《信息论与编码理论基础》
第一章
熵与互信息

马猛 副教授
北京大学信息科学技术学院

提纲

- 一、渐进均分性
- 二、熵
- 三、互信息

一、 渐进均分性

- **大数定理**：针对独立同分布随机变量，当n很大时$\frac{1}{n}\sum_1^n X_i$近似于期望值$\mathrm{E}(X)$
- **渐进均分性（AEP）定理**：若$X_1, X_2, \cdots, X_n$为i.i.d~$p(x)$，则

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

  - 渐进均分性可以概括为“几乎一切事件都令人同等意外”
- 证明：由于X_i为i.i.d，独立随机变量的函数依然是独立同分布的，因而$\log p(X_i)$也是i.i.d. 因此，由大数定理得：

$$-\frac{1}{n}\log p(X_1, X_2, \cdots, X_n) = -\frac{1}{n}\sum_i \log p(X_i)$$

$$\text{依概率} \to E[\log p(X_i)] \triangleq H(X)$$

一、渐进均分性

- 考虑n个独立同分布（i.i.d.）的随机变量构成的序列$X_1, X_2, \cdots, X_n$，我们将全体可能出现的序列组成的集合划分为为两个子集，其一是典型集；其余为非典型集。

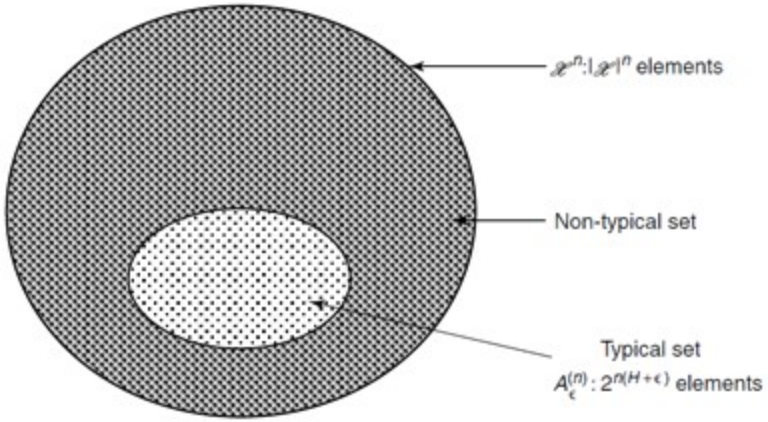


FIGURE 3.1. Typical sets and source coding.

- **典型集(typical set)定义**：关于$p(x)$的典型集$A_\varepsilon^{(n)}$是序列$(x_1, x_2, \cdots, x_n) \in X^n$的集合，且满足性质：

$$2^{-n(H(X)+\varepsilon)} \leq p(x_1, x_2, \cdots, x_n) \leq 2^{-n(H(X)-\varepsilon)}$$

其中$H(X) = -\sum_{x\epsilon X} p(x)\log p(x)$

- 说明：典型集中的序列的概率趋近于

$$2^{-nH(X)} = 2^{\Sigma_{x\epsilon X}\, np(x)\log p\,(x)} = \prod_{x\epsilon X} p(x)^{np(x)}$$

一、渐进均分性

- *例题1*：二元随机过程$P_r(x=1)=p=0.75$，$P_r(x=0)=q=0.25$。序列(1,1,1,0)是否属于典型集？(0,0,0,0)是否属于典型集？（$\varepsilon=0.0001$）

解：$p(1,1,1,0)=p^3q^1$，$p(0,0,0,0)=p^0q^4$

$$2^{-nH(X)}=\prod_{x\epsilon X}p(x)^{np(x)}=p^{np}q^{nq}=p^3q^1$$

序列(1,1,1,0)属于典型集,(0,0,0,0)不属于典型集

- *例题2*：二元随机过程$P_r(x=1)=0.5$，$P_r(x=0)=0.5$。序列(1,0,1,0)是否属于典型集？(1,1,1,1)是否属于典型集？（$H(X)=1$，$\varepsilon=0.0001$）

解：$p(1,0,1,0)=p(1,1,1,1)=0.5^4$

$$2^{-nH(X)}=p^{np}q^{nq}=0.5^20.5^2$$

序列(1,0,1,0)和(1,1,1,1)都属于典型集

一、渐进均分性

典型集的性质1：当n充分大时，$\Pr\left\{A_\varepsilon^{(n)}\right\} > 1-\varepsilon$

说明：该性质可由渐进均分定理直接得到。

- *习题3.13*（典型集的计算）考虑独立同分布（i.i.d.）的二值随机变量序列$X_1, X_2, \cdots, X_n$，其中$X_i = 1$的概率为0.6，如果$n = 25$，各种情况的数值结果如右表所示（第一列为n长序列中取值为1的元素个数，第二列为有k个元素取1的序列个数（即从n个不同的元素中取k个的组合数），第三列为有k个元素取1的事件概率），取$\varepsilon = 0.1$，哪些序列落入典型集$A_\varepsilon^{(n)}$中？典型集概率为多大？典型集中有多少个元素

解： $H(X)$ =0.970951，典型集序列包含$n = 11,12,\cdots,19$的所有序列。

典型集概率为0.936，典型集元素个数26366510

k	$\binom{n}{k}$	$\binom{n}{k}p^k(1-p)^{n-k}$	$-\frac{1}{n}\log p(x^n)$
0	1	0.0000	1.321928
1	25	0.0000	1.298530
2	300	0.0000	1.275131
3	2300	0.0000	1.251733
4	12650	0.0000	1.228334
5	53130	0.0000	1.204936
6	177100	0.0002	1.181537
7	480700	0.0009	1.158139
8	1081575	0.0031	1.134740
9	2042975	0.0088	1.111342
10	3268760	0.0212	1.087943
11	4457400	0.0434	1.064545
12	5200300	0.0760	1.041146
13	5200300	0.1140	1.017748
14	4457400	0.1465	0.994349
15	3268760	0.1612	0.970951
16	2042975	0.1511	0.947552
17	1081575	0.1200	0.924154
18	480700	0.0800	0.900755
19	177100	0.0442	0.877357
20	53130	0.0199	0.853958
21	12650	0.0071	0.830560
22	2300	0.0019	0.807161
23	300	0.0004	0.783763
24	25	0.0000	0.760364
25	1	0.0000	0.736966

一、 渐进均分性

- **典型集的性质2：典型集$A_\varepsilon^{(n)}$中的任何事件都是等概率出现的**

说明：这可由典型集的定义直接得到，也可称为“几乎一切事件都令人同等的意外”

- 例题3：一副扑克牌有52张，从中随机地抽取一张，然后放回去洗牌再抽，这样共抽取了100次。求每次都抽到红心A的概率，这个概率比其它任何一种“随机”的概率小吗？

解：100个红心A出现的概率为$\frac{1}{52^{100}}$

一、渐进均分性

- *例题4*（典型集元素的概率计算）：设x是一个离散型随机变量，有k个取值，概率密度函数为$p_i = Pr(X = x_i)$，求典型集$A_\varepsilon^{(n)}$中元素的个数以及每个元素出现的概率？

解：我们考虑$\varepsilon = 0$情况下的序列个数，可以通过组合数的连乘计算如下：

$$\left|A_\varepsilon^{(n)}\right| \to C_n^{m_1} C_{n-m_1}^{m_2} \cdots C_{m_K}^{m_K} = \frac{n!}{m_1!\,(n-m_1)!} \frac{(n-m_1)!}{m_2!\,(n-m_1-m_2)!} \cdots \frac{m_K!}{m_K!}$$

$$= \frac{n!}{m_1!\,m_2!\,m_K!}$$

根据斯特令（Stirling）公式：$\ln n! \xrightarrow[n\to\infty]{} n\ln n$

因此：$\frac{1}{n}\ln\left|A_\varepsilon^{(n)}\right| = \frac{1}{n}\left(\ln n! - \sum_{i=1}^{k}\ln m_i!\right) = \sum_{i=1}^{k}\frac{m_i}{n}\ln\frac{n}{m_i} = -\sum_{i=1}^{k} p_i \ln p_i$

每个元素出现的概率$p(X_1, X_2, \cdots, X_n) = \frac{1}{\left|A_\varepsilon^{(n)}\right|} = e^{n\sum_{i=1}^{k} p_i \ln p_i}$

补充说明：当$\varepsilon > 0$时的典型集中元素个数相比$\varepsilon = 0$时的典型集中元素个数的增长倍数至多是关于n的多项式（详见型方法），在$n \to \infty$情况下对结果的影响趋近于0

一、渐进均分性

- 定义“熵”概念的必要性：
  - 在后面将要介绍的许多重要的分析场合中，我们需要在n很大时对典型集的数量$\left|A_\varepsilon^{(n)}\right|$进行描述，由例题4给出的描述$\left|A_\varepsilon^{(n)}\right| = e^{-n\sum_{i=1}^{k} p_i \ln p_i}$形式复杂且数值巨大。
  - 因为$\left|A_\varepsilon^{(n)}\right|$是随$n$指数增长的，为便于我们表达，可以取对数变为随$n$线性增长，增长的速率称为熵率表示为：$\frac{1}{n}\log\left|A_\varepsilon^{(n)}\right| = -\sum_{i=1}^{k} p_i \ln p_i$
- **熵（entropy）的定义**：一个离散型随机变量X 的熵$H(X)$定义为

$$H(X) = -\sum_{x\in X} p(x)\log p(x)$$

其中对数log所用的底是2时，熵的单位为比特（bit）；如果使用底为e时，熵的单位是奈特（nat）

- “熵”这个度量本质是字母表（事件集合）大小的对数

“熵”的由来

- 熵（希腊语：entropia 英语：entropy）的概念是由德国物理学家克劳修斯于1865年所提出。在希腊语源中意为“内在”，即“一个系统内在性质的改变”，公式中一般记为S。

- 1923年，德国科学家普朗克（Plank）来中国讲学用到entropy这个词，胡刚复教授翻译时，因为熵是Q除以T（温度）的商数，于是把“商”字加火旁来意译“entropy”这个字，创造了“熵”字，（音读同：商）。

- 香侬最初想用“信息”来表达这一概念，但这个词当时已经被用滥了。后来，他决定用“不确定性”（uncertainty）来表达这个意思。当他和冯·诺依曼讨论这个问题时，冯·诺依曼对香侬建议说：你应该把它称之为“熵”。

一、渐进均分性

- 比特与奈特单位的换算
  - 如果一个随机变量的熵为x bit，或者y nat。由熵的本质是字母表大小的度量这一概念，应满足：

$$\left|A_\varepsilon^{(n)}\right| = 2^{nx} = e^{ny}$$

因此：$x = \frac{1}{n}\log_2 e^{ny} = \frac{\log_e e^{ny}}{n\log_e 2} = \frac{y}{\log_e 2}$

  - 由上式可得：1bit=0.693nat

- 典型集性质总结：
  - 1. 典型集的概率近似为1
  - 2. 典型集中的所有元素几乎是等可能的
  - 3.（由上面两条可知）典型集的元素个数近似等于2^{nH}

一、渐进均分性

- 关于信息量与典型集数量间关系的一些说明：
  - 典型集中的每个事件都带给我们同等的“意外”。
  - 我们可以定义每个事件都带给我们的信息量为：$\left|A_\varepsilon^{(n)}\right|$个等概率事件中的一个发生了。
  - 信息量可以理解为事件出现带给人的“意外”程度。
  - 如100个红心A事件出现了，它带给我们的信息量为$\frac{1}{52^{100}}\times 100\%$的事件发生了。通过对典型集中的每个事件赋予某个特定含义（编码），可以从每个具体事件获得信息量为$\frac{1}{52^{100}}\times 100\%$的具有某种含义的信息（译码）。

二、熵

- X的熵可以解释为随机变量$\log\frac{1}{p(X)}$的期望值，即$H(X)=\mathrm{E}_p\log\frac{1}{p(X)}$
- **熵的性质1：$H(X)\geq 0$**

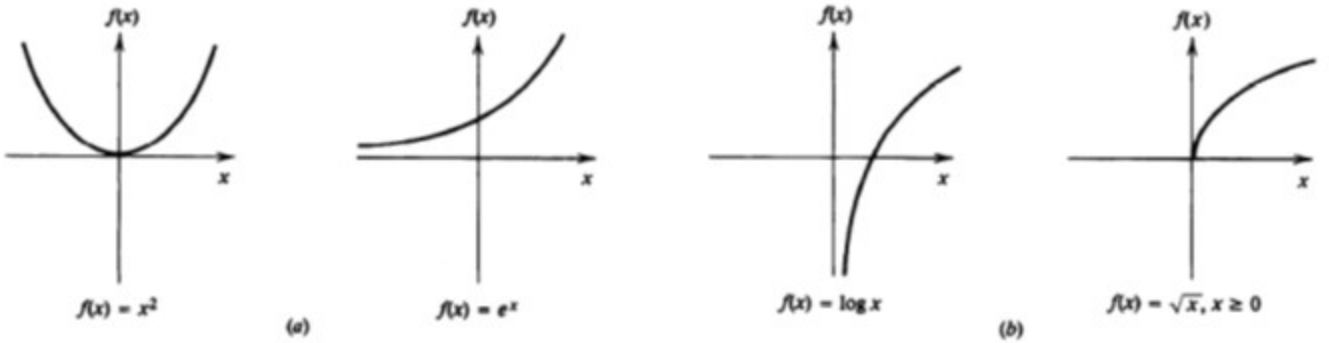
证明：对于**离散**随机变量，由$0\leq p(x)\leq 1$知$\log\frac{1}{p(x)}\geq 0$

- **熵的性质2：熵的凹性：$H(p)$是关于p的凹函数**

说明：即对于分布p_1和p_2有：$H(\lambda p_1+(1-\lambda)p_1)\geq\lambda H(p_1)+(1-\lambda)H(p_2)$

本课程定义凹函数与英文concave function一致，与中文凹的含义相反

证明略。



凸函数与凹函数的例子（a）凸函数（b）凹函数

二、熵

- *例2.1.1* 二元随机变量的熵，设

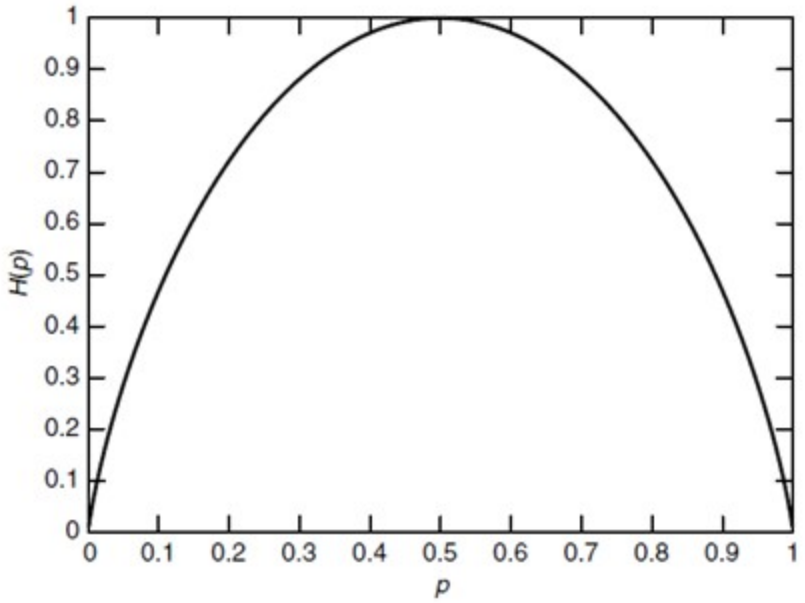
$$X=\begin{cases}1 & \text{概率为}p\\ 0 & \text{概率为}1-p\end{cases}$$

于是$H(X)=-p\log p-(1-p)\log(1-p)\triangleq H(p)$

当$p=0.5$时，$H(X)=1$比特。不确定度最大，此时熵也最大。

当$p=0$或1时，$H(X)=0$。因为此时变量不再是随机的，从而不具有不确定度。

下图显示了熵的凹性



二、熵

- 例2.1.2设

$$X=\begin{cases} a & \text{概率为}0.5 \\ b & \text{概率为}0.25 \\ c & \text{概率为}0.125 \\ d & \text{概率为}0.125 \end{cases}$$

则X的熵为$H(X)=-\frac{1}{2}\log\frac{1}{2}-\frac{1}{4}\log\frac{1}{4}-\frac{1}{8}\log\frac{1}{8}-\frac{1}{8}\log\frac{1}{8}=\frac{7}{4}$比特

通过二元问题确定变量X的值，需要最少如下三个问题：1. $X=a$吗？2. $X=b$吗？3. $X=c$吗？

所需的二元问题数目的期望值为$0.5\times1+0.25\times2+0.23\times3=1.75$。可以证明这是为确定变量$X$的值所需的二元问题的最小期望值，即描述$X$最少需要1.75比特。这将在下一章信源编码中介绍。

二、熵

- 定理2.6.4：$H(X) \le \log|\chi|$，其中$|\chi|$表示X的字母表χ中元素的个数，当且仅当服从均匀分布，等号成立。

- 证明：设$u(x) = \frac{1}{|\chi|}$为χ上均匀分布的概率密度函数，是随机变量的概率密度函数。于是

$$\log|\chi| - H(X) = \sum p(x) log \frac{p(x)}{u(x)} \triangleq D(p\|u)$$

$$D(p\|u) = -\sum p(x) log \frac{u(x)}{p(x)} \ge -log \sum p(x) \frac{u(x)}{p(x)} = log \sum u(x) = 0$$

- 上面的不等式利用了$\log(x)$是凹函数
- 上述定理也可通过观察文氏图3.1获得：仅当典型集与全体序列集合重合时典型集中元素个数最多，此时服从均匀分布。
- 等概率分布时熵最大（对于QAM信号，各星座图点为何种分布时信号的熵最大？）
- **$D(p\|u) \ge 0$称为相对熵（relative entropy）**

二、熵

- **联合熵（joint entropy）定义：**对于服从联合分布为$p(x,y)$的一对离散随机变量(X,Y)，其联合熵$H(X,Y)$定义为：

$$H(X,Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

  - 联合熵是对单个随机变量的熵推广到两个随机变量情形下的定义。
- **条件熵（conditional entropy）定义：**

$$H(Y|X) = -\sum_{x\in X} p(x)H(Y|X=x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$$

  - 该定义是一个随机变量在给定另一随机变量下的条件熵，它是条件分布熵关于起条件作用的那个随机变量取平均之后的期望值。

二、熵

- **定理2.2.1**（链式法则）

$$\boldsymbol{H(X,Y) = H(X) + H(Y|X)}$$

证明：$H(X,Y)$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x,y)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x)p(y|x)$$

$$= -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$$

$$= -\sum_{x\in X} p(x)\log p(x) - \sum_{x\in X}\sum_{y\in Y} p(x,y)\log p(y|x)$$

$$= H(X) + H(Y|X)$$

信息熵与热力学熵的关系

- 热力学第二定律（又称熵增定律）：对于孤立系统，熵永远增加。
- 麦克斯韦妖（Maxwell's demon）：在一含有气体的容器里中间加一隔板，在隔板中间开一小门，由一小精灵把守。他让快速的气体分子从一个方向通过小门，慢速的分子则从另一方向通过小门，这样小精灵就把气体按速度分为两部分，于是该容器系统的热熵减小了。
- 一种学术观点解释为：热熵的减少相当于信息的增加，所以信息熵相当于负热熵。
- 修改的第二形式热力学定律：在任何系统中，热熵和信息熵的总和是恒定的。

三、互信息

- **互信息（mutual information）定义：**考虑两个随机变量X和Y，它们的联合概率密度函数为$p(x,y)$，其边际概率密度函数分别是$p(x)$和$p(y)$，互信息为

$$I(X;Y) = -\sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

- 熵与互信息的关系

$$I(X;Y) = H(X) - H(X|Y)$$
$$I(X;Y) = H(Y) - H(Y|X)$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$
$$I(X;Y) = I(Y;X)$$
$$I(X;X) = H(X)$$

- $2^{nH(X)}$是X的字符集个数，$2^{nH(X|Y)}$是给定Y后X的字符集的个数，$2^{nI(X;Y)}$是给定Y后X的字符集个数的减少量
- 互信息是一个随机变量包含另一个随机变量信息量的度量；它是在给定另一个随机变量知识的条件下，原随机变量不确定度的缩减

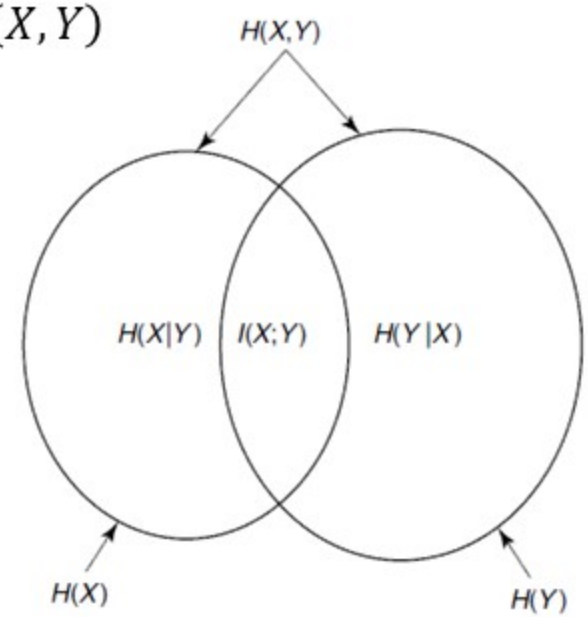


FIGURE 2.2. Relationship between entropy and mutual information.

三、互信息

- **定理：（互信息的非负性）对任意两个随机变量X和Y，**

$$I(X;Y) \geq 0$$

当且仅当X与Y互相独立*，等号成立。

- 上述定理可以通过观察文氏图2.2获得。
- **定理：（条件作用使熵减小）（信息不会有负面影响）**

$$H(X|Y) \leq H(X)$$

  - 说明：给定Y后X的字符集的个数可能减小或不变（当Y与X 独立时），但不会增加。

*注:如果$p(X;Y) = p(X)p(Y)$则称X与Y互相独立

三、互信息

- **马尔可夫（Markov）链定义：**如果Z的条件分布仅依赖于Y的分布，而（在给定Y时）与X是条件独立的，则称随机变量X,Y,Z依序构成马尔可夫链（记为$X \to Y \to Z$）。若X,Y,Z的联合概率密度函数可写为

$$p(x,y,z) = p(x)p(y|x)p(z|y)$$

则X,Y,Z构成马尔可夫链$X \to Y \to Z$

- **定理2.8.1（数据处理不等式）若$\boldsymbol{X \to Y \to Z}$，则有$\boldsymbol{I(X;Y) \geq I(X;Z)}$**

证明：$I(X;Y) - I(X;Z) = [H(X) - H(X|Y)] - [H(X) - H(X|Z)] = H(X|Z) - H(X|Y) = [H(X|Z) - H(X|Y,Z)] - [H(X|Y) - H(X|Y,Z)] = I(X;Y|Z) - I(X;Z|Y) = I(X;Y|Z) \geq 0$

在给定Y的情况下，X与Z是条件独立的，因此$I(X;Z|Y) = 0$

- 这说明对数据Y的后处理不会增加关于X的信息量。

类似的可以证明$I(Y;Z) \geq I(X;Z)$，即预处理也不会增加互信息。

主要知识点

- 典型集概念
- 典型集的性质（1）（2）（3）
- 熵的概念和性质
- 渐进均分性（AEP）定理
- 条件熵、联合熵
- 链式法则
- 互信息
- 马尔可夫链
- 数据处理不等式

习题

- 阅读教材2.1-2.4，2.8，3.1 3.2
- 证明例题2.1.1中当$p = 0.5$时熵最大 $\max H(p) = 1$
- 2.3，2.11，2.12，2.29（a），3.7（a，b）