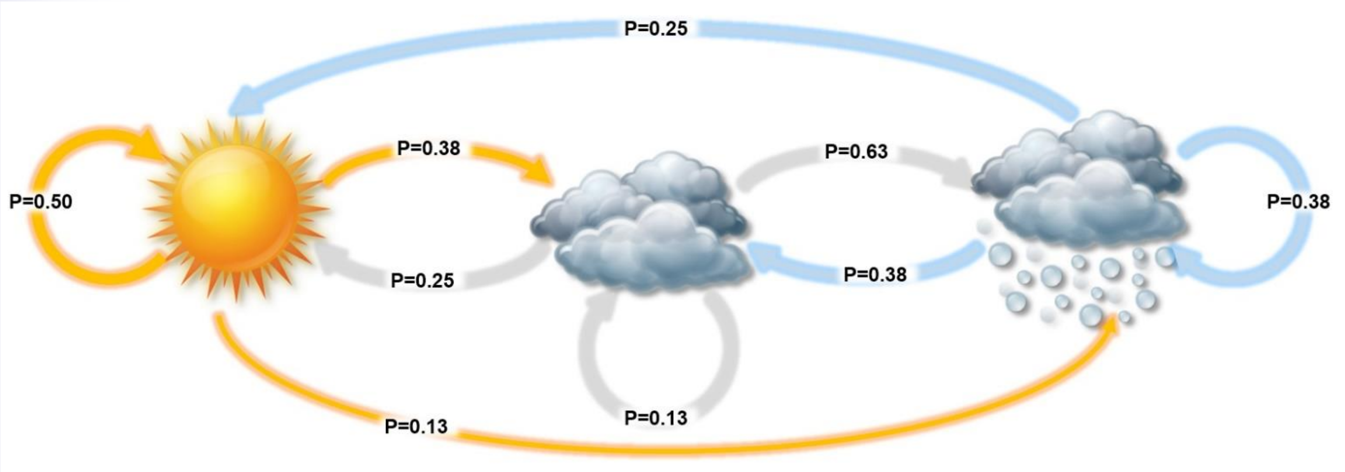
基础信息论

马尔可夫信源

华中科技大学电信学院

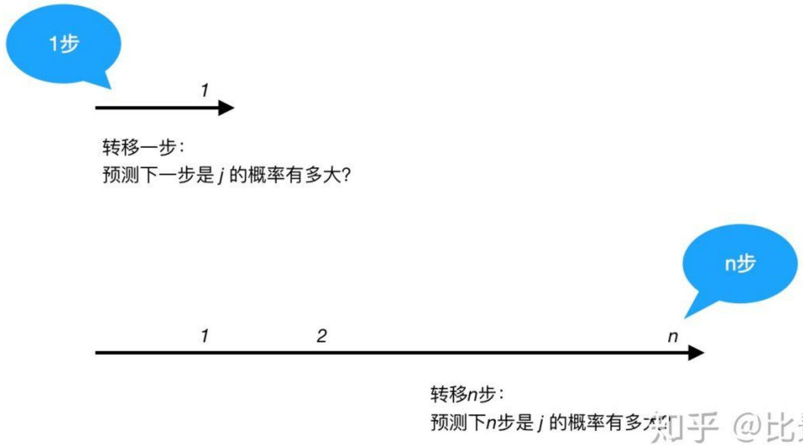
天气预报



$$P(X_n|X_{n-1},\dots,X_0)=P(X_n|X_{n-1})$$

- 明天是什么天气，后天是什么天气，大后天是什么天气。
- 每天（独立的天）的天气，在数学上可以用随机变量表达。

已知当前状态为
i，和转移矩阵



学习目标

- 构建马尔科夫信源的数学模型
- 计算马尔科夫信源的信息熵
- 分析马尔科夫的信息熵性质

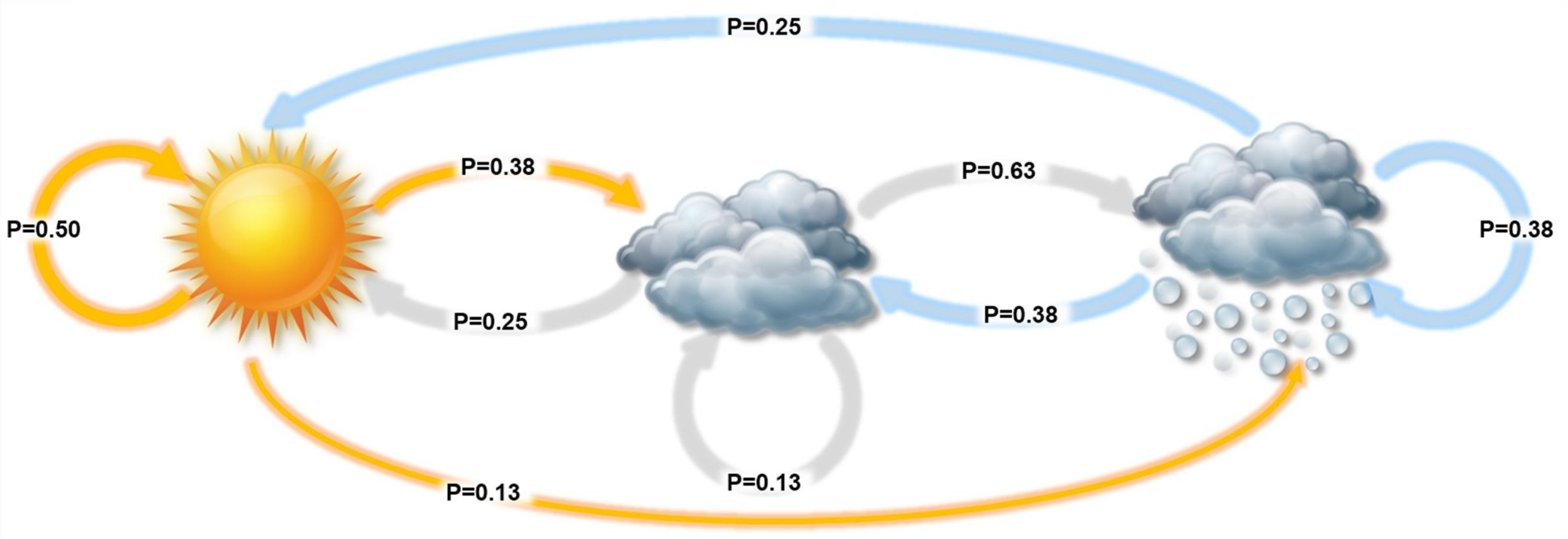
阅读：陈运，信息论与编码（第3版）第2章2.3.4节

答题

微助教

天气预报

- 基于下述天气的变化构建马氏链模型，回答
  1. 该马氏链多少个状态？
  2. 如果今天是阴天，明天转为晴天的概率多少？
  3. 如果今天是雨天，明天继续是雨天的概率多少？



马尔可夫信源

- 在很多信源的输出序列中，符号之间的依赖关系是**有限**的，任何时刻信源符号发生的概率只与前边已经发出的若干个符号有关，而与更前面的符号无关。

$\boxed{X_1X_2\ldots X_m}$ X_{m+1} $m+1$时刻的输出

X_1 $\boxed{X_2X_3\ldots X_{m+1}}$ X_{m+2} $m+2$时刻的输出

X_1X_2 $\boxed{X_3X_4\ldots X_{m+2}}$ X_{m+3} $m+3$时刻的输出

$\vdots$

为了描述这类信源除了信源符号集外还要引入状态。

马尔可夫信源的状态

- 何谓状态：
  - 在马尔可夫链中，唯一决定下一时刻输出符号概率分布的量。
  - 马尔可夫信源的状态：当前输出符号之前的m个符号。
  - 状态集：$S \in (e_1, e_2, \cdots e_J)$ 符号集：$X \in (x_1, x_2, \cdots x_n)$

$\boldsymbol{e_1}$ $\quad p(x_1/e_1)\ \ p(x_2/e_1)\ \ \cdots\ \ p(x_n/e_1)\quad \sum_{i=1}^{n} p(x_i/e_1) = 1$

$\boldsymbol{e_2}$ $\quad p(x_1/e_2)\ \ p(x_2/e_2)\ \ \cdots\ \ p(x_n/e_2)\quad \sum_{i=1}^{n} p(x_i/e_2) = 1$

…

$\boldsymbol{e_J}$ $\quad p(x_1/e_J)\ \ p(x_2/e_J)\ \ \cdots\ \ p(x_n/e_J)\quad \sum_{i=1}^{n} p(x_i/e_J) = 1$

马尔可夫信源定义

- 若一个信源满足下面两个条件，则称为马尔可夫信源：
- （1）某一时刻信源输出的符号的概率只与当前所处的状态有关，而与以前的状态无关；

$$p(X_l = x_k \,|\, S_l = e_j, X_{l-1} = x_{k_1}, S_{l-1} = e_i, \cdots) = p(X_l = x_k \,|\, S_l = e_j) = p(x_k/e_j)$$

其中，x_k、$x_{k_1} \in A$；e_i、$e_j \in S$

- （2）信源的下一个状态由当前状态和下一刻的输出唯一确定。

$$p(S_l = e_i \,|\, X_l = x_k, S_{l-1} = e_j) = \begin{cases} 1 \\ 0 \end{cases}$$

相关知识

- 符号输出概率：
  - 当马尔可夫链处于状态e_i时，发出符号集中某一符号x_k的概率，记为

$$\boxed{p(x_k/e_i)}$$

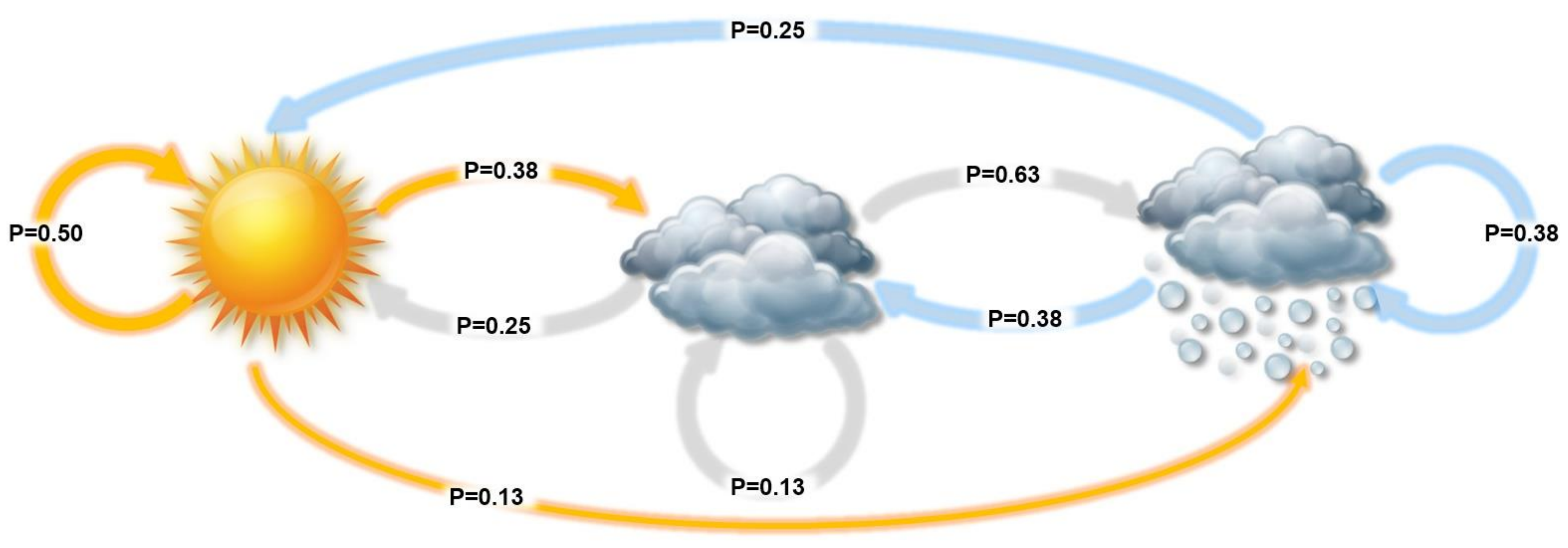
- 状态转移
  - 定义：每一时刻，当信源发出一个符号后，信源所处状态将发生变化，转入一个新的状态。所以，可将信源的输出符号系列变换成状态系列，将信源输出符号的不确定性问题变成信源状态的转换问题。
  - 状态一步转移概率：当马尔可夫链处于状态e_i时，发出某一符号后，状态转移为e_j的概率，记为 $\boxed{p(e_j/e_i)}$
  - 状态 k步转移概率：经过 k 步转移以后，马尔可夫链由状态 e_i转移到 e_j的概率，记为 $\boxed{p^k(e_j/e_i)}$

答题

微助教

提问

- 1）晴天的概率多少?
- 2）阴天的概率多少?
- 3）雨天的概率多少?

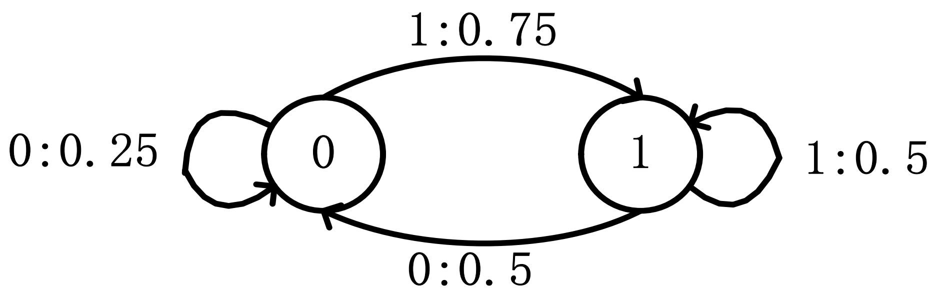


状态转移图

- 描述马尔可夫链状态转移过程的一种图形。圆圈代表状态，有向线段(弧)代表从状态的转移，用线段(弧)一侧的符号和数字代表发出的符号e_i和符号输出概率$p(x_k/e_i)$

- 例1：一个二进制一阶马尔可夫信源，符号集为$X=\{0,1\}$
- 条件概率为 $p(0|0)=0.25, p(0|1)=0.5$ $p(1|0)=0.75,\ p(1|1)=0.5$ $q=2, m=1$,所以$e1=0, e2=1.$
- 解：由条件概率可求得状态转移概率：

$$p(e1|e1)=0.25, p(e1|e2)=0.5$$
$$p(e2|e1)=0.75,\ p(e2|e2)=0.5$$



一阶马尔可夫信源状态转移图

例题2

- 设有一个二进制二阶马尔可夫信源，信源符号集为{0,1}

该信源符号数n=2，则共有4个状态，分别为：

$$e_1 = 00, e_2 = 01, e_3 = 10, e_4 = 11$$

条件概率为

$$p(0|00) = p(1|11) = 0.8,$$
$$p(1|00) = p(0|11) = 0.2,$$
$$p(0|01) = p(0|10) = p(1|01) = p(1|10) = 0.5$$

解：容易求出转移概率为

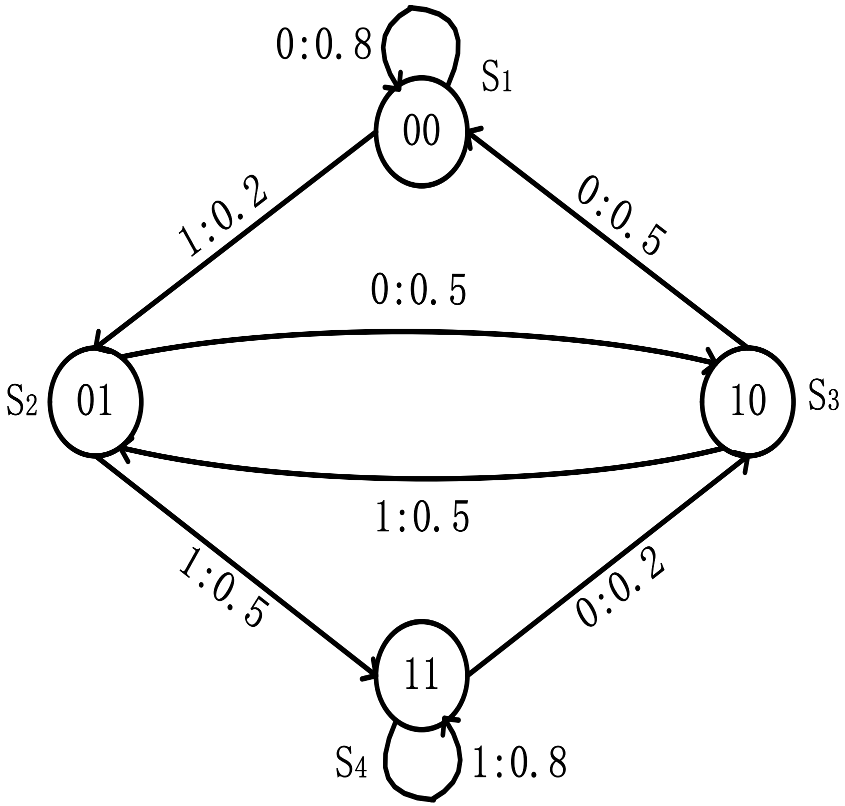
$$p(e_1|e_1) = p(e_4|e_4) = 0.8$$
$$p(e_2|e_1) = p(e_3|e_4) = 0.2$$
$$p(e_3|e_2) = p(e_1|e_3) = p(e_4|e_2) = p(e_2|e_3) = 0.5$$

例题2（续）

- 信源的状态转移矩阵为：

$$\Pi=\begin{bmatrix}0.8 & 0.2 & 0 & 0\\ 0 & 0 & 0.5 & 0.5\\ 0.5 & 0.5 & 0 & 0\\ 0 & 0 & 0.2 & 0.8\end{bmatrix}$$

- 状态转移图为：



例题3

例：设一信源，它在开始时以$P(a) = 0.6, P(b) = 0.3, P(c) = 0.1$的概率发出$X_1$。

如果X_1为a, 则X_2为a, b, c的概率为$\frac{1}{3}$；

如果X_1为b, 则X_2为a, b, c的概率为$\frac{1}{3}$；

如果X_1为c, 则X_2为a, b的概率为$\frac{1}{2}$, 为c的概率为0。

其后发出X_i的概率只与X_{i-1}有关，

且$P(X_i|X_{i-1}) = P(X_2|X_1)$，$i \geq 3$，请画出其状态转移图。

例题3（续）

解：由题意知，信源在开始发出信号后，后面发出什么符号只与前一个所发符号有关，即

$P(X_i|X_{i-1}) = P(X_2|X_1) i \geq 3$且

$$P(X_2 = a|X_1 = a) = P(X_2 = b|X_1 = a) = P(X_2 = c|X_1 = a) = \frac{1}{3}$$

$$P(X_2 = a|X_1 = b) = P(X_2 = b|X_1 = b) = P(X_2 = c|X_1 = b) = \frac{1}{3}$$

$$P(X_2 = a|X_1 = c) = P(X_2 = b|X_1 = c) = \frac{1}{2}$$

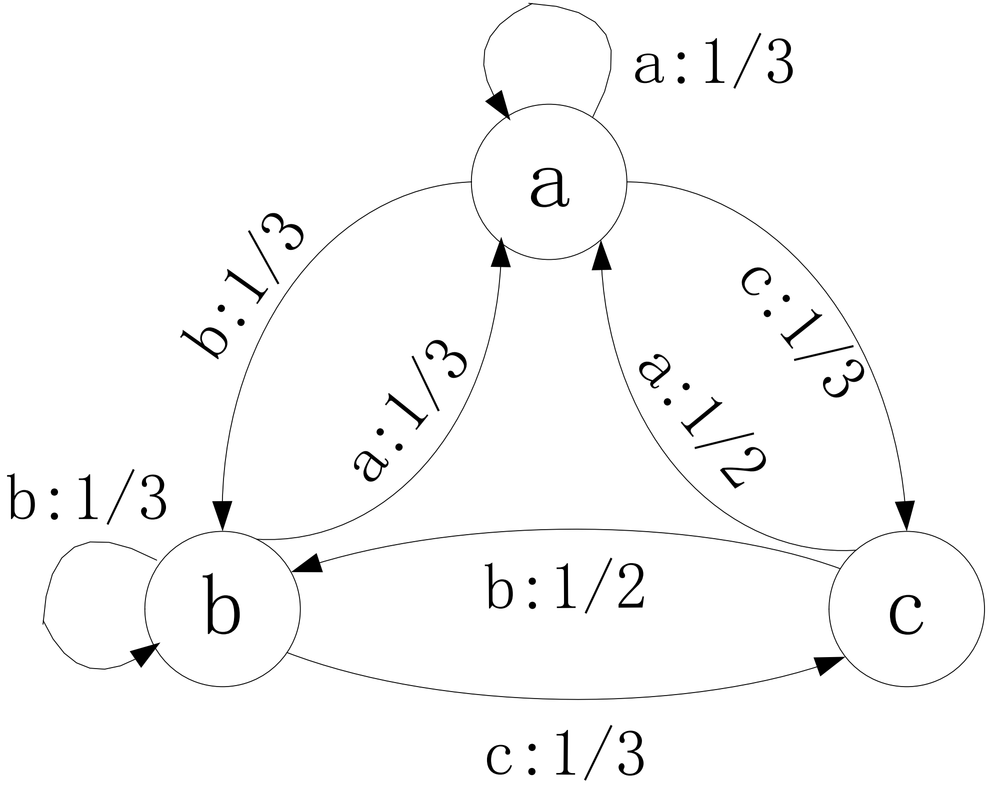
$$P(X_2 = c|X_1 = c) = 0$$

例题3（续）

- 一步转移矩阵为P

$$\mathbf{P}=\begin{bmatrix}\frac{1}{3} & \frac{1}{3} & \frac{1}{3}\\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3}\\ \frac{1}{2} & \frac{1}{2} & 0\end{bmatrix}$$

由此可见该信源是一阶马尔可夫信源，状态空间就等于信源符号集$E=\{a,b,c\}$,其状态转移图如下



马尔可夫信源的极限熵

- 因为信源发出的符号只与最近的m个符号有关，所以极限熵为

$$= H(X_{m+1}|X_1,X_2,\cdots,X_m)$$
$$= H_{m+1}$$

- 即：m阶马尔可夫信源的极限熵等于m阶条件熵

$$H(x_{m+1}|x_m,\cdots,x_1) = H_{m+1}$$

$$\boxed{p(x_{m+1}|x_m,\cdots,x_1) = p(e_j|e_i)}$$

$$= -\sum_{m+1,\cdots,1} p(x_{m+1},\cdots,x_1) \bullet \log p\,(x_{m+1}|x_m,\cdots,x_1)$$

$$= -\sum_i \sum_j p(e_i)p(e_j|e_i) \log p\,(e_j|e_i)$$

马尔可夫信源的极限熵

$$H_\infty = H_{m+1}$$

$$H_\infty = H_{m+1} = -\sum_i \sum_j p(e_i) p(e_j|e_i) \log p(e_j|e_i)$$

一步转移概率是给定的

$p(e_j)$：信源的平稳分布（稳定后各状态的极限概率）

有限齐次马尔可夫链满足以下条件：

$$p(e_j) = \sum_{i=1}^{n^m} p(e_i) p(e_j/e_i) \quad (j = 1,2,\dots,n^m)$$

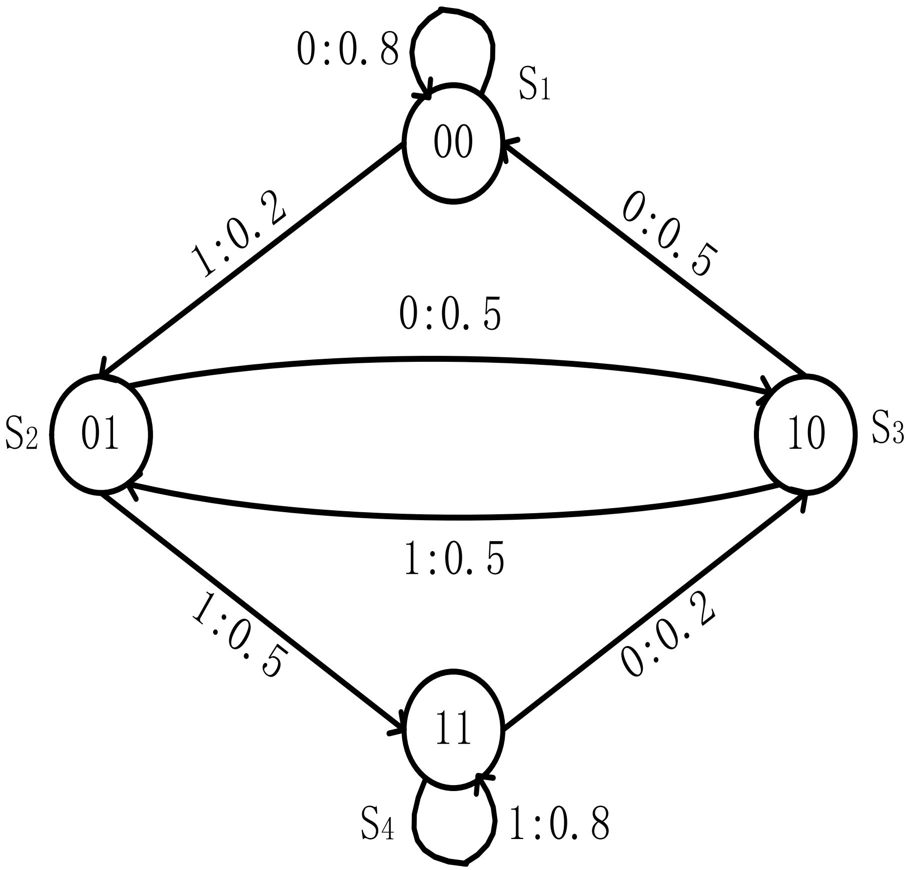
$$p(e_j) > 0, \sum_{j=1}^{n^m} p(e_j) = 1$$

注意

- 1. 极限熵并非一定存在。
- 对于n元m阶马尔可夫信源，要求：
- a)平稳信源（如果不平稳则先把其变成分段平稳的）
- b）$p(s_j)$存在，其中$j = 1,2,\cdots,n^m$

例题1

- 信源的状态转移图如下所示，求极限熵



信源的状态转移矩阵为：

$$\Pi=\begin{bmatrix}0.8 & 0.2 & 0 & 0\\ 0 & 0 & 0.5 & 0.5\\ 0.5 & 0.5 & 0 & 0\\ 0 & 0 & 0.2 & 0.8\end{bmatrix}$$

例题1（续）

信源的状态转移矩阵为：

$$\Pi=\begin{bmatrix}0.8 & 0.2 & 0 & 0\\ 0 & 0 & 0.5 & 0.5\\ 0.5 & 0.5 & 0 & 0\\ 0 & 0 & 0.2 & 0.8\end{bmatrix}$$

- 解：先求信源的极限概率
- 因为：

$$p(s_j)=\sum_i p(s_i)\,p(s_j|s_i)$$

$$p(s_1)=p(s_1)p(s_1|s_1)+p(s_2)p(s_1|s_2)+p(s_3)p(s_1|s_3)+p(s_4)p(s_1|s_4)$$

$$=0.8p(s_1)+0.5p(s_3)$$

- 同理有：

$$p(s_2)=0.2p(s_1)+0.5p(s_3)$$
$$p(s_3)=0.5p(s_2)+0.2p(s_4)$$
$$p(s_4)=0.5p(s_2)+0.8p(s_4)$$

- 解上述方程组

$$p(s_1)=p(s_4)=\frac{5}{14}\qquad p(s_2)=p(s_3)=\frac{2}{14}$$

马尔可夫信源熵-例题(续)

信源的状态转移矩阵为：

$$\mathbf{II}=\begin{bmatrix} 0.8 & 0.2 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.2 & 0.8 \end{bmatrix}$$

极限熵可求得为：

$$H_\infty = H_{m+1} = H_{2+1} = H_3 =$$

$$= -\sum_i \sum_j p(e_i)p(e_j|e_i)\log p\,(e_j|e_i)$$

$$= \frac{5}{14}H(0.8,0.2) + \frac{1}{7}H(0.5,0.5) + \frac{1}{7}H(0.5,0.5) + \frac{5}{14}H(0.8,0.2)$$

$$= \frac{5}{7}\times 0.7219 + \frac{2}{7}\times 1 = 0.80$$ 比特/符号

m阶马尔可夫与一般有记忆信源的区别

1. 马尔可夫信源发出一个个的符号，有限长度有记忆信源发出一组组符号；
2. 一般有记忆信源用联合概率描述符号间的关联关系，马尔可夫信源用条件概率（状态转移概率）来描述符号间的关联关系；
3. 马尔可夫信源记忆长度虽然有限，但依赖关系延伸到无穷远；长为m的有限记忆信源符号间的依赖关系仅限于每组内，组与组之间没有依赖关系；
4. 马尔可夫信源的极限熵是条件熵，m长有记忆信源的极限熵是平均符号熵。

信源冗余度

信源冗余度： 信源的冗余程度或重复程度。

实际信源通常存在着冗余，所以才有可能对其进行压缩。

信源压缩： 利用对信源进行**编码**的方法，用**尽可能少的码符号数携带同样多的信息量**。

实际信源： 严格来讲，大多是关联（记忆）长度为无穷大的多符号信源。

对实际信源，其所提供的信息量应该用H_∞衡量。

但涉及到求解无穷维联合概率分布的问题。

将实际信源近似为 多符号信源 或m阶马尔可夫信源。

近似为马尔可夫信源

当近似为马尔可夫信源后，显然阶数 m 越大，越接近实际情况。因此有：

$$H_\infty \le H_{m+1} \triangleq H(X_{m+1}|X_1X_2\cdots X_m) \le \cdots \le H_{1+1} \triangleq H(X_2|X_1)$$

$$\le \underbrace{H_{0+1} \triangleq H(X)}_{} \le \underbrace{H_0 = \log n}_{}$$

1阶马尔可夫信源

0阶马尔可夫信源，无记忆信源

等概率分布的无记忆信源

注意： **为和多符号信源中的平均符号熵相区分，采用H_{m+1}的记法。**

实例：英语信源。

英语信源分析1

英语中包含26个英文字母，假设不区分大小写，并只有空格一个标点符号。

分析1：对英语信源，最粗略的近似可以如何进行处理?

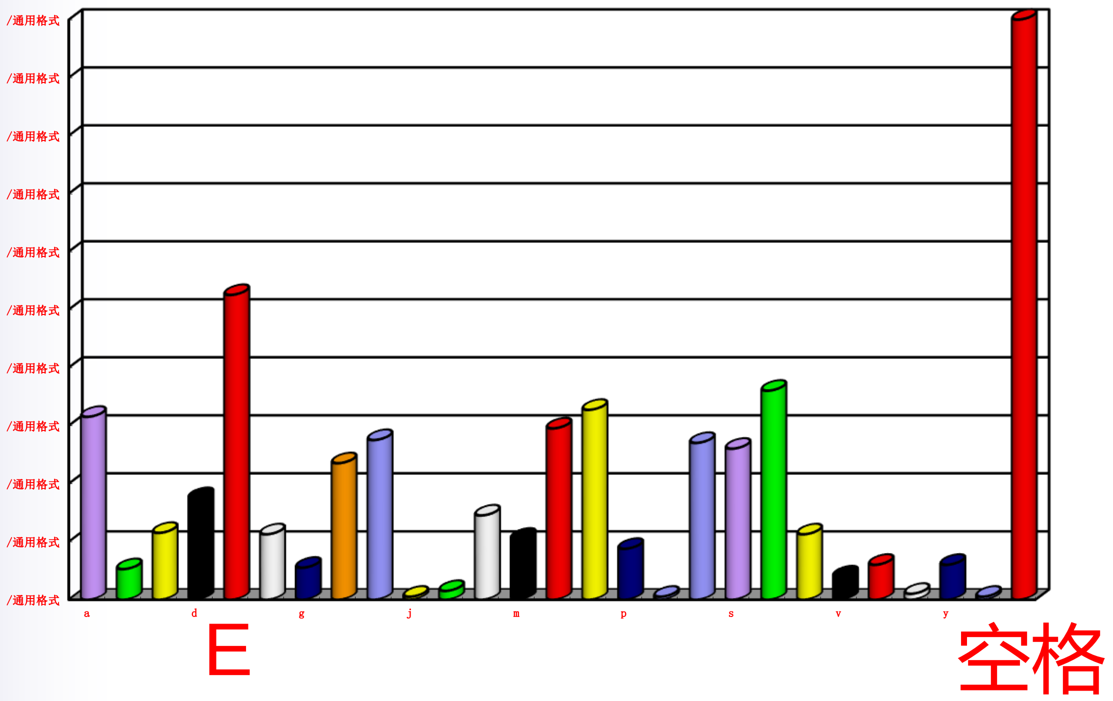
回答：假设认为前后符号间**不相关**，并且所有27个符号**等概率分布**。

$$H_0 = \log 27 \approx 4.76 \quad \text{比特/符号}$$

H_0**为信源的最大熵**

英语符号概率分布

实际英语信源，并非等概率分布



符号	概率	符号	概率	符号	概率
空格	0.2	S	0.052	Y, W	0.012
E	0.105	H	0.047	G	0.011
T	0.072	D	0.035	B	0.0105
O	0.0654	L	0.029	V	0.008
A	0.063	C	0.023	K	0.003
N	0.059	F, U	0.0225	X	0.002
I	0.055	M	0.021	J, Q	0.001
R	0.054	P	0.0175	Z	0.001

英文字母出现概率统计

英语信源分析2

分析2：考虑英语符号概率分布，不考虑符号间依赖关系的情况下，平均符号熵等于多少?

$$H_{0+1} = -p(a)\cdot\log p\,(a) - p(b)\cdot\log p\,(b) - \cdots - p(\text{空格})\cdot\log p\,(\text{空格})$$

$$H_{0+1} \approx 4.03\text{比特/符号}$$

问题：上述信源与实际情况近似到何种程度?

分析：按表的概率分布，随机选择英语字母得到一个信源输出序列为。

AI_NGAE_ITE_NNR_ASAEV_OTE_BAINTHA_HYROO_POER_SETRYGAIETRWCO_EHDUARU_EUEU_C_FT_NSREM_DIY_EESE_F_O_SRIS_R_UNNASHOR…

英语信源分析3

分析3：考虑符号间依赖关系，可近似为马尔可夫信源。

1. 近似为一阶马尔可夫信源

前一个字母	后一个字母	条件概率
A	A	$P(A/A)$
	B	$P(B/A)$
	$\vdots$	$\vdots$
	空格	$P($空格$/A)$
B	A	$P(A/B)$
	B	$P(B/B)$
	$\vdots$	$\vdots$
	空格	$P($空格$/B)$

$$H_{1+1} = H(X_2|X_1)$$

$$= -\sum_{i=1}^{27}\sum_{j=1}^{27} p(x_i)\cdot p(x_j|x_i)\cdot \log p(x_j|x_i)$$

$$\approx 3.32$$ **比特/符号**

方法： 首字母可以任意选择。

首字母选定后，按条件概率选第二个字母。

第二个字母选定后，再按条件概率选第三个。

……

英语信源分析3（续）

2. 类似地，近似为二阶马尔可夫信源。

$$H_{2+1} = H(X_3|X_2X_1) \approx 3.1 \quad \text{比特/符号}$$

输出结果实例：

IANKS_CAN_OU_ANG_RLER_THTTED_OF_TO_SHOR_OF_TO_HAVEMEM_A_I_MAND_AND_BUT_WHISS_ITABLY_THERVEREER…

3. 类似地，可将英语信源近似为三阶、四阶… 。

$$\vdots$$

$$H_\infty \approx 1.4 \ \text{比特/符号}$$

依赖关系越多，及马尔科夫信源的阶数越高，输出的序列越接近实际情况。

英语信源分析3（续）

$$H_\infty \approx 1.4 \quad \le \cdots \le \quad H_{2+1} \approx 3.1 \qquad H_{1+1} \approx 3.32 \qquad H_{0+1} \approx 4.03 \quad \le H_0 \approx 4.76$$

上述结果，验证了随着阶数m的增加，符号相关性增加，熵值（平均每个符号所携带的信息量）会降低。

实际英语：

$$\underbrace{\text{Hello, My name is Lai. How are you}}_{L\text{个字符}}$$

问题：携带的信息量?

$$L \cdot H_\infty$$

例

假设有一个27元（27种可能的符号）、等概率分布、无记忆的符号序列。

符号集： $\{a, b, c, d, \cdots, \_\}$

$$\underbrace{\text{bcnmlas\_giovdwphueftzyqrxjk}}_{K\text{个码符号}}$$

携带的信息量 = $K \cdot \log n = K \cdot H_0$

$\therefore \quad K \cdot H_0 = L \cdot H_\infty \quad \Rightarrow \quad K = \dfrac{H_\infty}{H_0} \cdot L$

H_∞ **与** H_0 **越接近，可压缩的程度越小，冗余度越小。**

当 $H_\infty = H_0$ **时，已无法实现压缩，冗余度等于零。**

冗余度的定义

信源的冗余度： $\xi = 1 - \frac{H_\infty}{H_0}$ *

信息变差

对上式通分后，可得： $\xi = \frac{H_0 - H_\infty}{H_0} = \frac{I_{0\infty}}{H_0}$ *

英语信源的冗余度： $\xi = 1 - \frac{H_\infty}{H_0} \approx 1 - \frac{1.4}{4.76} = 79\%$

问题： 79%代表什么含义?

回答：从平均意义而言，一大段英语文字中有79%的信息都是多余的，是由英语的语法结构、表达习惯决定的。只有21%的内容是作者可以自由选择的。理论上讲，通信时只需传送21%的内容，其余内容可依据英语信源的统计特性推算得出。

思考

讨论

微助教

- 除了统计分布，语言中还有哪些冗余因素?

Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

研表究明，汉字序顺并不定一影阅响读。比如当你看完这句话后，才发这现里的字全是都乱的。

冗余度与传输效率

- 信源有冗余，可进行压缩
- 信源编码：
- 信源编码是通过尽可能压缩信源冗余度的手段，实现提高通信有效性的目的。

例：中华人民共和国 $\xrightarrow{\text{压缩}}$ **中国　　效率最高**

评论：

1. 压缩后信源的冗余度越低，通信的有效性越好。

2. 信源冗余度过低，甚至没有冗余度，又会带来通信可靠性方面的问题。

冗余度与传输可靠性

若通信过程中出现错误：

1. 当信源无冗余度

中国 → ×国　美国 法国 德国 …?

中国 → 中 ×　中国 中央 中间 …?

2. 当信源存在一定冗余度

中华人民共和国 → ×华人民×和国 $\xrightarrow{\text{恢复}}$ 中华人民共和国

结论：通信有效性（信源编码）与可靠性（信道编码）往往是一对矛盾。

例

用α, β, γ三个字符组字，设组成的字有以下三种情况：

(1) 只用α一个字母的单字母字。

(2) 用α开头或结尾的两字母字。

(3) 把α夹在中间的三字母字。

假定由这三种字组成一种简单语言，试计算当所有字等概率出现的语言的冗余度。

解：求解思路

$$\xi = 1 - \frac{H_\infty}{H_0}$$

H_∞ 考虑字符前后联系、以及概率分布时，平均每个字符的熵。

H_0 前后字符独立、等概率出现时的熵。

$H_0 = \log 3$ 比特/符号

例题（续）

- 接下来求：

$$H_\infty = \frac{\log(\text{字的个数})}{\text{每个字包含的平均字符数}}$$

- 分析该语言有哪些字：

(1)单字母字: α **1个**

(2)双字母字: **5个**

$\begin{cases} \textbf{以}\alpha\textbf{开头:} & C_3^1 = 3 \ \textbf{个} \quad \alpha\alpha, \alpha\beta, \alpha\gamma \\ \textbf{以}\alpha\textbf{结尾:} & C_3^1 = 3 \ \textbf{个} \quad \alpha\alpha, \beta\alpha, \gamma\alpha \end{cases}$

(3)三字母字: $C_3^1 \cdot C_3^1 = 9$ **9个**

$\alpha\alpha\alpha, \alpha\alpha\beta, \alpha\alpha\gamma,\quad \beta\alpha\alpha, \beta\alpha\beta, \beta\alpha\gamma,\quad \gamma\alpha\alpha, \gamma\alpha\beta, \gamma\alpha\gamma$

例题（续）

- 统计可得：

字的个数$= 1 + 5 + 9 = 15$**个**

每个字包含的平均字符数$= 1 \cdot \frac{1}{15} + 2 \cdot \frac{5}{15} + 3 \cdot \frac{9}{15} = \frac{38}{15}$**个**

$$\Rightarrow \quad H_\infty = \frac{\log 15}{\frac{38}{15}} \approx 1.542$$ **比特/符号**

- 该语言的冗余度为： $\therefore \ \xi = 1 - \frac{H_\infty}{H_0} \approx 0.027$

总结

- 讨论具有平稳性和遍历性的马尔可夫信源。
- 证明马尔可夫信源极限熵的存在条件：信源的状态极限概率存在；给出了马尔可夫信源极限熵的求解方法。
- 设计实际通信系统时，信源剩余度的存在对传输是不利的，应当尽量压缩信源剩余度，以使信源发出的每个符号携带的平均信息量最大。
- 若考虑通信中的抗干扰问题时，则信源剩余度是有利的，常常人为的加入某种特殊的剩余度，以增强通信系统的抗干扰能力。

谢谢！

黑晓军
华中科技大学
电子信息与通信学院
Email: heixj@hust.edu.cn
网址： http://eic.hust.edu.cn/aprofessor/heixiaojun

参考资料

- 陈运，信息论与编码（第3版）第2章2.3.4，电子工业出版社出版，2012
- 马尔可夫链，https://zhuanlan.zhihu.com/p/37847722