第2章 随机过程和Markov信源熵（第三部分）

樊平毅 教授

清华大学电子工程系 WIST LAB.

Email: fpy@tsinghua.edu.cn

2022年9月14日

随机过程的信息熵

- 对于独立同分布的随机变量，可以直接利用其概率分布计算其信息熵， 而当随机变量序列不是独立同分布的，是否有工具可以研究其熵。

- 在随机过程中，最直接的随机过程就是严平稳过程，利用其任何有限维分布都是时间推移不变的特征加以刻画。 于是，Markov 过程作为一种典型的随机过程被首先加以研究，而其他类型的平稳随机过程相对比较难，需要借助随机场的概念进行处理。

- 一个Markov 过程，如果其初始分布是平稳分布，那么该过程就是平稳的；对于有限状态的过程，它是严平稳的。此时可以借助联合熵的概念计算其熵率，而非信息熵。

- **开放问题： 一般的平稳过程的熵率如何计算？ 在平稳随机场中是否可以沿用计算马尔科夫链的方法进行？算法的收敛性条件是什么？**

Acta Mathematica Scientia
Estimation of Average Differential Entropy for Stationary Ergodic Space-Time Random Field on Bounded Area

随机过程$\{X_i\}$是一个带下标的随机变量序列。一般允许随机变量间具有任意的相关性。刻画一个过程需要知道所有有限的联合概率密度函数

$$\Pr\{(X_1,X_2,\cdots,X_n)=(x_1,x_2,\cdots,x_n)\}=p(x_1,x_2,\cdots,x_n)$$

其中$(x_1,x_2,\cdots,x_n)\in\mathcal{X}^n$，$n=1,2,\cdots$。

定义 如果随机变量序列的任何有限子集的联合分布关于时间下标的位移不变，即对于每个 n 和位移 l，以及任意的 $x_1,x_2,\cdots,x_n\in\mathcal{X}$，均满足

$$\begin{aligned}&\Pr\{X_1=x_1,X_2=x_2,\cdots,X_n=x_n\}\\&=\Pr\{X_{1+l}=x_1,X_{2+l}=x_2,\cdots,X_{n+l}=x_n\}\end{aligned}\tag{4-1}$$

则称该随机过程是平稳的。

严平稳过程的定义

定义 如果对 $n=1,2,\cdots$，及所有的 $x_1,x_2,\cdots,x_n\in\mathcal{X}$，有

$$\Pr(X_{n+1}=x_{n+1}\mid X_n=x_n,X_{n-1}=x_{n-1},\cdots,X_1=x_1)$$
$$=\Pr(X_{n+1}=x_{n+1}\mid X_n=x_n)$$

则称离散随机过程 $X_1,X_2,\cdots$ 为马尔可夫链或马尔可夫过程。

随机变量的联合概率密度函数可以写为

$$p(x_1,x_2,\cdots x_n)=p(x_1)p(x_2\mid x_1)p(x_3\mid x_2)\cdots p(x_n\mid x_{n-1})$$

定义 如果条件概率 $p(x_{n+1}\mid x_n)$ 不依赖于 n，即对 $n=1,2,\cdots$，有

$$\Pr\{X_{n+1}=b\mid X_n=a\}=\Pr\{X_2=b\mid X_1=a\}\quad \text{对任意 } a,b\in\mathcal{X}$$

则称马尔可夫链是时间不变的。

利用严平稳的时间推移不变的特征

如果$\{X_i\}$为马尔可夫链，则称 X_n 为 n 时刻的状态。一个时间不变的马尔可夫链完全由其初始状态和概率转移矩阵 $P=[P_{ij}]$所表征，其中 $P_{ij}=\Pr\{X_{n+1}=j/X_n=i\}$，$i,j\in\{1,2,\cdots,m\}$。

若马尔可夫链可以从任意状态经过有限步转移到另一任意状态，且其转移概率为正，则称此马尔可夫链是不可约的。如果从一个状态转移到它自身的不同路径长度的最大公因子为 1，则称马尔可夫链是非周期的。

$$p(x_{n+1})=\sum_{x_n}p(x_n)P_{x_nx_{n+1}} \tag{4-5}$$

若在 $n+1$ 时刻，状态空间上的分布与在 n 时刻的分布相同，则称此分布为平稳分布。如果马尔可夫链的初始状态服从平稳分布，那么该马尔可夫链为平稳过程，这也正是平稳分布的称谓由来。

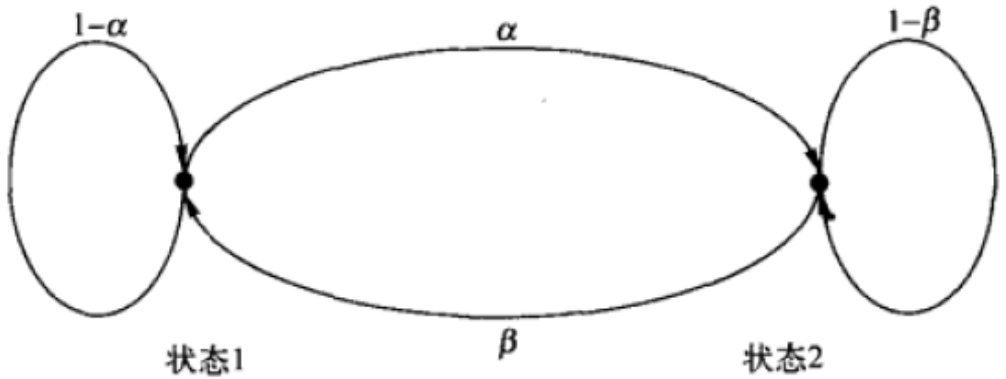
重点关注的这类随机过程

若有限状态马尔可夫链是不可约的和非周期的，则它的平稳分布惟一，从任意的初始分布出发，当 $n\to\infty$时，X_n 的分布必趋向于此平稳分布。

例题

考虑两状态的一个马尔可夫链，其概率转移矩阵为

$$P=\begin{bmatrix}1-\alpha & \alpha \\ \beta & 1-\beta\end{bmatrix}$$



其平稳分布为 $\pi P=\pi$ ⟶ $\mu_1\alpha=\mu_2\beta$

平稳分布为

$$\mu_1=\frac{\beta}{\alpha+\beta},\mu_2=\frac{\alpha}{\alpha+\beta}$$

在 n 时刻的状态 X_n 的熵

$$H(X_n)=H\left(\frac{\beta}{\alpha+\beta},\frac{\alpha}{\alpha+\beta}\right)$$

这并非熵 $H(X_1,X_2,\cdots,X_n)$ 的增长速率。

由于 X_i 之间存在着相关性

熵的增长速率需要扣除这种相关性的影响

熵率的定义

如果给定一个长度为 n 的随机变量序列，我们自然会问：该序列的熵随 n 如何增长？下面定义这个增长率，我们称为熵率。

定义 当如下极限存在时，随机过程 $\{X_i\}$ 的熵率定义为

$$H(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \cdots, X_n)$$

实际上这是定义的线性熵率，是否存在其他形式的熵率？

假定一台打字机可输出 m 个等可能的字母 由此打字机可产生长度为 n 的序列数为m^n 并且都等可能出现

因此，$H(X_1, X_2, \cdots, X_n) = \log m^n$

熵率为 $H(\mathcal{X}) = \log m$ 比特/字

特点： 独立同分布序列， 熵率等于熵

i.i.d. 随机变量序列 $X_1, X_2, \cdots, X_n$。此时，有

$$H(\mathcal{X}) = \lim \frac{H(X_1, X_2, \cdots, X_n)}{n} = \lim \frac{nH(X_1)}{n} = H(X_1)$$

这正是我们所期望的每字符的熵率。

独立但非同分布的随机变量序列。在此情形下，有

$$H(X_1, X_2, \cdots, X_n) = \sum_{i=1}^{n} H(X_i)$$

但 $H(X_i)$不全相等

需要考察 $\frac{1}{n}\sum H(X_i)$的极限

例如取二值随机分布序列，其中 $p_i = P(X_i = 1)$不是常数，而为 i 的函数

例如，对 $k = 0,1,2\cdots$，取

$$p_i = \begin{cases} 0.5 & 2k < \log\log i \leqslant 2k+1 \\ 0 & 2k+1 < \log\log i \leqslant 2k+2 \end{cases}$$

此时，该序列的情况是，满足 $H(X_i) = 1$ 的随机变量序列(可以任意长)之后，紧接着是更长以指数变化的序列满足 $H(X_i) = 0$。所以，$H(X_i)$的累积平均值将在 0 与 1 之间振荡，从而不存在极限。因此，该过程的 $H(\mathcal{X})$无定义。

也可以定义熵率的一个相关的量(如果下列极限存在)：

直接定义熵率 $$H'(\mathcal{X}) = \lim_{n\to\infty} H(X_n \mid X_{n-1}, X_{n-2}, \cdots, X_1)$$

定理： 对于平稳过程，如果极限存在，则 $H(\mathcal{X}) = H'(\mathcal{X})$

说明在特定场景下，在两种熵率定义中，选择任何一种极限定义都是可以的

定理 4.2.2 对于平稳随机过程，$H(X_n|X_{n-1},\cdots,X_1)$随 n 递减且存在极限$H'(\mathcal{X})$.

证明：

$$H(X_{n+1}|X_1,X_2,\cdots,X_n)\leqslant H(X_{n+1}|X_n,\cdots,X_2)$$
$$=H(X_n|X_{n-1},\cdots,X_1)$$

其中的不等式由条件作用使熵减小这个性质得到，而等式由该过程的平稳性得到.

利用单调有界序列的极限存在性质可得所要结论

若 $a_n\to a$，且 $b_n=\frac{1}{n}\sum_{i=1}^{n}a_i$，则 $b_n\to a$。

利用

$$\frac{H(X_1,X_2,\cdots,X_n)}{n}=\frac{1}{n}\sum_{i=1}^{n}H(X_i\mid X_{i-1},\cdots,X_1)$$

$$H(\mathcal{X})=\lim\frac{H(X_1,X_2,\cdots,X_n)}{n}=\lim H(X_n|X_{n-1},\cdots,X_1)$$
$$=H'(\mathcal{X})$$

对应的数值计算方法，利用大数定律和上述定理

$$-\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) \to H(\mathcal{X})$$

马尔可夫链 对于平稳的马尔可夫链，熵率为

$$H(\mathcal{X}) = H'(\mathcal{X}) = \lim H(X_n \mid X_{n-1}, \cdots, X_1) = \lim H(X_n \mid X_{n-1})$$
$$= H(X_2 \mid X_1)$$

平稳分布 μ 为下列方程组的解 $\mu_j = \sum_i \mu_i P_{ij}$ 对任意的 j

两步法：
先计算平稳分布；
在计算熵率

定理：设 $\{X_i\}$ 为平稳马尔可夫链，其平稳分布为 μ，转移矩阵为 P。则熵率为

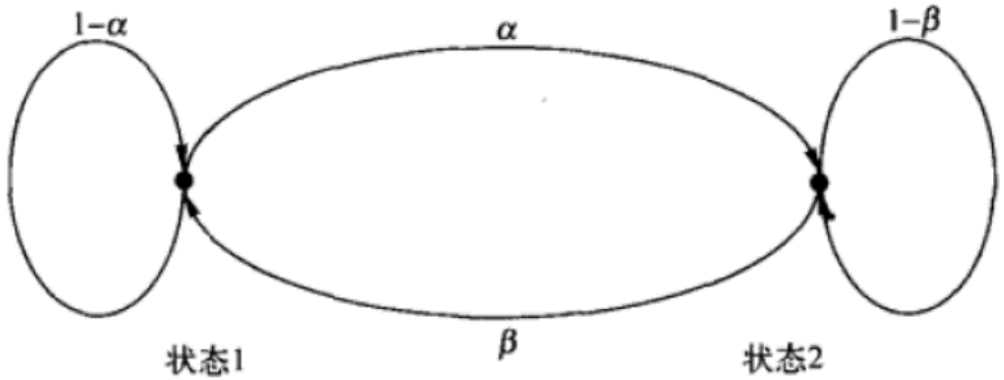
$$H(\mathcal{X}) = -\sum_{ij} \mu_i P_{ij} \log P_{ij}$$

证明：$H(\mathcal{X}) = H(X_2 \mid X_1) = \sum_i \mu_i \left(\sum_j -P_{ij} \log P_{ij} \right)$

例题

考虑两状态的一个马尔可夫链，其概率转移矩阵为

$$P=\begin{bmatrix}1-\alpha & \alpha \\ \beta & 1-\beta\end{bmatrix}$$



（1）平稳分布为

$$\mu_1=\frac{\beta}{\alpha+\beta},\mu_2=\frac{\alpha}{\alpha+\beta}$$

（两状态的马尔可夫链）如图4.1所示的两状态马尔可夫链的熵率为

（2）

$$H(\mathcal{X})=H(X_2|X_1)=\frac{\beta}{\alpha+\beta}H(\alpha)+\frac{\alpha}{\alpha+\beta}H(\beta)$$

思考与分析

- 注意：我们此处讨论的都是非周期不可约的Markov 过程的求解

➢如果是周期的，如何求解？

➢如果非周期，但存在多个平稳分布，如何求解？

➢应如何讨论，给出相应的定义？

➢这类问题在讨论博弈论中是否有价值？存在多个纳什均衡点的情况？

例题：加权图上的熵率计算

考虑一个连通图(图4-2)上的随机游动

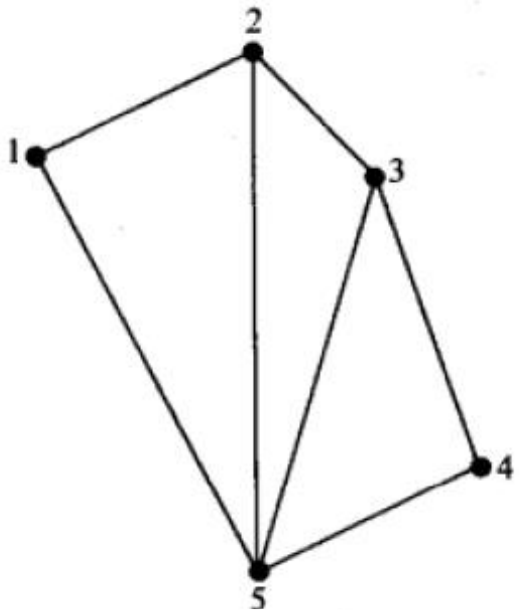
假定该图有 m 个标记 为 $\{1,2,\cdots,m\}$ 的节点

其中连接节点 i 和 j 的边权重为 $W_{ij} \geqslant 0$

假定此图是无向的 $W_{ij} = W_{ji}$

若节点 i 和 j 没有连接边，则设 $W_{ij} = 0$

有一个粒子在图中由一个节点到另一个节点作随机游动。设随机游动 $\{X_n\}$，$X_n \in \{1,2,\cdots,m\}$ 为图的一个顶点序列。若 $X_n = i$，那么下一个顶点 j 只可能是与节点 i 相连的所有节点中的一个，且转移概率为连接 i 和 j 的边权重所占所有与 i 相连的边的权重之和的比例。因此，$P_{ij} = W_{ij} / \sum_k W_{ik}$。



一个图上的随机游动

建立一个Markov 模型，
给出状态空间和转移概率矩阵

平稳分布为 $\mu_i = \frac{W_i}{2W}$

平稳分布是链接边权重的比例因子

其中 $W_i = \sum_j W_{ij}$

$W = \sum_{i,j:j>i} W_{ij}$ → $\sum_i W_i = 2W$

$$\sum_i \mu_i P_{ij} = \sum_i \frac{W_i}{2W}\frac{W_{ij}}{W_i} = \sum_i \frac{1}{2W} W_{ij} = \frac{W_j}{2W} = \mu_j$$

因此，状态 i 的平稳概率为连接节点 i 的各边权重总和占所有的边权重总和的比例.

$$H(\mathcal{X}) = H(X_2 | X_1)$$
$$= -\sum_i \mu_i \sum_j P_{ij} \log P_{ij}$$

例题

（棋盘上的随机游动） 假定一个“王”在 8×8 的（国际象棋）棋盘上作随机游动

“王”这个棋子在棋盘内部时可有 8 个移位，在边缘时有 5 个移位，在角落时有 3 个移位.

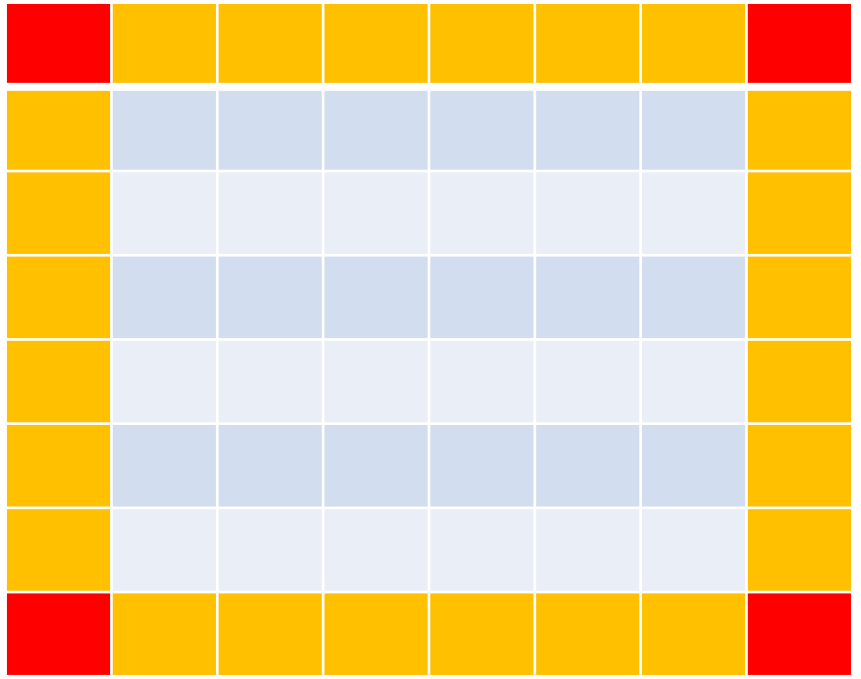
平稳概率分别是$\frac{8}{420}$，$\frac{5}{420}$和$\frac{3}{420}$

大家可以计算4*3+24*5+36*8=420;

转移概率1/3, 1/5, 1/8

从而，熵率为 0.92 log8.

边缘效应



应用：热力学第二定律

热力学第二定律是物理学中的基本定律之一，表明孤立系统的熵总是不减的

在统计热力学中，熵通常定义为物理系统的微观状态数的对数值

现在我们建立模型，将孤立系统视为一个马尔可夫链

其中状态的转移规律由控制该系统的 物理定律所决定

关于第二定律的 4 种不同解释　　从信息论的角度去看问题

1. 相对熵 $D(\mu_n \| \mu_n')$ 随 n 递减

2. 在 n 时刻状态空间上的分布 μ_n 与平稳分布 μ 之间的相对熵 $D(\mu_n \| \mu)$ 随 n 递减

3. 若平稳分布是均匀分布，则熵增加

4. 对于平稳的马尔可夫过程，条件熵 $H(X_n | X_1)$ 随 n 递增

1. 相对熵 $D(\mu_n \| \mu_n')$ 随 n 递减 ← 越来越逼近平稳分布

设 μ_n 和 μ_n' 为 n 时刻的马尔可夫链状态空间上的两个概 率分布

而 μ_{n+1} 和 μ_{n+1}' 是时刻 $n+1$ 时的相应分布　对应的联合概率密度分别记为 p 和 q，

$$p(x_n, x_{n+1}) = p(x_n) r(x_{n+1} | x_n)$$

$$q(x_n, x_{n+1}) = q(x_n) r(x_{n+1} | x_n)$$

$$D(p(x_n) \| q(x_n)) \geqslant D(p(x_{n+1}) \| q(x_{n+1}))$$

其中 $r(\cdot|\cdot)$ 表示马尔可夫链的概率转移函数

$$D(p(x_n, x_{n+1}) \| q(x_n, x_{n+1})) = D(p(x_n) \| q(x_n) + D(p(x_{n+1} | x_n) \| q(x_{n+1} | x_n))$$

=0

$$= D(p(x_{n+1}) \| q(x_{n+1})) + D(p(x_n | x_{n+1}) \| q(x_n | x_{n+1}))$$

条件概率密度函数 $p(x_{n+1}|x_n)$ 和 $q(x_{n+1}|x_n)$ 都是 $r(x_{n+1}|x_n)$ > 0

2. 在 n 时刻状态空间上的分布 μ_n 与平稳分布 μ 之间的相对熵 $D(\mu_n \| \mu)$ 随 n 递减

在上面的讨论中 μ_n' 是 n 时刻状态空间上的分布，

若设 μ_n' 是任意平稳分布 μ

那么下一时刻的分布 μ'_{n+1} 也为 μ。因而，

$$D(\mu_n \| \mu) \geqslant D(\mu_{n+1} \| \mu)$$

随着时间的流逝，状态分布将会愈来愈接近于每个平稳分布。

3. 若平稳分布是均匀分布，则熵增加

一般来说，相对熵减小并不表示熵增加　　非均匀平稳分布的Markov链

如果马尔可夫链的初始状态服从均匀分布

即已经是最大熵分布　　随着时间的推移，逐渐趋于平稳分布，

此平稳分布的熵必定低于均匀分布的熵。因而，熵随着时间而减少

如果平稳分布是均匀分布，则可将相对熵表示为

$$D(\mu_n \| \mu) = \log|\mathcal{X}| - H(\mu_n) = \log|\mathcal{X}| - H(X_n)$$　　单调递减

此时，相对熵的单调递减蕴含了熵的单增性　　H 递增

统计热力学：微观状态都是等可能发生的　　解释了为什么用状态数定义系统熵

4. 对于平稳的马尔可夫过程，条件熵 $H(X_n|X_1)$ 随 n 递增

$$H(X_n|X_1) \geqslant H(X_n|X_1,X_2) \quad \text{（条件作用使熵减小）}$$
$$= H(X_n|X_2) \quad \text{（由马尔可夫性）}$$
$$= H(X_{n-1}|X_1) \quad \text{（由平稳性）}$$

数据处理不等式应用于马尔可夫链 $X_1 \to X_{n-1} \to X_n$，则有

$$I(X_1;X_{n-1}) \geqslant I(X_1;X_n)$$

$$H(X_{n-1}) - H(X_{n-1}|X_1) \geqslant H(X_n) - H(X_n|X_1)$$

由平稳性，$H(X_{n-1}) = H(X_n)$ ⟶ $H(X_{n-1}|X_1) \leqslant H(X_n|X_1)$

洗牌操作增加熵

如果 T 是一副扑克牌的一次洗牌(置换)操作

X 表示这副牌的初始(随　机的)排列，假定洗牌操作 T 的选取独立于 X

$$H(TX)\geqslant H(X)$$

其中 TX 表示由洗牌 T 作用于初始排列 X 而获得的新排列

注意：T 作为一个操作，等价一个变换

对于洗牌操作 T 的任何分布和扑克牌的排列 X 的任意分布，有

$$\begin{aligned} H(TX) &\geqslant H(TX|T) \\ &= H(T^{-1}TX|T) \\ &= H(X|T) \\ &= H(X) \end{aligned}$$

思考：
这个是否与数据处理不等式的结论相矛盾？
差别在哪？

设 $X_1, X_2, \cdots, X_n, \cdots$ 为平稳马尔可夫链

设 $Y_i = \phi(X_i)$ 是一个随机过程 问题： 此时熵率 $H(\mathcal{Y})$ 为多少

难度： 新的随机过程可能不是一个马尔科夫链

好的思路不是利用定义1去处理，而是讨论 $H(Y_n \mid Y_{n-1}, \cdots, Y_1)$

引理：

$$H(Y_n \mid Y_{n-1}, \cdots, Y_2, X_1) \leqslant H(\mathcal{Y})$$

$$H(Y_n \mid Y_{n-1}, \cdots, Y_1) - H(Y_n \mid Y_{n-1}, \cdots, Y_1, X_1) \to 0$$

定理

解决了计算问题

若 $X_1, X_2, \cdots, X_n$ 构成平稳的马尔可夫链，且 $Y_i = \phi(X_i)$，那么

$$H(Y_n \mid Y_{n-1}, \cdots, Y_1, X_1) \leqslant H(\mathcal{Y}) \leqslant H(Y_n \mid Y_{n-1}, \cdots, Y_1)$$

$$\lim H(Y_n \mid Y_{n-1}, \cdots, Y_1, X_1) = H(\mathcal{Y}) = \lim H(Y_n \mid Y_{n-1}, \cdots, Y_1)$$

证明：对 $k=1,2,\cdots$，有

$$H(Y_n|Y_{n-1},\cdots,Y_2,X_1)$$

$$\overset{(a)}{=} H(Y_n|Y_{n-1},\cdots,Y_2,Y_1,X_1)$$

(a)成立是由于 Y_1 为 X_1 的函数

$$\overset{(b)}{=} H(Y_n|Y_{n-1},\cdots,Y_1,X_1,X_0,X_{-1},\cdots,X_{-k})$$

(b)可由 X 的马尔可夫性得到

$$\overset{(c)}{=} H(Y_n|Y_{n-1},\cdots,Y_1,X_1,X_0,X_{-1},\cdots,X_{-k},Y_0,\cdots,Y_{-k})$$

(c)由于 Y_i 为 X_i 的函数

$$\overset{(d)}{\leqslant} H(Y_n|Y_{n-1},\cdots,Y_1,Y_0,\cdots,Y_{-k})$$

(d)由于条件作用使熵减小

$$\overset{(e)}{=} H(Y_{n+k+1}|Y_{n+k},\cdots,Y_1)$$

(e)根据平稳性可得

由于对任意的 k，不等式成立

$$H(Y_n|Y_{n-1},\cdots,Y_1,X_1)\leqslant \lim_k H(Y_{n+k+1}|Y_{n+k},\cdots,Y_1)$$

$$=H(\mathcal{Y})$$

证明：上述区间长度可重新写为

$$H(Y_n|Y_{n-1},\cdots,Y_1)-H(Y_n|Y_{n-1},\cdots,Y_1,X_1)$$
$$=I(X_1;Y_n|Y_{n-1},\cdots,Y_1)$$

由互信息的性质，可得

$$I(X_1;Y_1,Y_2,\cdots,Y_n)\leqslant H(X_1)$$

且 $I(X_1;Y_1,Y_2,\cdots,Y_n)$ 随 n 递增

因此，$\lim I(X_1;Y_1,Y_2,\cdots,Y_n)$ 存在且满足

$$\lim_{n\to\infty} I(X_1;Y_1,Y_2,\cdots,Y_n)\leqslant H(X_1)$$

$$H(X_1)\geqslant \lim_{n\to\infty} I(X_1;Y_1,Y_2,\cdots,Y_n)$$
$$=\lim_{n\to\infty}\sum_{i=1}^{n} I(X_1;Y_i\mid Y_{i-1},\cdots,Y_1)$$
$$=\sum_{i=1}^{\infty} I(X_1;Y_i|Y_{i-1},\cdots,Y_1)$$

且每一项均为非负值，则其通项必趋向于 0

$$\lim I(X_1;Y_n|Y_{n-1},\cdots,Y_1)=0$$

综合可得

$$H(Y_n|Y_{n-1},\cdots,Y_1,X_1)\leqslant H(\mathcal{Y})\leqslant H(Y_n|Y_{n-1},\cdots,Y_1)$$

$$\lim H(Y_n|Y_{n-1},\cdots,Y_1,X_1)=H(\mathcal{Y})=\lim H(Y_n|Y_{n-1},\cdots,Y_1)$$

通过添加初始条件，可以完成计算

定义新过程 $Y_1, Y_2, \cdots, Y_n$，其中每个 Y_i 服从 $p(y_i|x_i)$

$$p(x^n, y^n) = p(x_1)\prod_{i=1}^{n-1} p(x_{i+1}|x_i)\prod_{i=1}^{n} p(y_i|x_i)$$

这样的过程称为隐马尔可夫模型(HMM)

它已广泛应用于语音识别、手写体识别等

在现代的机器学习方法中也可引入类似的模型加以研究；
只要有马尔科夫建模，增加新任务，就可采用类似的模型
也可沿用一些熵率计算技巧

小结

- 这部分主要就是解决了马尔科夫链的熵率计算问题
- 实际上就是一类特定的随机过程的熵率如何计算；
- 也谈到一个处理技巧：增加条件量，起到控制条件熵的目的