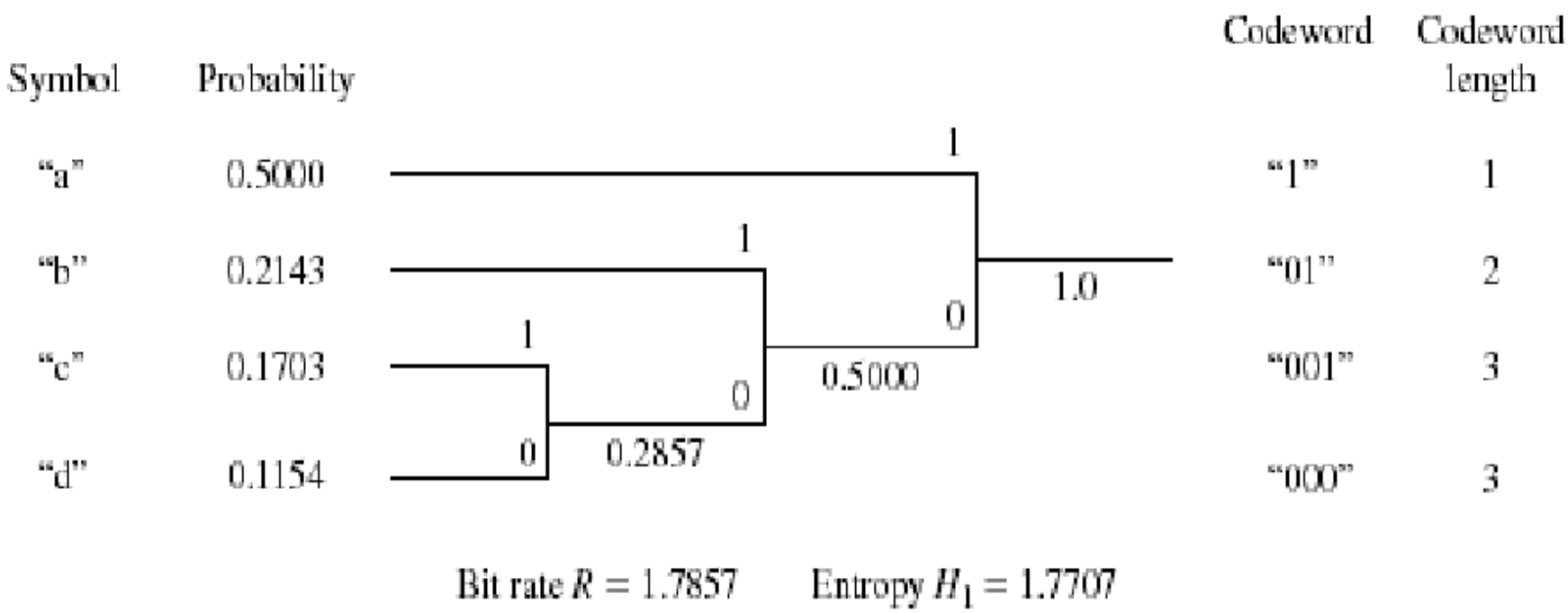
$$[\mathbf{Q}] = \begin{bmatrix} 0.6250 & 0.3750 & 0.3750 & 0.3750 \\ 0.1875 & 0.3125 & 0.1875 & 0.1875 \\ 0.1250 & 0.1875 & 0.3125 & 0.1250 \\ 0.0625 & 0.1250 & 0.1250 & 0.3125 \end{bmatrix}$$

- 2nd order pmf:

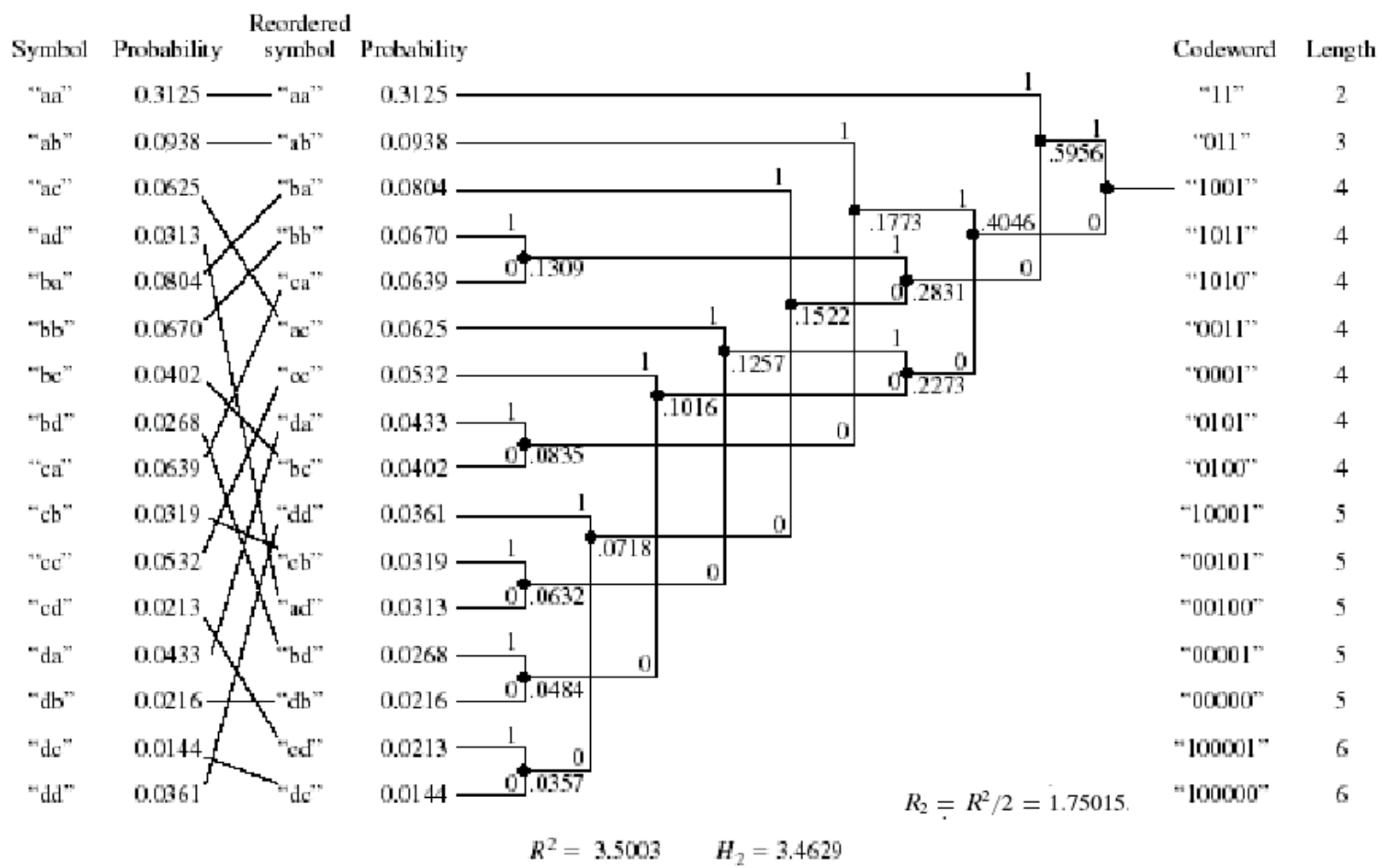
$$p(f_{n-1}, f_n) = p(f_{n-1}) q(f_n \mid f_{n-1}).$$

Ex. $p("ab") = p("a") q("b"/"a") = 0.5 * 0.1875 = 0.0938$

赫夫曼码


Symbol
Probability
Codeword
Codeword length
"a"
0.5000
"b"
0.2143
"c"
0.1703
"d"
0.1154
1
1
1
0
0
0
1.0
0.5000
0.2857
"1"
"01"
"001"
"000"
1
2
3
3
Bit rate $R = 1.7857$
Entropy $H_1 = 1.7707$

赫夫曼码


Symbol
Probability
Reordered symbol
Probability
Codeword
Length
"aa" 0.3125 "aa" 0.3125 "11" 2
"ab" 0.0938 "ab" 0.0938 "011" 3
"ac" 0.0625 "ba" 0.0804 "1001" 4
"ad" 0.0313 "bb" 0.0670 "1011" 4
"ba" 0.0804 "ca" 0.0639 "1010" 4
"bb" 0.0670 "ac" 0.0625 "0011" 4
"bc" 0.0402 "cc" 0.0532 "0001" 4
"bd" 0.0268 "da" 0.0433 "0101" 4
"ca" 0.0639 "bc" 0.0402 "0100" 4
"cb" 0.0319 "dd" 0.0361 "10001" 5
"cc" 0.0532 "cb" 0.0319 "00101" 5
"cd" 0.0213 "ad" 0.0313 "00100" 5
"da" 0.0433 "bd" 0.0268 "00001" 5
"db" 0.0216 "db" 0.0216 "00000" 5
"dc" 0.0144 "cd" 0.0213 "100001" 6
"dd" 0.0361 "dc" 0.0144 "100000" 6
.1309
.1773
.4046
.5956
.2831
.1522
.1257
.2273
.1016
.0835
.0718
.0632
.0484
.0357
$R_2 = R^2/2 = 1.75015.$
$R^2 = 3.5003$
$H_2 = 3.4629$

赫夫曼码

- **举例：** 随机变量 X={1, 2, 3, 4, 5}，概率为0.25, 0.25, 0.2, 0.15, 0.15， D=3 （如不足D=3个字符，可用空字符代替）

码字	X	概率			
12	1	0.25	0.3	0.7	1
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			

$$L = 2\ ternarydigit$$

赫夫曼码的讨论

- 赫夫曼编码和香农码（码字长度 $l(x)=\left\lceil \log \frac{1}{q(x)} \right\rceil$）

- 对单个字符来说，香农码可能具有比赫夫曼码更短的码长

  举例：2个字符，概率为0.9999和0.00001，香农码码长分别为1和14比特

- 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

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

$L_{Huffman} = 2$比特

$L_{Shannon} = 2.17$比特

IAIR2023

赫夫曼码的最优性——典则码

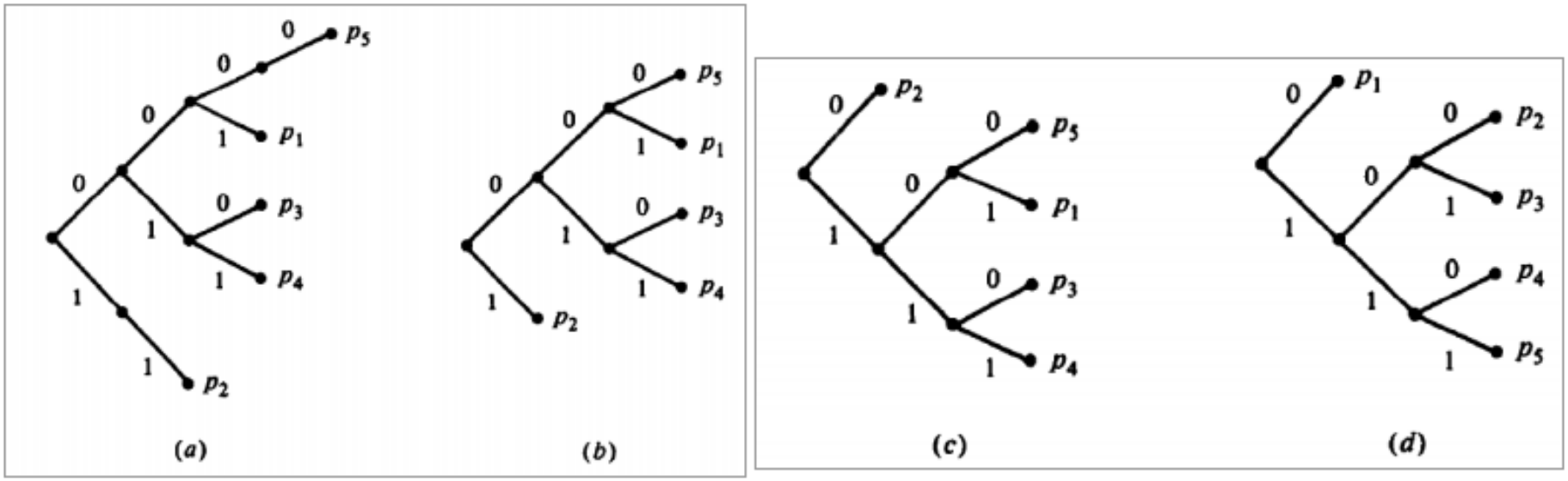
- 对任意一个分布，必然存在满足如下性质的一个最优前缀码（最小期望长度）：

  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（canonical code）

赫夫曼码的最优性——典则码

- 已知分布$p_1 \geq p_2 \geq \cdots \geq p_m$，给定任一前缀码，可通过修剪、交换、重排等操作得到典则码



(a)前缀码　(b)修剪无兄弟的分支　(c)按码长重排　(b)交换概率

赫夫曼码的最优性

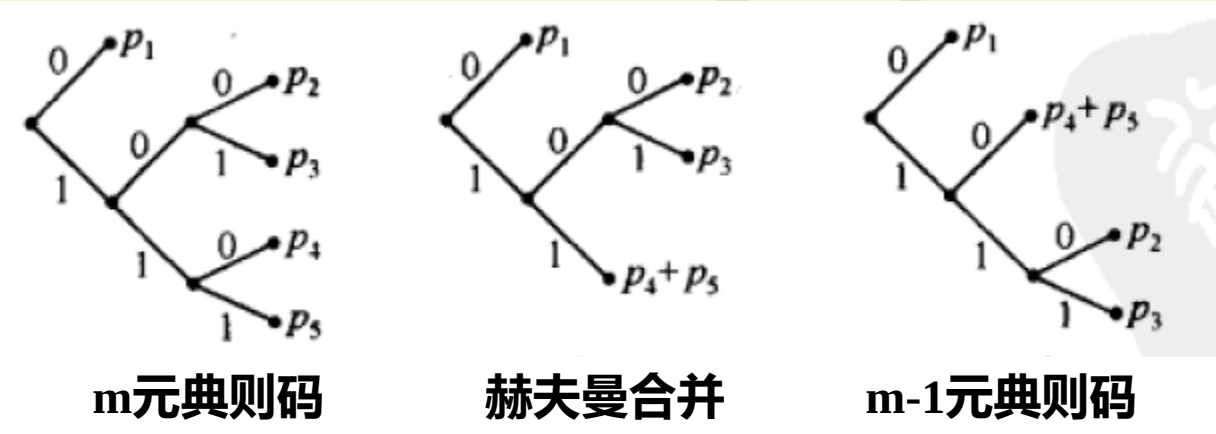
- **定理**：赫夫曼码是最优的，它提供了最短的期望码长

$$\boxed{L\left(C^*\right) \le L\left(C'\right)}$$

归纳法证明（了解）：

m维概率密度函数 $p = (p_1, p_2, \ldots, p_m),\ p_1 \ge p_2 \ge \cdots \ge p_m$

m-1维赫夫曼合并 $p' = (p_1, p_2, \ldots, p_{m-2}, p_{m-1} + p_m)$



赫夫曼码的最优性

- 基于m-1维概率密度函数p'的最优码C_{m-1}^*，取权重为$p_{m-1}+p_m$的码字，在尾部加0或1，构造m元扩展码，由码长计算$\sum p_i l_i$得到

$$\boxed{L(p)=L^*(p')+p_{m-1}+p_m}$$

- 基于m维概率密度函数p的最优码C_m^*，将两个最小概率p_{m-1}和p_m对应的码字（兄弟节点）合并，构造m-1元概率密度函数p'的码字，可得码长

$$\boxed{L(p')=L^*(p)-p_{m-1}-p_m}$$

IAIR 2026

赫夫曼码的最优性

- 相加即得

$$\boxed{[L(p') - L^*(p')] + [L(p) - L^*(p)] = 0}$$

- 因为$L(p') \geq L^*(p')$且$L(p) \geq L^*(p)$，（最优码的平均码字最短），得到

$$\boxed{L(p) = L^*(p)}$$

- 因此，从m-1维概率密度函数p'的最优码出发，扩展得到的码字，是关于m维概率密度函数p最优的。

IAIR2023

本章小结

- 数据压缩与信息论的关系
- Kraft不等式与最优码
- 香农第一定理
- 赫夫曼码的构造
- 赫夫曼码的最优性

IAIR2023

第5章作业

■ 题目

5.1、5.3、5.4、5.5、5.6、5.12、5.14、5.15、5.17、5.18、5.33、5.38

■ 要求

提交方式： 纸质版作业

提交日期： 2022年5月25日（周四）

IAIR2023