信息论第二讲作业解答

中国科学技术大学《信息论 A》006125.01 班助教组

2025 年 3 月 31 日

第 1 题

For random variables X and Y, prove that $H(X+Y) \le H(X) + H(Y)$ holds.

证明:

$$\begin{aligned}
H(X+Y) &\le H(X+Y) + H(X,Y|X+Y) \\
&= H(X,Y,X+Y) \\
&= H(X,Y) + H(X+Y|X,Y) \\
&= H(X,Y) \\
&\le H(X) + H(Y). \qquad \square
\end{aligned}$$

注 1. 两处不等号取等条件分别为 $X+Y$ 到 (X,Y) 是单射, 以及 X,Y 相互独立. 以及此处的 $X+Y$ 可以替换为 $f(X,Y)$, 因为我们总是有 $H(X,Y) \ge H(f(X,Y))$.

第 2 题

For random variables X_1, X_2, …, X_n, Y_1, Y_2, …, Y_n, when does

$$H(X_1, X_2, \ldots, X_n | Y_1, Y_2, \ldots, Y_n) = H(X_1|Y_1) + H(X_2|Y_2) + \cdots + H(X_n|Y_n)$$

hold?

解: 我们先定义几个命题:

1. $H(X_1, X_2, \cdots, X_n|Y_1, Y_2, \cdots, Y_n) = H(X_1|Y_1) + H(X_2|Y_2) + \cdots + H(X_n|Y_n)$;
2. $H(X_1|Y_1, Y_2, \cdots, Y_n) = H(X_1|Y_1)$ 且对每个 $i \in \{2, 3, \cdots, n\}$ 有

$$H(X_i|X_{i-1}, \cdots, X_1, Y_1, Y_2, \cdots, Y_n) = H(X_i|Y_i);$$

3. 对所有 x_1, y_1, y_2, $\cdots$, y_n, 只要

$$P_{X_1|Y_1,Y_2,\cdots,Y_n}(x_1|y_1,y_2,\cdots,y_n) = P_{X_1|Y_1}(x_1|y_1) \tag{1}$$

   中条件的概率大于 0, 1 式就成立; 对所有 $i \in \{2,3,\cdots,n\}$ 和 x_1, x_2, $\cdots$, x_i, y_1, y_2, $\cdots$, y_n, 只要

$$P_{X_i|X_{i-1},\cdots,X_1,Y_1,Y_2,\cdots,Y_n}(x_i|x_{i-1},\cdots,x_1,y_1,y_2,\cdots,y_n) = P_{X_i|Y_i}(x_i|y_i) \tag{2}$$

   中条件的概率大于 0, 2 式就成立;

4. 对所有 x_1, x_2, $\cdots$, x_n, y_1, y_2, $\cdots$, y_n, 如果 $P_{Y_1,Y_2,\cdots,Y_n}(y_1,y_2,\cdots,y_n) > 0$ 则

$$P_{X_1,X_2,\cdots,X_n|Y_1,Y_2,\cdots,Y_n}(x_1,x_2,\cdots,x_n|y_1,y_2,\cdots,y_n) = \prod_{i=1}^{n} P_{X_i|Y_i}(x_i|y_i).$$

因为

$$\begin{aligned} & H(X_1,X_2,\cdots,X_n|Y_1,Y_2,\cdots,Y_n) \\ = {} & H(X_1|Y_1,Y_2,\cdots,Y_n) + \sum_{i=2}^{n} H(X_i|X_{i-1},\cdots,X_1,Y_1,Y_2,\cdots,Y_n), \end{aligned} \tag{3}$$

$H(X_1|Y_1,Y_2,\cdots,Y_n) \le H(X_1|Y_1)$, 对每个 $i \in \{2,3,\cdots,n\}$ 有

$$H(X_i|X_{i-1},\cdots,X_1,Y_1,Y_2,\cdots,Y_n) \le H(X_i|Y_i),$$

所以命题 1 等价于命题 2. 用讲义的 Theorem 3.6 可以证明命题 2 等价于命题 3. 用形式与 3 式相同的概率的链式法则可以证明命题 3 等价于命题 4. 因此命题 1 等价于命题 4. □

第 3 题

We know from Theorem 3.6 that conditioning reduces entropy. For mutual information $I(X;Y)$ and conditional mutual information $I(X;Y|Z)$, does an analogous property hold?

解: 设 X 和 Y 独立, 取值于 $\{0,1\}$ 且都以 1/2 的概率取 1. 定义随机变量 $Z = X \oplus Y$. 这样 $I(X;Y) = 0 < 1 = I(X;Y|Z)$.

设随机变量 X, Z, Y 都取值于 $\{0,1\}$ 且形成一条 Markov 链, $P_X(1) = 1/2$, 对所有 x, $z \in \{0,1\}$ 有

$$P_{Z|X}(z|x) = \begin{cases} 0.9, & z = x \\ 0.1, & z \neq x \end{cases},$$

对所有 $z, y \in \{0,1\}$ 有

$$P_{Y|Z}(y|z) = \begin{cases} 0.9, & y = z \\ 0.1, & y \neq z \end{cases}$$

这样 $I(X;Y) > 0 = I(X;Y|Z)$. □

有同学把 $I(X;Y)$ 写成了 $I(x;y)$. 注意大写字母和小写字母的含义是不同的.

有同学给出的理由是: $I(X;Y;Z) = I(X;Y) - I(X;Y|Z)$, 而 $I(X;Y;Z)$ 可能大于 0 也可能小于 0. 如果要用这个结论, 我们需要先证明它. 最好不要用这个结论, 因为我们没有定义过 $I(X;Y;Z)$.

第 4 题

Using the non-negativity of relative divergence, prove the log-sum inequality: for non-negative numbers $\{a_i\}_{i=1,\ldots,n}$ *and* $\{b_i\}_{i=1,\ldots,n}$,

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b},$$

where $a = \sum_{i=1}^n a_i$, $b = \sum_{i=1}^n b_i$, *with equality holding if and only if there exists* c *such that* $a_i = cb_i$ *for all* i.

证明: 用 $p(i) = a_i/a$ 和 $q(i) = b_i/b$ 定义 $\{1, 2, \cdots, n\}$ 上的概率函数 p 和 q. 这样

$$0 \leq D(p\|q) = \sum_{i=1}^{n} \frac{a_i}{a} \log_2 \left(\frac{a_i/a}{b_i/b} \right) = \frac{1}{a} \sum_{i=1}^{n} a_i \log_2 \left(\frac{a_i}{b_i} \right) + \log_2 \left(\frac{b}{a} \right), \tag{4}$$

所以

$$\sum_{i=1}^{n} a_i \log_2 \left(\frac{a_i}{b_i} \right) \geq a \log_2 \left(\frac{a}{b} \right). \tag{5}$$

如果 5 式成立等号, 则 4 成立等号, $p = q$, 对每个正整数 $i \leq n$ 有 $a_i = (a/b)b_i$. 反之, 如果存在 c 使 $a_i = cb_i$ 对每个正整数 $i \leq n$ 成立, 则 $a = \sum_{i=1}^n a_i = c\sum_{i=1}^n b_i = cb$,

$$\sum_{i=1}^{n} a_i \log_2 \left(\frac{a_i}{b_i} \right) = \sum_{i=1}^{n} a_i \log_2(c) = a \log_2 \left(\frac{a}{b} \right).$$

□

注 2. 也可以用 $x \log x$ 的凹凸性和 *Jensen* 不等式或者 $ln(1+x) \leq x$ 证明，但此处题目要求使用相对熵的非负性质。

第 5 题

In this exercise we apply Corollary 3.2 to a guessing problem due to James L. Massey [13]. Suppose that we want to guess the value of a random variable X over $X(\Omega) = \{1, 2, \dots\}$. How many times do we need to guess, on average? Without loss of generality, we can always relabel the random variable so that $P_X(1) \geq P_X(2) \geq \dots$ holds. Prove that, on average, we need to guess no less than $e^{H(X)-1}$ times, where the unit of entropy is nat.

证明: 为了使猜的平均次数最小, 我们第 1 次应该猜 1, 第 2 次应该猜 2, 以此类推. 此时猜的平均次数为 $\sum_{x=1}^{\infty} x P_X(x) = \mathbf{E}[X]$. 用讲义推论 3.2 和对每个正数 t 成立的不等式 $\ln(t) \leq t - 1$,

$$\begin{aligned} H(X) &\leq \mathbf{E}[X] \ln(\mathbf{E}[X]) - (\mathbf{E}[X] - 1) \ln(\mathbf{E}[X] - 1) \\ &= \ln(\mathbf{E}[X]) + (\mathbf{E}[X] - 1) \ln\left(\frac{\mathbf{E}[X]}{\mathbf{E}[X] - 1}\right) \\ &\leq \ln(\mathbf{E}[X]) + (\mathbf{E}[X] - 1) \frac{1}{\mathbf{E}[X] - 1} \\ &= \ln(\mathbf{E}[X]) + 1, \end{aligned}$$

所以猜的平均次数的最小值 $\mathbf{E}[X] \geq e^{H(X)-1}$. □

第 6 题

Consider a random variable X over $X(\Omega) = \{1, 2, \dots\}$.

- *a) Prove that if $\mathbf{E}X$ is finite, then $H(X)$ is also finite.*
- *b) Prove that if $\mathbf{E}\log X$ is finite, then $H(X)$ is also finite.*
- *c) Prove that if $H(X)$ is finite and $P_X(x)$ is monotonically non-increasing with x, then $\mathbf{E}\log X$ is finite.*
- *d) Give an example to illustrate that the monotonically non-increasing condition of $P_X(x)$ in the previous statement is necessary.*

a) 证明: 如果 $\mathbf{E}[X] = 1$, 则 X 以概率 1 取 1, $H(X) = 0$. 如果 $\mathbf{E}[X] > 1$, 则根据讲义推论 3.2,

$$H(X) \leq \mathbf{E}[X] \log_2(\mathbf{E}[X]) - (\mathbf{E}[X] - 1) \log_2(\mathbf{E}[X] - 1) < \infty. \qquad \square$$

b) 证明: 由讲义推论 3.6 得 $H(\log_2(X)) = H(X)$. 如果 $\mathbf{E}[\log_2(X)]$ 有限, 则 $H(\log_2(X))$ 有限, $H(X)$ 有限.

另一种方法是, 对随机变量 X 构造另一概率分布 $Q_X(x)=\frac{6}{\pi^2 x^2}$, 可以验证 $\sum_1^\infty Q_X(x)=1$. 接着我们有：

$$\begin{aligned}
D(P_X||Q_X) &= \sum_{x=1}^{\infty} P_X(x)\log\left(\frac{P_X(x)}{Q_X(x)}\right)\\
&= -H(X)-\sum_{x=1}^{\infty} P_X(x)\log Q_X(x)\\
&= -H(X)+\log\frac{\pi^2}{6}+2\sum_{x=1}^{\infty} P_X(x)\log x\\
&= -H(X)+\log\frac{\pi^2}{6}+2\mathbf{E}\log X \geq 0.
\end{aligned}$$

所以我们可得，$H(X)\leq \log\frac{\pi^2}{6}+2\mathbf{E}\log X<\infty$. □

c) 证明: 对每个正整数 x, $xP_X(x)\leq \sum_{x'=1}^{x} P_X(x')\leq 1$, 所以 $x\leq 1/P_X(x)$. 这样

$$\mathbf{E}[\log_2(X)]=\sum_{x=1}^{\infty} P_X(x)\log_2(x)\leq \sum_{x=1}^{\infty} P_X(x)\log_2\left(\frac{1}{P_X(x)}\right)=H(X)<\infty. \qquad \square$$

d) 证明: 记 $A=\sum_{i=1}^{\infty}(1/i^2)$. 设对每个正整数 i, 随机变量 X 取 2^i 的概率是 $1/Ai^2$. 这样

$$H(X)=-\sum_{i=1}^{\infty}\frac{1}{Ai^2}\log_2\left(\frac{1}{Ai^2}\right)=\sum_{i=1}^{\infty}\frac{\log_2(A)+2\log_2(i)}{Ai^2}<\infty,$$

但

$$\mathbf{E}[\log_2(X)]=\sum_{i=1}^{\infty}\frac{1}{Ai^2}\log_2(2^i)=\sum_{i=1}^{\infty}\frac{1}{Ai}=\infty. \qquad \square$$

第 7 题

Consider a uniform random variable X over $\{0,1,\ldots,m-1\}$, and its observation Y is drawn uniformly from $\{(X-1) \bmod m, X, (X+1) \bmod m\}$. Define $P_e=P(Y\neq X)$.

- *a) Give a lower bound of P_e using the Fano inequality.*
- *b) Find the gap between the lower bound and the exact value of P_e of the MAP decision.*
- *c) Can you resolve the gap by inspecting the proof of the Fano inequality and improving it?*

这道题应该假设了 $m\geq 3$.

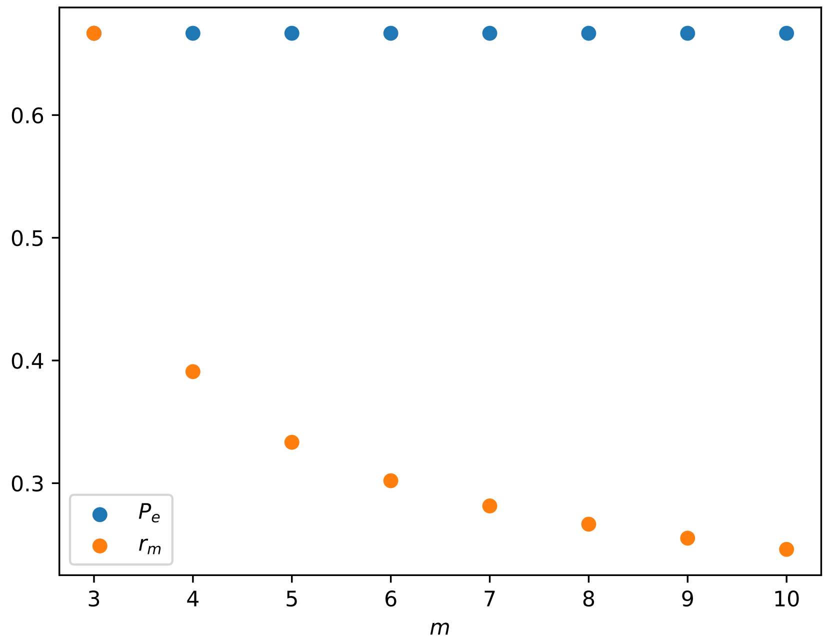

图 1: Fano 不等式给出的 P_e 的下界 r_m 和 P_e.

a) 解: 对每个自然数 $y < m$, 可以看出 X 在 $Y = y$ 的条件下服从 $\{(y-1) \bmod m, y, (y+1) \bmod m\}$ 上的均匀分布, 所以 $H(X|Y=y) = \log_2(3)$. 这样 $H(X|Y) = \log_2(3)$, 我们可以把 Fano 不等式写成

$$\log_2(3) \le h_2(P_e) + P_e \log_2(m-1).$$

所以 P_e 大于等于关于 p 的方程 $h_2(p) + p\log_2(m-1) - \log_2(3) = 0$ 的最小正根 r_m. □

b) 解: 可以看出 $P_e = 2/3$. 图 1 对比了 P_e 和 r_m. □

c) 解: 令随机变量 Z 在 $X \ne Y$ 时取 1, 否则取 0. 类似于讲义中 Fano 不等式的推导,

$$H(X|Y) = H(X|Y) + H(Z|X,Y) = H(X,Z|Y) = H(Z|Y) + H(X|Z,Y),$$
$$H(Z|Y) \le H(Z) = h_2(P_e).$$

对使 $P_{Z,Y}(0,y) > 0$ 的每个 y, X 在 $Z=0$ 且 $Y=y$ 的条件下以概率 1 取 y, 所以 $H(X|Z=0,Y=y) = 0$. 对使 $P_{Z,Y}(1,y) > 0$ 的每个 y, X 在 $Z=1$ 且 $Y=y$ 的条件下以概率 1 属于 $\{(y-1) \bmod m, (y+1) \bmod m\}$, 所以 $H(X|Z=1,Y=y) \le \log_2(2) = 1$. 这样

$$H(X|Z,Y) \le \sum_{y, P_{Z,Y}(1,y)>0} P_{Z,Y}(1,y) \cdot 1 = P_Z(1) = P_e,$$

$$H(X|Y) \le h_2(P_e) + P_e. \tag{6}$$

如果 $m=3$ 则 6 式就是 Fano 不等式. 如果 $m>3$ 则 6 式比 Fano 不等式紧. 由 6 式得 P_e 大于等于关于 p 的方程 $h_2(p)+p-\log_2(3)=0$ 的最小正根 2/3. □

第 8 题

Construct an example where equality holds in the Fano inequality.

解: 我们仔细考察 Fano 不等式的推导过程以及取等条件，如下：

$$\begin{aligned} H(X|\hat{X}) + H(E|X,\hat{X}) &= H(E|\hat{X}) + H(X|\hat{X},E) \\ H(X|\hat{X}) + 0 &\le h_2(P_e) + (1-P_e)\cdot 0 + P_e \cdot \log(|X(\Omega)|-1) \\ H(X|\hat{X}) &\le h_2(P_e) + P_e \log(|X(\Omega)|-1) \end{aligned}$$

从中可以看出，两处放缩分别使用

$$H(E|\hat{X}) \le H(E) = h_2(P_e), \qquad H(X|\hat{X},E=1) \le \log(|X(\Omega)|-1).$$

换言之，即:

- $\hat{X}$ 的具体值对译码错误率没有影响。
- 译码错误时，不论译为任何值，X 的分布总是均匀的。

基于此，我们构造如下两个例子：

例子一：设 $X \in \{1,2,3,...m\}$ 并且 $p_1 \ge p_2 \ge ... \ge p_m$，我们对 X 的最佳估计是 $\hat{X}=1$，此时产生的误差概率为 $p_e = 1-p_1$。此时 Fano 不等式变为:

$$h_2(p_e) + p_e \log(m-1) \ge H(X)$$

并且概率密度分布为：

$$(p_1,p_2,...,p_m) = \left(1-p_e, \frac{p_e}{m-1}, ..., \frac{p_e}{m-1}\right),$$

据此概率分布，我们可以计算出 $H(X) = h_2(p_e) + p_e \log(m-1)$, 此时等号成立, Fano 不等式是精确的。

例子二：设 $X \in \{1,2,3,...m\}$ 并且 $p_1 = p_2 = ... = p_m = \frac{1}{m}$，然后我们构造概率转移矩阵如下：

$$p_{Y|X}(y_j|x_i) = \begin{cases} \frac{p_e}{m-1} & \text{if } i \ne j \\ 1-p_e & \text{if } i = j \end{cases}.$$

并且在译码端保证 $\hat{X}=Y$。此时根据 $P_{X|\hat{X}}=\frac{P_X P_{\hat{X}|X}}{P_{\hat{X}}}=P_{Y|X}$ 有

$$\begin{aligned}H(X|\hat{X}) &= -\frac{p_e}{m-1}\log\frac{p_e}{m-1}(m-1)+(1-p_e)\log\frac{1}{1-p_e}\\ &= p_e\log\frac{1}{p_e}+(1-p_e)\log\frac{1}{1-p_e}+p_e\log(m-1)\\ &= h_2(p_e)+p_e\log(m-1).\end{aligned}$$

□

第 9 题

If the estimate $\hat{X}$ is a size-L subset of $X(\Omega)$, and define the error event to be $\{X\notin\hat{X}\}$, establish an extension of the Fano inequality.

解: 设 X 是离散随机变量, 正整数 $L<|X(\Omega)|$, $\hat{X}$ 是 $X(\Omega)$ 的一个随机的子集, $|X(\Omega)|=L$ 以概率 1 成立. 用 P_e 表示 $X\notin\hat{X}$ 的概率. 我们来证明

$$H(X|\hat{X})\le h_2(P_e)+(1-P_e)\log_2(L)+P_e\log_2(|X(\Omega)|-L). \tag{7}$$

令随机变量 Z 在 $X\notin\hat{X}$ 时取 1, 否则取 0. 类似于讲义中 Theorem 3.9 关于 Fano 不等式的推导,

$$H(X|\hat{X})=H(X|\hat{X})+H(Z|X,\hat{X})=H(X,Z|\hat{X})=H(Z|\hat{X})+H(X|Z,\hat{X}),$$
$$H(Z|\hat{X})\le H(Z)=h_2(P_e).$$

对使 $P_{Z,\hat{X}}(0,\hat{x})>0$ 的每个 $\hat{x}$, X 在 $Z=0$ 且 $\hat{X}=\hat{x}$ 的条件下以概率 1 属于有 L 个元素的集合 $\hat{x}$, 所以

$$H(X|Z=0,\hat{X}=\hat{x})\le\log_2(L).$$

对使 $P_{Z,\hat{X}}(1,\hat{x})>0$ 的每个 $\hat{x}$, X 在 $Z=1$ 且 $\hat{X}=\hat{x}$ 的条件下以概率 1 属于有 $|X(\Omega)|-L$ 个元素的集合 $X(\Omega)\setminus\hat{x}$, 所以

$$H(X|Z=1,\hat{X}=\hat{x})\le\log_2(|X(\Omega)|-L).$$

这样

$$\begin{aligned}H(X|Z,\hat{X}) &\le \sum_{\hat{x},P_{Z,\hat{X}}(0,\hat{x})>0}P_{Z,\hat{X}}(0,\hat{x})\log_2(L)+\sum_{\hat{x},P_{Z,\hat{X}}(1,\hat{x})>0}P_{Z,\hat{X}}(1,\hat{x})\log_2(|X(\Omega)-L|)\\ &= P_Z(0)\log_2(L)+P_Z(1)\log_2(|X(\Omega)|-L)\\ &= (1-P_e)\log_2(L)+P_e\log_2(|X(\Omega)|-L),\end{aligned}$$

7 式成立. □

第 10 题

Prove the Csiszár identity:

$$\sum_{i=1}^{n} I(X_{i+1},\dots,X_n;Y_i|Y_1,\dots,Y_{i-1}) = \sum_{i=1}^{n} I(Y_1,\dots,Y_{i-1};X_i|X_{i+1},\dots,X_n),$$

where X_{n+1} and Y_0 are understood as degenerated.

这样说可能更好理解: 如果 X_1, X_2, $\cdots$, X_n, Y_1, Y_2, $\cdots$, Y_n 是离散随机变量, X_{n+1} 和 Y_0 是常数, 则

$$\sum_{i=1}^{n} I(X_{i+1},\cdots,X_{n+1};Y_i|Y_0,\cdots,Y_{i-1}) = \sum_{i=1}^{n} I(Y_0,\cdots,Y_{i-1};X_i|X_{i+1},\cdots,X_{n+1}). \tag{8}$$

方法一:

$$\begin{aligned}
&\sum_{i=1}^{n} I(X_{i+1},\cdots,X_{n+1};Y_i|Y_0,\cdots,Y_{i-1}) \\
\overset{\text{(a)}}{=}\ &\sum_{i=1}^{n-1} I(X_{i+1},\cdots,X_n;Y_i|Y_0,\cdots,Y_{i-1},X_{n+1}) \\
\overset{\text{(b)}}{=}\ &\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(X_j;Y_i|Y_0,\cdots,Y_{i-1},X_{j+1},\cdots,X_{n+1}) \\
\overset{\text{(c)}}{=}\ &\sum_{j=2}^{n} \sum_{i=1}^{j-1} I(X_j;Y_i|Y_0,\cdots,Y_{i-1},X_{j+1},\cdots,X_{n+1}) \\
\overset{\text{(d)}}{=}\ &\sum_{j=2}^{n} I(X_j;Y_1,\cdots,Y_{j-1}|Y_0,X_{j+1},\cdots,X_{n+1}) \\
\overset{\text{(e)}}{=}\ &\sum_{j=1}^{n} I(X_j;Y_0,\cdots,Y_{j-1}|X_{j+1},\cdots,X_{n+1}).
\end{aligned}$$

上式第一行 $i=n$ 的一项等于 0, 对每个正整数 $i<n$ 有

$$I(X_{i+1},\cdots,X_{n+1};Y_i|Y_0,\cdots,Y_{i-1}) = I(X_{i+1},\cdots,X_n;Y_i|Y_0,\cdots,Y_{i-1},X_{n+1}),$$

所以 (a) 成立. 同理可得 (e). (b) 和 (d) 用了互信息的链式法则. 通过交换求和顺序可以得到 (c). □

方法二: 对每个正整数 $i \le n-1$, 我们可以以两种方式展开 $I(X_{i+1},\cdots,X_{n+1};Y_0,\cdots,Y_i)$ 得到

$$I(X_{i+1},\cdots,X_{n+1};Y_0,\cdots,Y_{i-1}) + I(X_{i+1},\cdots,X_{n+1};Y_i|Y_0,\cdots,Y_{i-1})$$

$$= I(X_{i+2},\cdots,X_{n+1};Y_0,\cdots,Y_i) + I(X_{i+1};Y_0,\cdots,Y_i|X_{i+2},\cdots,X_{n+1}). \tag{9}$$

因为 $i \in \{1,n\}$ 时 $I(X_{i+1},\cdots,X_{n+1};Y_0,\cdots,Y_{i-1}) = 0$, 所以

$$\begin{aligned}\sum_{i=1}^{n-1} I(X_{i+1},\cdots,X_{n+1};Y_0,\cdots,Y_{i-1}) &= \sum_{i=2}^{n} I(X_{i+1},\cdots,X_{n+1};Y_0,\cdots,Y_{i-1}) \\ &= \sum_{i=1}^{n-1} I(X_{i+2},\cdots,X_{n+1};Y_0,\cdots,Y_i).\end{aligned}$$

求 9 式对所有正整数 $i \leq n-1$ 的和得

$$\begin{aligned}\sum_{i=1}^{n-1} I(X_{i+1},\cdots,X_{n+1};Y_i|Y_0,\cdots,Y_{i-1}) &= \sum_{i=1}^{n-1} I(X_{i+1};Y_0,\cdots,Y_i|X_{i+2},\cdots,X_{n+1}) \\ &= \sum_{i=2}^{n} I(X_i;Y_0,\cdots,Y_{i-1}|X_{i+1},\cdots,X_{n+1})\end{aligned}$$

即 8 式. □

第 11 题

In this exercise, we provide an information-theoretic proof of the well known number-theoretic result that there are infinitely many prime numbers. For this, consider an arbitrary integer n, and denote the number of primes no greater than n by $\pi(n)$. Take a random variable N uniformly distributed over $\{1,2,\ldots,n\}$, and write it in its unique prime factorization, $N = p_1^{X_1}p_2^{X_2}\ldots p_{\pi(n)}^{X_{\pi(n)}}$, where $\{p_1,p_2,\cdots,p_{\pi(n)}\}$ are primes no greater than n, and each X_i is the largest power $k \geq 0$ such that p_i^k divides N. By inspecting $H(N)$, prove that $\pi(n) \to \infty$ as $n \to \infty$. For further reading, refer to [14].

证明: 由素数分解的唯一性得知，一个 N 的抽样结果可以和一组 $x_1,x_2,...,x_{\pi(n)}$ 形成一一对应，从而依次根据均匀分布熵的表达、链式法则和条件减少熵的性质有

$$\log_2(n) = H(N) = H(X_1,X_2,\cdots,X_{\pi(n)}) \leq \sum_{i=1}^{\pi(n)} H(X_i).$$

对每个正整数 $i \leq \pi(n)$, 均有 $2^{X_i} \leq p_i^{X_i} \leq N \leq n$, 从而 $0 \leq X_i \leq \lfloor \log_2(n) \rfloor$. 这样对每个正整数 $i \leq \pi(n)$, X_i 字母表大小为 $\lfloor \log_2(n) \rfloor + 1$, 所以有 $H(X_i) \leq \log_2(\lfloor \log_2(n) \rfloor + 1) \leq \log_2(\log_2(n)+1)$, 即有

$$\log_2(n) \leq \pi(n)\log_2(\log_2(n)+1).$$

$n \to \infty$ 时, 因为 $\log_2(n)/\log_2(\log_2(n)+1) \to \infty$, 所以 $\pi(n) \to \infty$. □

第 12 题

For integer set $[n] := \{1, 2, \ldots, n\}$*, drawing each of its elements independently with probability* p *leads to a random subset of* $[n]$*. For two such subsets,* A *and* B*, generated independently, calculate* $H(A)$ *and* $H(A \cup B)$*, and show that* $H(A \cup B) > H(A)$ *when* $p \le \frac{3-\sqrt{5}}{2}$*.*
This is related to the so-called union-closed sets conjecture, for which the first constant lower bound was established using an information-theoretic argument; for further reading, refer to [15].

这道题应该假设了 $0 < p < (3-\sqrt{5})/2$.

证明: 对每个 $i \in [n]$ 定义随机变量

$$X_i = \begin{cases} 0, & i \in A \\ 1, & i \notin A \end{cases}, \quad Y_i = \begin{cases} 0, & i \in B \\ 1, & i \notin B \end{cases}.$$

这样以概率 1 有 $A = \{i \in [n] | X_i = 0\}$ 和 $A \cup B = \{i \in [n] | X_i Y_i = 0\}$, X_1, X_2, $\cdots$, X_n, Y_1, Y_2, $\cdots$, Y_n 独立,

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

$$H(A \cup B) = H(X_1 Y_1, X_2 Y_2, \cdots, X_n Y_n) = \sum_{i=1}^{n} H(X_i Y_i).$$

对每个 $i \in [n]$, 因为 $P_{X_i}(1) = 1 - p$, $P[X_i Y_i = 1] = P_{X_i}(1) P_{Y_i}(1) = (1-p)^2$, 所以 $H(X_i) = h_2(1-p)$, $H(X_i Y_i) = h_2((1-p)^2)$. 可以验证 $(2-\sqrt{2})/2 < (3-\sqrt{5})/2$.

如果 $0 < p \le (2-\sqrt{2})/2$, 则 $\sqrt{2}/2 \le 1 - p < 1$, $1/2 \le (1-p)^2 < 1 - p$,

$$H(A) = nh_2(1-p) < nh_2((1-p)^2) = H(A \cup B).$$

如果 $(2-\sqrt{2})/2 < p < (3-\sqrt{5})/2$, 则 $p < (1-p)^2 < 1/2$,

$$H(A) = nh_2(p) < nh_2((1-p)^2) = H(A \cup B). \qquad \square$$

另外如果 $p = 0$ 或$(3-\sqrt{5})/2$, 可以验证 $h_2(1-p) = h_2((1-p)^2)$, 所以有 $H(A) = H(A \cup B)$.

第 13 题

Consider a random variable X *generated as follows: conditioned upon a random variable* Z *taking values in* $\{1, 2, \ldots\}$*, let* X *be a geometric random variable (see Example 2.3) with parameter* 2^{-Z}*.*

a) Show that if $\mathbf{E}[Z]=\infty$ then $H(X)=\infty$.

b) Define a random variable Y as follows: $Y=0$ with probability $1-\epsilon$ and $Y=X$ with probability ϵ. Let $\hat{Y}=0$ with probability one. Show that if $H(X)=\infty$ then $H(Y|\hat{Y})$ does not tend to zero, no matter how small the decision error probability $P_e=P(Y\neq\hat{Y})=\epsilon>0$ is. This example illustrates the delicacy when applying Fano's inequality when the alphabet is infinite ([8, Example 2.49]).

a) 证明:

$$\begin{aligned}
H(X) &\geq H(X|Z) \\
&= \sum_{z,P_Z(z)>0} P_Z(z)H(X|Z=z) \\
&= \sum_{z,P_Z(z)>0} P_Z(z)\frac{h_2(2^{-z})}{2^{-z}} \\
&\geq \sum_{z,P_Z(z)>0} P_Z(z)\frac{-2^{-z}\log_2(2^{-z})}{2^{-z}} \\
&= \mathbf{E}[Z].
\end{aligned}$$

□

b) 证明: 可以认为随机变量 W 以概率 $1-\epsilon$ 取 0, 以概率 ϵ 取 1, 独立于 X, 且 Y 由

$$Y=\begin{cases}0, & W=0\\ X, & W=1\end{cases}$$

定义.

由于 $Y\neq\hat{Y}$ 当且仅当 $W=1$, 所以 $P_e=\epsilon=P_W(1)$.

$$\begin{aligned}
H(Y|\hat{Y}) &= H(Y) \\
&\geq H(Y|W) \\
&= P_W(0)H(Y|W=0)+P_W(1)H(Y|W=1) \\
&= (1-P_e)H(0|W=0)+P_eH(X|W=1) \\
&= (1-P_e)\cdot 0+P_e\cdot\infty,
\end{aligned}$$

所以 $P_e\to 0$ 时 $H(Y|\hat{Y})$ 不趋于 0. □

第 14 题

Prove the submodularity property of entropy: for any two sets of random variables $\mathbf{S}_1$ and $\mathbf{S}_2$, $H(\mathbf{S}_1\cup\mathbf{S}_2)+H(\mathbf{S}_1\cap\mathbf{S}_2)\leq H(\mathbf{S}_1)+H(\mathbf{S}_2)$.

证明: 记 $X = \mathbf{S}_1 \setminus \mathbf{S}_2$, $Y = \mathbf{S}_1 \cap \mathbf{S}_2$, $Z = \mathbf{S}_2 \setminus \mathbf{S}_1$. 这样 $\mathbf{S}_1 = (X,Y)$, $\mathbf{S}_2 = (Y,Z)$, $\mathbf{S}_1 \cup \mathbf{S}_2 = (X,Y,Z)$. 因为

$$\begin{aligned} H(X,Y,Z) + H(Y) &= H(X,Y) + H(Z|X,Y) + H(Y) \\ &\leq H(X,Y) + H(Z|Y) + H(Y) \\ &= H(X,Y) + H(Y,Z), \end{aligned}$$

所以 $H(\mathbf{S}_1 \cup \mathbf{S}_2) + H(\mathbf{S}_1 \cap \mathbf{S}_2) \leq H(\mathbf{S}_1) + H(\mathbf{S}_2)$. □

第 15 题

For random variables X and Y and a mapping f, under what condition does $H(X|f(Y)) = H(X|Y)$ hold?

解: 因为 $I(X;f(Y)) = H(X) - H(X|f(Y))$, $I(X;Y) = H(X) - H(X|Y)$, 所以

$$H(X|f(Y)) = H(X|Y) \tag{10}$$

当且仅当

$$I(X;f(Y)) = I(X;Y). \tag{11}$$

因为 $X \leftrightarrow Y \leftrightarrow f(Y)$, 根据讲义中 Theorem 3.5, 11 式成立当且仅当 $X \leftrightarrow f(Y) \leftrightarrow Y$. 因此 10 式成立当且仅当 $X \leftrightarrow f(Y) \leftrightarrow Y$. □

第 16 题

Suppose that $\Theta \in (0,1)$ is a random variable over the unit interval, and conditioned upon Θ, $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ consists of n i.i.d. random variables $X_i \sim$ Bernoulli(Θ). Define $T = \sum_{i=1}^n X_i$. Is T a sufficient statistic for Θ?

解: 根据题目描述，已存在 Markov chain: $\Theta \leftrightarrow \mathbf{X} \leftrightarrow T$, 根据充分统计量的定义，我们需要证明的是 Markov chain: $\Theta \leftrightarrow T \leftrightarrow \mathbf{X}$ 是否成立。

当 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ 时, 且 $\sum_{i=1}^n x_i = k$ 时，有

$$\begin{aligned} P_{\mathbf{X}|T,\Theta}(\mathbf{X} = \mathbf{x}|T = k, \Theta = \theta) &= \frac{P_{\mathbf{X},T|\Theta}(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n, T = k|\Theta = \theta)}{P_{T|\Theta}(T = k|\Theta = \theta)} \\ &= \frac{P_{\mathbf{X}|\Theta}(X_1 = x_1, X_2 = x_2, \cdots, X_n = k - \sum_{i=1}^{n-1} x_i|\Theta = \theta)}{P_{T|\Theta}(T = k|\Theta = \theta)} \\ &= \frac{\theta^k (1-\theta)^{n-k}}{\binom{n}{k}\theta^k(1-\theta)^{n-k}} = \frac{1}{\binom{n}{k}} \end{aligned} \tag{12}$$

因此有

$$P_{\mathbf{X}|T,\Theta}(\mathbf{X}=\mathbf{x}|T=k,\Theta=\theta)=\begin{cases}\frac{1}{\binom{n}{k}}, & \text{if } \sum_{i=1}^{n}x_i=k,\\ 0, & \text{otherwise.}\end{cases} \tag{13}$$

由 13 式可以看出上述条件分布与 Θ 无关，与 T 的取值有关，所以 Markov chain: $\Theta \leftrightarrow T \leftrightarrow \mathbf{X}$ 成立，即 T 是 Θ 的充分统计量。

□

第 17 题

For the two-state Markov chain in Example 3.5, if we undersample it to obtain a new stochastic process X_1, X_3, X_5, ..., is it still a Markov chain? Under stationarity, evaluate its entropy rate and compare with that of the original Markov chain X_1, X_2, X_3, ...

解: 设 n 是正整数. 定义随机变量 $Y=(X_1,X_3,\cdots,X_{2n-1})$. 如果 $P_{X_{2n+2},X_{2n+1},Y}(x_2,x_1,y)>0$ 且 $x_3\in\{0,1\}$ 则

$$\begin{aligned}
&P_{X_{2n+3},X_{2n+2}|X_{2n+1},Y}(x_3,x_2|x_1,y)\\
=&P_{X_{2n+2}|X_{2n+1},Y}(x_2|x_1,y)P_{X_{2n+3}|X_{2n+2},X_{2n+1},Y}(x_3|x_2,x_1,y)\\
=&P_{X_{2n+2}|X_{2n+1}}(x_2|x_1)P_{X_{2n+3}|X_{2n+2},X_{2n+1}}(x_3|x_2,x_1)\\
=&P_{X_{2n+3},X_{2n+2}|X_{2n+1}}(x_3,x_2|x_1).
\end{aligned}$$

等式两边对 x_2 求和得 $P_{X_{2n+3}|X_{2n+1},Y}(x_3|x_1,y)=P_{X_{2n+3}|X_{2n+1}}(x_3|x_1)$. 因此 X_1, X_3, X_5, $\cdots$ 是 Markov 链.

根据平稳 Markov 链的熵率的定义, Markov 链 X_1, X_2, $X_3,\ldots$ 和 X_1, X_3, X_5, $\cdots$ 的熵率分别为 $H(X_3|X_2)$ 和 $H(X_3|X_1)$. 依据数据处理不等式，我们有

$$\begin{aligned}
I(X_2;X_3)&\geq I(X_1;X_3)\\
H(X_3)-H(X_3|X_2)&\geq H(X_3)-H(X_3|X_1)
\end{aligned}$$

即 $H(X_3|X_2)\leq H(X_3|X_1)$，说明 Markov 链 X_1, X_2, X_3, $\cdots$ 的熵率小于等于 Markov 链 X_1, X_3, X_5, $\cdots$ 的熵率.

我们可以通过以下方法进一步计算 Markov 链 X_1, X_3, X_5, $\cdots$ 的熵率. 用 Q 表示 Markov 链 X_1, X_2, X_3, $\cdots$ 的一步转移概率矩阵

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

用 π 表示它的平稳分布. X_1, X_3, X_5, $\cdots$ 的一步转移概率矩阵等于

$$Q^2 = \begin{bmatrix} 1-2\alpha+\alpha^2+\alpha\beta & 2\alpha-\alpha^2-\alpha\beta \\ 2\beta-\alpha\beta-\beta^2 & 1-2\beta+\alpha\beta+\beta^2 \end{bmatrix}.$$

因为 $[\pi(0),\pi(1)]Q = [\pi(0),\pi(1)]$, 所以 $[\pi(0),\pi(1)]Q^2 = [\pi(0),\pi(1)]$, π 也是 X_1, X_3, X_5, $\cdots$ 的平稳分布. 由于我们假设了 X_1, X_3, X_5, $\cdots$ 是平稳的, X_1 服从 π. 这样 X_1, X_3, X_5, $\cdots$ 的熵率等于

$$\begin{aligned} H(X_3|X_1) &= \pi(0)H(X_3|X_1=0) + \pi(1)H(X_3|X_1=1) \\ &= \frac{\beta}{\alpha+\beta}h_2(2\alpha-\alpha^2-\alpha\beta) + \frac{\alpha}{\alpha+\beta}h_2(2\beta-\alpha\beta-\beta^2). \end{aligned}$$

□

我们也可以对每个正整数 n 证明 $I(X_1, X_3, \cdots, X_{2n-1}; X_{2n+3}|X_{2n+1}) = 0$, 从而证明 X_1, X_3, X_5, $\cdots$ 是一条 Markov 链.

$$\begin{aligned} I(Y; X_{2n+2}, X_{2n+3}|X_{2n+1}) &= I(Y; X_{2n+2}|X_{2n+1}) + I(Y; X_{2n+3}|X_{2n+1}, X_{2n+2}) \\ &= I(Y; X_{2n+3}|X_{2n+1}) + I(Y; X_{2n+2}|X_{2n+1}, X_{2n+3}) \end{aligned}$$

又因为 $Y \leftrightarrow X_{2n+1} \leftrightarrow X_{2n+2}$ 和 $Y \leftrightarrow X_{2n+2} \leftrightarrow X_{2n+3}$, 所以我们有 $I(Y; X_{2n+2}|X_{2n+1}) = 0$, $I(Y; X_{2n+3}|X_{2n+1}, X_{2n+2}) = 0$ 和 $I(Y; X_{2n+2}|X_{2n+1}, X_{2n+3}) = 0$, 从而可得 $I(Y; X_{2n+3}|X_{2n+1}) = 0$, 即 $X_1, \cdots, X_{2n-1} \leftrightarrow X_{2n+1} \leftrightarrow X_{2n+2}$ 成立.

用类似的方法可以证明如果正整数 $k_1 \le n_1 < k_2 \le n_2 < \cdots$ 则

$$\{(X_{k_j}, X_{k_j+1}, \cdots, X_{n_j})\}_{j=1}^{\infty}$$

是一条 Markov 链. 见 [1] 推论 3.10.

第 18 题

Define an "almost Markov" relationship for three random variables (X, Y, Z) if they satisfy

$$p(z|x,y) = p(z|y)(1+\epsilon(x,y,z)),$$

where $|\epsilon(x,y,z)| \le \delta$ for any (x,y,z) tuple. Prove that for such an "almost Markov" relationship, we have the following "δ-approximate DPI" hold:

$$I(X;Z) \le I(X;Y) + \delta^2.$$

这道题中互信息的底应该是 e.

证明: 类似于数据处理不等式的推导,

$$I(X;Z) \leq I(X;Z) + I(X;Y|Z) = I(X;Y,Z) = I(X;Y) + I(X;Z|Y). \tag{14}$$

根据条件互信息的定义，我们有：

$$\begin{aligned} I(X;Z|Y) &= \sum_{x,y,z} P_{X,Y,Z}(x,y,z) \ln \frac{P_{X,Z|Y}(x,z|y)}{P_{X|Y}(x|y)P_{Z|Y}(z|y)} \\ &= \sum_{x,y,z} P_{X,Y,Z}(x,y,z) \ln \frac{P_{Z|X,Y}(z|x,y)}{P_{Z|Y}(z|y)} \\ &= \sum_{x,y,z} P_{X,Y,Z}(x,y,z) \ln(1+\epsilon(x,y,z)) \end{aligned} \tag{15}$$

在开始后续分析之前，我们可以得到以下事实：

$$\begin{aligned} \sum_{x,y,z} P_{X,Y,Z}(x,y,z) &= 1 \\ \sum_{x,y,z} P_{X,Y}(x,y)P_{Z|X,Y}(z|x,y) &= 1 \\ \sum_{x,y,z} P_{X,Y}(x,y)P_{Z|Y}(z|y)(1+\epsilon(x,y,z)) &= 1 \\ \sum_{x,y,z} P_{X,Y}(x,y)P_{Z|Y}(z|y) + \sum_{x,y,z} P_{X,Y}(x,y)P_{Z|Y}(z|y)\epsilon(x,y,z) &= 1 \end{aligned}$$

又 $\sum_{x,y,z} P_{X,Y}(x,y)P_{Z|Y}(z|y) = \sum_{x,y} P_{X,Y}(x,y) \sum_z P_{Z|Y}(z|y) = 1$，所以有

$$\sum_{x,y,z} P_{X,Y}(x,y)P_{Z|Y}(z|y)\epsilon(x,y,z) = 0 \tag{16}$$

接着从 15 式出发，我们有

$$\begin{aligned} I(X;Z|Y) &= \sum_{x,y,z} P_{X,Y,Z}(x,y,z) \ln(1+\epsilon(x,y,z)) \\ &\leq \sum_{x,y,z} P_{X,Y,Z}(x,y,z)\epsilon(x,y,z) \\ &= \sum_{x,y,z} P_{X,Y}(x,y)P_{Z|Y}(z|y)(1+\epsilon(x,y,z))\epsilon(x,y,z) \\ &= \sum_{x,y,z} P_{X,Y}(x,y)P_{Z|Y}(z|y)\epsilon(x,y,z) + \sum_{x,y,z} P_{X,Y}(x,y)P_{Z|Y}(z|y)\epsilon^2(x,y,z) \\ &\leq \sum_{x,y,z} P_{X,Y}(x,y)P_{Z|Y}(z|y)\epsilon(x,y,z) + \delta^2 \sum_{x,y,z} P_{X,Y}(x,y)P_{Z|Y}(z|y) \\ &= \delta^2 \end{aligned} \tag{17}$$

其中第一个不等式是因为 $\ln(1+x) \leq x$，最后一个等号基于 16 式的结果。综合 14 式和 17 式，最终证得 $I(X;Z) \leq I(X;Y) + \delta^2$. □

第 19 题

For random variables $V, W_1, W_2, \ldots, W_n$, prove that

$$H(V) \geq \sum_{i=1}^{n} I(V;W_i), \tag{18}$$

when $W_1, W_2, \ldots, W_n$ are mutually independent.

解: 方法一：

$$\begin{aligned} H(V) &= I(V;W_1) + H(V|W_1) \\ &= I(V;W_1) + I(V;W_2|W_1) + H(V|W_1,W_2) \\ &= \sum_{i=1}^{n} I(V;W_i|W_1,\ldots,W_{i-1}) + H(V|W_1,\ldots,W_n) \end{aligned} \tag{19}$$

进一步有

$$\begin{aligned} I(V;W_i|W_1,\ldots,W_{i-1}) &= H(W_i|W_1,\ldots,W_{i-1}) - H(W_i|V,W_1,\ldots,W_{i-1}) \\ &\overset{(a)}{\geq} H(W_i) - H(W_i|V) \\ &= I(V;W_i) \end{aligned} \tag{20}$$

其中 W_0 被视为常数，即有 $H(W_1|W_0) = H(W_1)$; (a) 是由条件减小熵定理 (Theorem 3.6) 得到的, 即

$$H(W_i|W_1,\ldots,W_{i-1}) \overset{(b)}{=} H(W_i)$$
$$H(W_i|V,W_1,\ldots,W_{i-1}) \overset{(c)}{\leq} H(W_i|V)$$

其中 (b) 等式恒成立的原因是 $W_1, W_2, \ldots, W_n$ 相互独立, (c) 取等的条件为当且仅当 Markov 链 $W_1, W_2, \ldots, W_{i-1} \leftrightarrow V \leftrightarrow W_i$ 存在.

结合 19 式和 20 式得

$$\begin{aligned} H(V) &= \sum_{i=1}^{n} I(V;W_i|W_1,\ldots,W_{i-1}) + H(V|W_1,\ldots,W_n) \\ &= \sum_{i=1}^{n} I(V;W_i) + H(V|W_1,\ldots,W_n) \\ &\geq \sum_{i=1}^{n} I(V;W_i) \end{aligned}$$

方法二：

$$\begin{aligned}H(V) &= H(V|W_1, W_2, \ldots, W_n) + I(V; W_1, W_2, \ldots, W_n)\\ &\geq I(V; W_1, W_2, \ldots, W_n)\\ &= \sum_{i=1}^{n} I(V; W_i|W_1, W_2, \ldots, W_{i-1})\\ &\geq \sum_{i=1}^{n} I(V; W_i)\end{aligned}$$

其中最后一个不等号基于 (20) 式。 □

参考文献

[1] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 2nd ed. Cambridge University Press, 2011.