

第 5 讲 习题课 1

例 0.1

不断抛掷一枚均匀硬币，直到第一次出现正面为止。令 X 表示所需的抛掷次数。

(a) 求熵 $H(X)$ (以比特为单位)。以下表达式可能会有用：

$$\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}, \quad \sum_{n=0}^{\infty} nr^n = \frac{r}{(1-r)^2}.$$

(b) 根据该分布抽取了一个随机变量 X 。构造一个形式为 “ X 是否包含在集合 S 中?” 的 “有效” 的 “是-否” 问题序列，以确定 X 的值。比较 $H(X)$ 与确定 X 所需的期望问题数。

证明.

- (a) 抛掷到首次出现正面时的次数 X 服从参数 $p = 1/2$ 的几何分布, 其中 $P(X = n) = pq^{n-1}$, $n \in \{1, 2, \dots\}$. 因此 X 的熵为:

$$\begin{aligned}
 H(X) &= - \sum_{n=1}^{\infty} pq^{n-1} \log(pq^{n-1}) \\
 &= - \left[\sum_{n=0}^{\infty} pq^n \log p + \sum_{n=0}^{\infty} npq^n \log q \right] \\
 &= \frac{-p \log p}{1-q} - \frac{pq \log q}{p^2} \\
 &= \frac{-p \log p - q \log q}{p} \\
 &= H(p)/p \text{ 比特 (bits).}
 \end{aligned}$$

如果 $p = 1/2$, 则 $H(X) = 2$ 比特.

证明.

- (b) 直观来看, 最好的问题显然是那些获得“是”或“否”的几率各占一半的问题。因此, 我们可以猜测, 最“有效”的一系列问题为: $X = 1$ 吗? 如果不是, $X = 2$ 吗? 如果不是, $X = 3$ 吗? 得到的期望问题数等于 $\sum_{n=1}^{\infty} n(1/2^n) = 2$ 。这应该增强了这样一种直觉: $H(X)$ 是衡量 X 的不确定性的一种度量。事实上, 在此情况下, 熵恰好等于确定 X 所需的平均问题数, 且在一般情况下 $E(\text{问题数}) \geq H(X)$ 。这个问题可以理解为信源编码问题。令 0 = 否, 1 = 是, X = 信源, 而 Y = 编码后的信源。那么上述过程中的问题集可以写成 (X, Y) 对的集合: $(1, 1), (2, 01), (3, 001)$ 等。事实上, 这个通过直觉得出的代码正是最小化预期问题数的最佳 (哈夫曼 Huffman) 码。 □

例 0.2: 课本习题 2.10

设离散随机变量 X_1 和 X_2 的概率质量函数分别为 $p_1(\cdot)$ 和 $p_2(\cdot)$, 字母表分别为 $\mathcal{X}_1 = \{1, 2, \dots, m\}$, $\mathcal{X}_2 = \{m+1, \dots, n\}$. 设

$$X = \begin{cases} X_1 & \text{概率为 } \alpha \\ X_2 & \text{概率为 } 1 - \alpha. \end{cases}$$

- (1) 试求 $H(X)$ 关于 $H(X_1)$ 、 $H(X_2)$ 和 α 的表达式.
- (2) 试关于 α 最大化 $H(X)$, 证明 $2^{H(X)} \leq 2^{H_1(X)} + 2^{H(X_2)}$.

证明.

注意到

$$X = \begin{cases} X_1 & \text{概率为 } \alpha \\ X_2 & \text{概率为 } 1 - \alpha, \end{cases}$$

我们定义

$$\theta = f(X) = \begin{cases} 1 & \text{若 } X = X_1, \\ 2 & \text{若 } X = X_2, \end{cases}$$

那么我们有

$$\begin{aligned} H(X) &= H(X, f(X)) = H(\theta) + H(X|\theta) \\ &= H(\theta) + p(\theta = 1)H(X|\theta = 1) + p(\theta = 2)H(X|\theta = 2) \\ &= H(\alpha) + \alpha H(X_1) + (1 - \alpha)H(X_2) \end{aligned}$$

证明.

其中 $H(\alpha) = -\alpha \log \alpha = (1 - \alpha) \log(1 - \alpha)$. 记
 $f(\alpha) = H(\alpha) + \alpha H(X_1) + (1 - \alpha) H(X_2)$, 则

$$f'(\alpha) = \log\left(\frac{1 - \alpha}{\alpha}\right) + H(X_1) - H(X_2).$$

解方程 $f'(\alpha_0) = 0$ 可得 $\alpha_0 = \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}}$. 不难看出 f 在 α_0 处取到最大值, 于是

$$\begin{aligned} H(X) &\leq -\frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}} \log \frac{2^{H(X_2)}}{2^{H(X_1)} + 2^{H(X_2)}} - \frac{2^{H(X_2)}}{2^{H(X_1)} + 2^{H(X_2)}} \log \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}} \\ &\quad + \frac{2^{H(X_2)}}{2^{H(X_1)} + 2^{H(X_2)}} H(X_1) + \frac{2^{H(X_1)}}{2^{H(X_1)} + 2^{H(X_2)}} H(X_2) \\ &= \log(2^{H(X_1)} + 2^{H(X_2)}). \end{aligned}$$

从而 (2) 得证.



例 0.3: (课本习题 2.28)

设 $p_i > 0$, $i = 1, 2, \dots, m$ 。证明服从分布 $(p_1, \dots, p_i, \dots, p_j, \dots, p_m)$ 的随机变量的熵, 小于或等于服从分布 $(p_1, \dots, \frac{p_i + p_j}{2}, \dots, \frac{p_i + p_j}{2}, \dots, p_m)$ 的随机变量的熵。

解.

这个问题取决于对数函数的凸性。设

$$P_1 = (p_1, \dots, p_i, \dots, p_j, \dots, p_m)$$

$$P_2 = (p_1, \dots, \frac{p_i + p_j}{2}, \dots, \frac{p_i + p_j}{2}, \dots, p_m)$$

那么,

$$\begin{aligned} H(P_2) - H(P_1) &= -2 \left(\frac{p_i + p_j}{2} \right) \log \left(\frac{p_i + p_j}{2} \right) + p_i \log p_i + p_j \log p_j \\ &= -(p_i + p_j) \log \left(\frac{p_i + p_j}{2} \right) + p_i \log p_i + p_j \log p_j. \end{aligned}$$

由于 $f(x) = x \log x$ 在 $x > 0$ 时是严格凸函数, 以下不等式成立:

$$\frac{f(p_i) + f(p_j)}{2} = \frac{1}{2}(p_i \log p_i + p_j \log p_j) \geq f\left(\frac{p_i + p_j}{2}\right) = \frac{p_i + p_j}{2} \log \frac{p_i + p_j}{2}.$$

因此, $H(P_2) \geq H(P_1)$.



例 0.4: (课本习题 2.39)

设 X, Y 和 Z 为三个服从 $\text{Bernoulli}(\frac{1}{2})$ 的二元随机变量, 且两两独立. 在上述约束条件下, $H(X, Y, Z)$ 的最小值是多少?

证明.
我们有

$$\begin{aligned} H(X, Y, Z) &= H(X, Y) + H(Z|X, Y) \\ &\geq H(X, Y) \\ &= 2 \text{ 比特}. \end{aligned}$$

下面我们验证这个下界可以达到, 从而 $H(X, Y, Z)$ 的最小值为 2 比特.
设 $X, Y \sim \text{Bernoulli}(1/2)$, $Z = X \oplus Y$, 其中 $\oplus$ 表示模 2 加法.

□

例 0.5

设 $Z_1, Z_2, \dots$ 为独立同分布的随机变量, 服从 $\{0, 1\}$ 上的均匀分布. 对于 $1 \leq i \leq n$, 设

$$X_i = \sum_{j=1}^i Z_j.$$

求 $I(X_1; X_2, X_3, \dots, X_n)$.

证明.

首先注意到 $X_1 \rightarrow X_2 \rightarrow \cdots \rightarrow X_n$ 构成一个马尔可夫链. (思考: 为什么?)

由互信息的链式法则, 我们有

$$\begin{aligned}
 I(X_1; X_2, X_3, \cdots, X_n) &= \sum_{i=2}^n I(X_1; X_i | X_2, \cdots, X_{i-1}) \\
 &= I(X_1; X_2). \\
 &= I(Z_1; Z_1 + Z_2) \\
 &= H(Z_1 + Z_2) - H(Z_1 + Z_2 | Z_1) \\
 &= \frac{3}{2} - 1 = 1/2 \text{ 比特}.
 \end{aligned}$$

□

例 0.6: (课本习题 4.33)

设 $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ 构成一个马尔科夫链. 证明:

$$I(X_1; X_3) + I(X_2; X_4) \leq I(X_1; X_4) + I(X_2; X_3).$$

证明.
注意到

$$\begin{aligned} I(X_1; X_3, X_4) &= I(X_1; X_3) + I(X_1; X_4|X_3) \\ &= I(X_1; X_4) + I(X_1; X_3|X_4) \end{aligned}$$

于是

$$I(X_1; X_4) - I(X_1; X_3) = I(X_1; X_4|X_3) - I(X_1; X_3|X_4).$$

类似地,

$$I(X_2; X_3) - I(X_2; X_4) = I(X_2; X_3|X_4) - I(X_2; X_4|X_3).$$

从而我们有 (由于 $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ 构成一个马尔可夫链,
 $I(X_1; X_4|X_3) = 0$, $I(X_2; X_4|X_3) = 0$.)

$$\begin{aligned} & I(X_1; X_4) + I(X_2; X_3) - I(X_1; X_3) + I(X_2; X_4) \\ &= I(X_1; X_4|X_3) - I(X_1; X_3|X_4) + I(X_2; X_3|X_4) - I(X_2; X_4|X_3) \\ &= I(X_2; X_3|X_4) - I(X_1; X_3|X_4) \end{aligned}$$

证明.

又由

$$\begin{aligned} I(X_1, X_2; X_3|X_4) &= I(X_1; X_3|X_4) + I(X_2; X_3|X_1, X_4) \\ &= I(X_2; X_3|X_4) + I(X_1; X_3|X_2, X_4), \end{aligned}$$

我们知

$$\begin{aligned} &I(X_1; X_4) + I(X_2; X_3) - I(X_1; X_3) + I(X_2; X_4) \\ &= I(X_2; X_3|X_1, X_4) - I(X_1; X_3|X_2, X_4). \end{aligned}$$

注意到 (这里我们再次用到 $X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$ 构成一个马尔可夫链这一事实) $I(X_1; X_3|X_2, X_4) = I(X_1; X_3, X_4|X_2) - I(X_1; X_4|X_2) = 0$, 于是

$$I(X_1; X_4) + I(X_2; X_3) - I(X_1; X_3) + I(X_2; X_4) = I(X_2; X_3|X_1, X_4) \geq 0.$$

从而得证.

例 0.7

记 $\pi(n)$ 为不超过 n 的素数个数. 注意到对于任一正整数 n , 其有唯一的形如下的素因子分解

$$n = \prod_{i=1}^{\pi(n)} p_i^{X_i}.$$

其中 $p_1, p_2, p_3, \dots$ 为素数. 即 $p_1 = 2, p_2 = 3, p_3 = 5, \dots$. $X_i = X_i(n)$ 为一个非负整数, 其表示 n 的素因子分解中 p_i 的重数. 设 N 为在 $\{1, 2, \dots, n\}$ 上均匀分布的一个随机变量.

(1) 证明 $X_i(N)$ 为一个整数值随机变量, 满足

$$0 \leq X_i(N) \leq \log n.$$

(2) 证明 $\log n = H(N) \leq \pi(n) \log(\log n + 1)$. 从而 $\pi(n) \geq \frac{\log n}{\log(\log n + 1)}$.
特别地当 $n \rightarrow \infty$ 时有 $\pi(n) \rightarrow \infty$.

证明.

(1) 由于 $p_i^{X_i} | N$, 所以 $2^{X_i(N)} \leq p_i^{X_i(N)} \leq N \leq n$. 从而 $X_i(N) \leq \log n$.

(2) 注意到对于每个 N , 其对应的 $(X_1(N), \dots, X_{\pi(n)}(N))$ 是唯一的, 所以

$$\begin{aligned} \log n &= H(N) \\ &= H(X_1(N), X_2(N), \dots, X_{\pi(n)}(N)) \\ &\leq H(X_1(N)) + \dots + H(X_{\pi(n)}(N)) \\ &\leq \pi(n) \log(\log(n+1)), \end{aligned}$$

从而 $\pi(n) \geq \frac{\log n}{\log(\log(n+1))}$. 特别地当 $n \rightarrow \infty$ 时有 $\pi(n) \rightarrow \infty$. □

注记 0.8

上例的结论可以进一步加强. 设 $N \sim \text{Uniform}\{1, 2, \dots, n\}$, 其可以唯一表示为

$$N = M^2 \cdot p_1^{Y_1} \cdot p_2^{Y_2} \cdot \dots \cdot p_{\pi(n)}^{Y_{\pi(n)}},$$

其中 M 是使得 $M^2 | N$ 的最大整数. 这样随机变量 Y_i 取值范围是 $\{0, 1\}$. 由于 $M^2 | N$, $1 \leq M \leq \sqrt{n}$. 从而我们有

$$\begin{aligned} \log n &= H(N) \\ &= H(M, Y_1, \dots, Y_{\pi(n)}) \\ &\leq H(M) + H(Y_1) + H(Y_2) + \dots + H(Y_{\pi(n)}) \\ &\leq \frac{1}{2} \log n + \pi(n). \end{aligned}$$

从而 $\pi(n) \geq \frac{1}{2} \log n$.

例 0.9

假设 X 是一个取值于自然数集 $\mathbb{N} = \{1, 2, \dots\}$ 上的随机变量, 其概率质量函数 (PMF) 为 P . 我们将尝试寻找使该随机变量的熵有限的条件.

- (1) 假设 $\mathbb{E}[X] < \infty$. 设 Q 为成功概率为 r 的几何随机变量的 PMF, 即对于所有 $x \in \mathbb{N}$, 有 $Q(x) = (1-r)^{x-1}r$, 其中 $r \in (0, 1)$. 请用 $H(X)$ 、 $\mathbb{E}[X]$ 和 r 来表示相对熵 $D(P||Q)$.
- (2) 利用上一问的结论证明:

$$H(X) \leq \mathbb{E}[X] \cdot h_2\left(\frac{1}{\mathbb{E}[X]}\right).$$

当 X 服从几何分布时等号成立. 由此我们知如果 $\mathbb{E}[X] < \infty$ 则 X 的熵是有限的. 其中, $h_2(\cdot)$ 表示二元熵函数, 定义为 $h_2(x) = -x \log_2 x - (1-x) \log_2 (1-x)$.

- (3) 现在假设我们不知道是否有 $\mathbb{E}[X] < \infty$ 但是我们知道 $\mathbb{E}[\log X] < \infty$. 证明这也能推出 $H(X) < \infty$.

证明.

(1) 展开相对熵, 我们有:

$$\begin{aligned}
 D(P||Q) &= \sum_{x \in \mathbb{N}} P(x) \log \frac{P(x)}{Q(x)} \\
 &= \sum_{x \in \mathbb{N}} P(x) \log P(x) + \sum_{x \in \mathbb{N}} P(x) \log \frac{1}{Q(x)} \\
 &= -H(X) + \sum_{x \in \mathbb{N}} P(x) \log \frac{1}{(1-r)^{x-1}r} \\
 &= -H(X) + (\mathbb{E}[X] - 1) \log \frac{1}{1-r} + \log \frac{1}{r}.
 \end{aligned}$$

证明.

(2) 设 Q 与上一问的分布相同。利用 KL 散度的非负性 $D(P||Q) \geq 0$, 我们有:

$$H(X) \leq (\mathbb{E}[X] - 1) \log \left(\frac{1}{1-r} \right) + \log \left(\frac{1}{r} \right).$$

该上界对任意 $r \in (0, 1)$ 均成立。因此, 我们选择使右式极小化的 r 值。对 r 求导并令导数为 0: $\frac{\mathbb{E}[X]-1}{1-r} - \frac{1}{r} = 0 \implies r = \frac{1}{\mathbb{E}[X]}$. 将此 r 代入上界表达式中, 可得:

$$\begin{aligned} H(X) &\leq (\mathbb{E}[X] - 1) \log \left(\frac{1}{1 - \frac{1}{\mathbb{E}[X]}} \right) + \log \left(\frac{1}{\frac{1}{\mathbb{E}[X]}} \right) \\ &= \mathbb{E}[X] \cdot \left[- \left(1 - \frac{1}{\mathbb{E}[X]} \right) \log \left(1 - \frac{1}{\mathbb{E}[X]} \right) - \frac{1}{\mathbb{E}[X]} \log \left(\frac{1}{\mathbb{E}[X]} \right) \right] \\ &= \mathbb{E}[X] \cdot h_2 \left(\frac{1}{\mathbb{E}[X]} \right). \end{aligned}$$

由于期望 $\mathbb{E}[X]$ 有限, 且二元熵函数受限于 1, 因此熵被一个有限值所界定, 故而是有限的。当 $D(P||Q) = 0$ 时取得上界, 这意味着 $P = Q$, 即 X 服从参数为 $r = \frac{1}{\mathbb{E}[X]}$ 的几何分布。

证明.

(3) 对于任意 $\alpha > 1$, 令:

$$\sum_{x=1}^{\infty} \frac{1}{x^{\alpha}} = c_{\alpha}$$

由于当 $\alpha > 1$ 时级数收敛, 我们知道存在这样一个有限的 c_{α} 。对于 $x \in \mathbb{N}$, 选取分布 Q 为 $Q(x) = \frac{1}{c_{\alpha} x^{\alpha}}$ 。我们可以将 P 与 Q 之间的 KL 散度展开为:

$$D(P||Q) = -H(X) + \mathbb{E} \left[\log \frac{1}{Q(X)} \right]$$

利用 KL 散度的非负性, 可得:

$$H(X) \leq \mathbb{E} [\log(c_{\alpha} X^{\alpha})] = \log c_{\alpha} + \alpha \mathbb{E} [\log X]$$

由于 c_{α} 、 α 和 $\mathbb{E} [\log X]$ 均为有限值, 因此熵被一个有限值所界定, 故也是有限的。 □

例 0.10: (课本习题 3.5)

设 $X_1, X_2, \dots$ 为 i.i.d. 离散随机变量序列, 熵为 $H(X)$. 设

$$C_n(t) = \{x^n \in \mathcal{X}^n : p(x^n) \geq 2^{-nt}\}$$

表示概率 $\geq 2^{-nt}$ 所有 n 长序列构成的子集.

- (a) 证明 $|C_n(t)| \leq 2^{nt}$.
- (b) 当 t 为何值时, 有 $P(\{X^n \in C_n(t)\}) \rightarrow 1$?

解.

(a) $1 = \sum_{x^n \in C_n(t)} p(x^n) \geq |C_n(t)| \min_{x^n \in C_n(t)} p(x^n)$, 从而 $|C_n(t)| \leq 2^{nt}$.

(b) 我们记 $H(X) = h$. 由 AEP, 对于任意 $\epsilon > 0$, $P(A_\epsilon^{(n)}) \rightarrow 1$.

如果 $t > H$, 我们有当 ϵ 充分小时, $C_n(t) \supseteq A_\epsilon^{(n)}$, 从而 $Pr\{x^n \in C_n(t)\} \rightarrow 1$.

如果 $t < H$, 我们有当 ϵ 充分小时, $C_n(t) \cap A_\epsilon^{(n)} = \emptyset$, 从而 $Pr\{x^n \in C_n(t)\} \rightarrow 0$. □

思考: 如果 $t = H$, 结论会是怎样?

例 0.11

我们将证明：如果 (Y, Y') 是独立同分布的离散随机变量，则：

$$H(Y - Y') - H(Y) \leq 2(H(Y' + Y) - H(Y))$$

我们将按照以下步骤证明该不等式。

- (a) **互信息的预处理不等式** 设 X_1, X_2 为离散随机变量。对于某些函数 $F(\cdot), G(\cdot)$ ，令 $Y_1 = F(X_1)$ 且 $Y_2 = G(X_2)$ 。证明：

$$I(X_1; X_2) \geq I(Y_1; Y_2).$$

- (b) **次模性** 假设存在函数 F, G, R 使得 $X_0 = F(X_1) = G(X_2)$ 且 $X_{12} = R(X_1, X_2)$ ，其中 X_1, X_2 为离散随机变量。利用前一部分证明：

$$H(X_{12}) + H(X_0) \leq H(X_1) + H(X_2).$$

- (c) **鲁扎三角形不等式** 设 X, Y, Z 为独立的离散随机变量。利用 (b) 部分证明：

$$1. H(X - Z) \leq H(X - Y) + H(Y - Z) - H(Y),$$

$$2. H(X - Z) \leq H(X + Y) + H(Y + Z) - H(Y).$$

[提示：在 (b) 部分中使用 $X_1 = (X - Y, Y - Z)$ ， $X_2 = (X, Z)$ ， $X_{12} = (X, Y, Z)$ 且 $X_0 = X - Z$ 。]

- (d) **熵的和与差** 利用前一部分得出结论：对于独立同分布的 (Y, Y') 随机变量，有：

$$H(Y - Y') - H(Y) \leq 2(H(Y' + Y) - H(Y)).$$

证明.

(a)

$$\begin{aligned}
 I(X_1; X_2) &= H(X_1) - H(X_1 | X_2) \\
 &\stackrel{(i)}{=} H(X_1) - H(X_1 | X_2, Y_2) \\
 &\stackrel{(ii)}{\geq} H(X_1) - H(X_1 | Y_2) \\
 &= I(Y_2; X_1) \\
 &= H(Y_2) - H(Y_2 | X_1) \\
 &\stackrel{(iii)}{=} H(Y_2) - H(Y_2 | X_1, Y_1) \\
 &\stackrel{(iv)}{\geq} H(Y_2) - H(Y_2 | Y_1) \\
 &= I(Y_1; Y_2)
 \end{aligned}$$

其中:

- (i) 来自 $Y_2 = G(X_2)$ 的事实。
- (ii) 来自 “条件化降低熵”。
- (iii) 来自 $Y_1 = F(X_1)$ 的事实。
- (iv) 来自 “条件化降低熵”。

证明.

(b) 显而易见 $H(X_{12}) \leq H(X_1, X_2)$, 因此

$$H(X_1) + H(X_2) - H(X_{12}) \geq H(X_1) + H(X_2) - H(X_1, X_2) = I(X_1; X_2).$$

结合上述 (a) 部分 $I(X_1; X_2) \geq I(Y_1; Y_2) = I(X_0; X_0) = H(X_0)$, 即完成了证明。

(c) 1. 令 $X_1 = (X - Y, Y - Z)$, $X_2 = (X, Z)$, $X_0 = X - Z$ 且 $X_{12} = (X, Y, Z)$ 。
因此对于函数 F, G, R , 我们有 $X_0 = F(X_1) = G(X_2)$ 且
 $X_{12} = R(X_1, X_2)$ 。利用上述 (b) 部分:

$$H(X, Y, Z) + H(X - Z) \leq H(X - Y, Y - Z) + H(X, Z)$$

根据独立性展开, 并利用条件化降低熵整理可得:

$$H(X - Z) \leq H(X - Y) + H(Y - Z) - H(Y)$$

2. 将 Y 替换为 $-Y$, 并注意到 $H(Y) = H(-Y)$, 即可得到该结果。

(d) 将 (c)-2 用于独立同分布的 X, Y, Z , 并利用 $H(Y) = H(X)$ 以及
 $H(X + Y) = H(Y + Z)$, 得到:

$$H(X - Z) + H(X) \leq 2H(X + Y)$$

现在替换 $X = Y'$ 且 $Z = Y''$, 即可得到: