CS225: Probability and Computing Homework 1

Zihao Ye

July 8, 2016

PROBLEM 1

(a) 用 0 表示原硬币生成了正面, 用 1 表示原硬币生成了反面.

每轮抛两次硬币, 如果为 01, 则视为正面, 如果为 10, 则视作反面, 否则进行下一轮.

如果这个过程终止, 容易证明 $Pr\{HEADS\} = Pr\{TAILS\}$.

需要进行的轮数为 $\frac{1}{2\rho(1-\rho)} < \frac{1}{\rho(1-\rho)}$

(b) 一个规则可以视为一个函数 $f : (X_1, X_2, \cdots, X_n) \longrightarrow (Z_1, Z_2, \cdots, Z_K, K)$.

其中 $\{X_i\}$ 为 biased coin 产生的正反面序列, $\{Z_i\}$ 表示由我们的规则产生的 unbiased 序列, 应当满足 $P(\mathbf{Z} = \mathbf{z}|K = k) = \frac{1}{2^k}$.

考虑输入的信息熵:

$$
\begin{aligned}
H(X_1, \cdots, X_n) &\geq H(Z_1, Z_2, \cdots, Z_K) = H(Z_1, Z_2, \cdots, Z_K, K) && (0.1)\\
&= H(K) + H(Z_1, \cdots, Z_K \mid K) && (0.2)\\
&= H(K) - E[\log Pr\{Z_1, \cdots, Z_k | K\}] && (0.3)\\
&= H(K) + E[K] && (0.4)\\
&\geq E[K] && (0.5)
\end{aligned}
$$

而 $X_1, \cdots, X_n$ 之间两两独立同分布, $H(X_1, \cdots, X_n) = nH(\rho)$, 由此得出:

$$E[K] \leq n(-\rho \log \rho - (1-\rho) \log(1-\rho))$$

下面考虑如何达到这个上界:

采用如下递归生成规则 g:

设输入的串为 A, 设串 X, Y 初始为空 (ϵ), 每次考虑 $A_{2k-1}A_{2k}$:

a) 若它为 01, 则生成出了一个 0, $Y_n = Y_n + 0$.

b) 若它为 10, 则生成出了一个 1, $Y_n = Y_n + 0$.

c) 若它为 00, 则 $X_n = X_n + 0$, $Y_n = Y_n + 1$.

d) 若它为 11, 则 $X_n = X_n + 1$, $Y_n = Y_n + 1$.

当整个串处理完成之后, 运行 $g(X), g(Y)$.

由于每次递归下去串的长度都减半, 此过程必然结束 (在 $\log n$ 层之后). 注意到, 假设输入串 A 每个字符满足 $B(1,\rho)$, 则 X 的每个字符满足 $B(1, \frac{\rho^2}{\rho^2+(1-\rho)^2})$, Y 的每个字符满足 $B(1, \rho^2+(1-\rho)^2)$.

用 $f(p)(p \in [0,1])$ 表示在每个字符满足 $B(1,p)$ 的情况下, 每个输入字符在此过程中生成出的均匀分布的字符数目的期望:

$$f(p) = pq + \frac{1}{2}\left(p^2+q^2\right) f\left(\frac{p^2}{p^2+q^2}\right) + \frac{1}{2} f\left(p^2+q^2\right)$$

满足 $f(p) = f(q), f(0) = f(1) = 0$, 令 S 为此类函数组成的函数空间.

令 $h : S \to S$ 为一高阶函数满足 $(hf) = \lambda p.pq + \frac{p^2+q^2}{2} f\left(\frac{p^2}{p^2+q^2}\right) + \frac{1}{2} f\left(p^2+q^2\right)$.

则 $H(p) = -p\log p - q\log q$ 为 h 的不动点: $h(H) = H$, 假设 h 存在其它的不动点 H', 则

$$(H-H') = \frac{p^2+q^2}{2}(H-H')\left(\frac{p^2}{p^2+q^2}\right) + \frac{1}{2}(H-H')\left(p^2+q^2\right)$$

若 $(H-H') \neq 0$, 则设 $M = \sup(H-H') > 0$, $(H-H')(x) \leq \frac{1+p^2+q^2}{2} M < M$, 矛盾.

故 h 存在唯一的不动点 H, 当 n 足够大时, 可以认为该规则 g 产生了最大可能数目的 unbiased coins.

PROBLEM 2

(a) 设抛 n 次硬币之后 HEADS 比 TAILS 多的数目最大值为 a_n, TAILS 比 HEADS 多的数目最大值为 b_n, $H = \max\{a_n, b_n\}$,

$$E[H] = E[\max\{a_n, b_n\}] \leq E[a_n] + E[b_n]$$

由于对称性 $E[a_n] = E[b_n]$, $E[H] \leq 2E[a_n]$.

设 X_n 为一个随机变量, 表示抛 n 次硬币之后 HEADS 比 TAILS 多的数目.

$$\begin{aligned}
E[a_n] &= \sum_{i=1}^{n} Pr\{a_n \geq i\} && (0.6)\\
&\leq \sum_{i=1}^{n} Pr\{X_n \geq i\} + Pr\{X_n \geq i+1\} && (0.7)\\
&= \left(2\sum_{i=1}^{n} Pr\{X_n \geq i\}\right) - Pr\{X_n \geq 1\} && (0.8)\\
&\leq 2\sum_{i=1}^{n} Pr\{X_n \geq i\} && (0.9)
\end{aligned}$$

其中(0.6)到(0.7)利用了对称性: 将 $a_n \ge i$ 关于 X_n 是否等于 a_n 分类, 如果等于, $Pr\{a_n \ge i\} = Pr\{X_n \ge i\}$, 否则由对称性 $Pr\{a_n \ge i\} = Pr\{X_n \ge i+1\}$.

由于 $E[X_n] = 0$, $Var[X_n] = nVar[X_1] = n$, $E[X_n^2] = Var[X_n] - E[X_n]^2 = n$.
由 Chebyshev Inequality 可以得到 $Pr\{X_n \ge i\}$ 的一个上界:

$$Pr\{X_n \ge i\} = Pr\{X_n \le -i\} = \frac{1}{2}Pr\{X_n^2 \ge i^2\} \le \frac{1}{2}\frac{E[X_n^2]}{i^2} = \frac{n}{2i^2} \tag{0.10}$$

将(0.10)代入(0.9), 得到:

$$\begin{aligned} E[a_n] &\le 2\sum_{i=1}^{\lfloor\sqrt{n}\rfloor} 1 + n\sum_{i=\lceil\sqrt{n}\rceil} n\frac{1}{i^2} \\ &\le 2\sqrt{n} + \frac{n}{\sqrt{n}-1} \\ &= O\left(\sqrt{n}\right) \end{aligned}$$

由于 $E[H] \le 2E[a_n]$, 得到

$$E[H] = O\left(\sqrt{n}\right) \tag{0.11}$$

.

从另一个方面考虑, 设 Y_n 为抛 n 次硬币之后 HEADS 比 TAILS 多的数目的绝对值.
有 $H \ge Y_n, E[H] \ge E[Y_n]$.

$$E[Y_n] = \frac{1}{2^{n-1}}\sum_{k=0}^{\lfloor\frac{n}{2}\rfloor}\binom{n}{k}(n-2k) \tag{0.12}$$

$$= \frac{n}{2^{n-1}}\sum_{k=0}^{\lfloor\frac{n}{2}\rfloor}\binom{n}{k} - \frac{n}{2^{n-2}}\sum_{k=0}^{\lfloor\frac{n}{2}\rfloor-1}\binom{n-1}{k} \tag{0.13}$$

$$= \begin{cases} n\binom{n-1}{\frac{n-1}{2}}\frac{1}{2^{n-1}} & 2\nmid n \\ n\binom{n}{\frac{n}{2}}\frac{1}{2^n} & 2\mid n \end{cases} \tag{0.14}$$

由于 $\binom{n}{\frac{n}{2}} = \Theta\left(\frac{2^n}{\sqrt{n}}\right)$, $E[Y_n] = \Theta\left(\sqrt{n}\right)$.
由此可推出 $E[H] = \Omega\left(\sqrt{n}\right)$, 联立(0.11), 得出 $E[H] = O\left(\sqrt{n}\right)$

(b) 设 $L(\pi), L'(\pi)$ 分别为 π 的最长上升子序列和最长不上升子序列的长度 (由于 π 中两两元素各不相同, 最长不上升子序列等价于最长下降子序列), 由偏序集的 Dilworth 定理,

$$L(\pi)\cdot L'(\pi) \ge n$$

因此对于每个 outcome 有: $L(\pi) + L'(\pi) \ge 2\sqrt{n}$,

$$2E[L(\pi)] \ge 2\sqrt{n} \Longrightarrow E[L(\pi)] \ge \sqrt{n}$$

PROBLEM 3

设 $S_1 = \{x_1, x_2, \cdots, x_n\}$, $S_2 = \{y_1, y_2, \cdots, y_n\}$ 令

$$FING(S_1, x) = (x - x_1)(x - x_2)\cdots(x - x_n)$$

$$FING(S_2, x) = (x - y_1)(x - y_2)\cdots(x - y_n)$$

令 p 为一个大于 n 的素数,每次取 $\mathbb{F}_p$ 中的一个随机变量 u,在 $\mathbb{F}_p$ 中计算 $FING(S_1, u), FING(S_2, u)$:
若 $S_1 = S_2$, 则一定有 $FING(S_1, u) = FING(S_2, u)$.
令 $G(x) = FING(S_1, x) - FING(S_2, x)$, 这是关于 x 的一个 $\leq n$ 次的首一多项式, 由唯一分解定理, 可以得出如果 $S_1 \neq S_2$, $G \neq 0$.
由代数基本定理 $G(x)$ 最多有 n 个根, 因此 $G(u) = 0$ 的概率 $\leq \frac{n}{p}$, 且 $Pr\{G(u) = 0, G(v) = 0\} = Pr\{G(u) = 0\}Pr\{G(v) = 0\}$.
所以我们重复这个算法 T 次, 每次选一个随机变量 u, 然后判断 $G(u)$ 是否等于 0, 如果要做到以 $(1 - \epsilon)$ 的概率 $S_1 = S_2$, 则需要 $(\frac{n}{p})^T < \epsilon$, 令 $T = \log(\epsilon)/(\log(n) - \log(p))$ 即可.
因此总的时间复杂度为 $O(T \times n)$, 传输的数据量为 $T\log(p)$.

PROBLEM 4

Proof. 仿照 Schwarz Zippel 定理的证明, 对 n 使用归纳法:

1. 当 $n = 1$ 时: $Q(r_1)$ 为一个关于 r_1 的 d_1 次多项式, 由代数基本定理, 有至多 d_1 个根, 因此:
$$Pr\{Q(r_1) = 0 \mid Q \neq 0\} \leq \frac{d_1}{S_1}$$

2. 当 $n > 1$ 时: $Q \neq 0$, 可以将 $Q(r_1, r_2, \cdots, r_n)$ 视作关于 r_n 的 k_n 次多项式,
$$Q(r_1, r_2, \cdots, r_n) = x_n^{d_k} f_{d_n}(r_1, r_2, \cdots, r_{n-1}) + \overline{f}(x_1, x_2, \cdots, x_n)$$
且 $f_{d_n} \neq 0$, 结合 I.H. 和可得:
$$\begin{aligned} Pr\{f(r_1, r_2, \cdots, r_n) = 0\} &= Pr\{f(r_1, r_2, \cdots, r_{n-1}) = 0 \mid f_{d_n}(r_1, r_2, \cdots, r_{n-1}) = 0\}Pr\{f_{d_n}(r_1, r_2, \cdots, r_{n-1}) = 0\} \\ &+ Pr\{f(r_1, r_2, \cdots, r_{n-1}) = 0 \mid f_{d_n}(r_1, r_2, \cdots, r_{n-1}) \neq 0\}Pr\{f_{d_n}(r_1, r_2, \cdots, r_{n-1}) \neq 0\} \\ &\leq \sum_{i=1}^{n-1} \frac{d_i}{|S_i|} + \frac{d_n}{|S_n|} \\ &= \sum_{i=1}^{n} \frac{d_i}{|S_i|} \end{aligned}$$

$\square$

PROBLEM 5

为了方便讨论,我们将 $S(v)$ 的 size 限定为 $10d$, 定义 bad event 如下:
$A=(x,y,c)$ 表示 $(x,y)\in E$ 且点 x 和点 y 都选择了颜色 c 的事件, $p(A)\le\frac{1}{10d}\times\frac{1}{10d}=\frac{1}{100d^2}$.
下面考虑所有这些事件组成的 dependency graph, 设事件 $A=(x_1,y_1,c)$, 与它有关的事件至少包含 x_1,y_1 其中之一, 与 x_1 有关的事件 $\le 10d\times d=10d^2$, 与 x_2 有关的事件 $\le 10d\times d=10d^2$, 因此 A 的 $deg\le 20d^2-1$. 由于 A 的任意性, dependency graph 中所有点的度 $\le 20d^2-1$.
由 Lovász Local Lemma, $ep(d+1)=\frac{e\cdot 20d^2}{100d^2}<1$, 存在合法的染色.

PROBLEM 6

PROBLEM 7

原问题等价于将 $10kn$ 个 ball 放入 n 个 bin 中, 要求以 ≥ 0.9 的概率每个 bin 中有 $\ge k$ 个 ball, 求 k 的范围.
考虑每个 bin, $E[X_i]=10k$, 由 *Chernoff Bound*,

$$Pr\{X_i<(1-0.9)10k\}<\left(\frac{e^{-0.9}}{0.1^{0.1}}\right)^{10k}$$

$$\Pr\{\exists i,X_i<k\}\le n\cdot Pr\{X_i<k\}\le n\left(\frac{e^{-0.9}}{0.1^{0.1}}\right)^{10k}$$

如果 $n\left(\frac{e^{-0.9}}{0.1^{0.1}}\right)^{10k}\le 0.1$, 则原题条件满足,
解以上不等式, 得出 $k\ge\frac{\log(n)+\log(10)}{9-\log(10)}\approx 0.149\log(n)+0.344$.

PROBLEM 8

使用 coupling 的思想.
首先证明: $\frac{n}{2}$ 个小球随机放入 n 个 bin 中, 则 max load 的期望为 $\Theta\left(\frac{\log n}{\log\log n}\right)$.

Proof. 考虑 n 个在 $[n]$ 中均匀分布的 mutually independent 的随机变量, 表示随机放入 n 个 bin 中的 n 个 ball, 他们中前半部分和后半部分分别构成 $\frac{n}{2}$ 个独立同分布的随机变量, 且前后部分之间独立. 对于任意一种 outcome:$\mathbf{r}=\mathbf{r^1}+\mathbf{r^2}$($\mathbf{r^1}$,$\mathbf{r^2}$ 分别代表前一半和后一半的随机变量, '+' 表示拼接, 下同), $\mathbf{r}$ 产生的 max load 为 $L(\mathbf{r})$, $\mathbf{r^1}$ 的 max load 为 $L(\mathbf{r^1})$, $\mathbf{r^2}$ 的 maxload 为 $L(\mathbf{r^2})$.
容易证明对于任意一个 outcome, $L(\mathbf{r})>L(\mathbf{r^1})$, 由此可以得出

$$E[L(\mathbf{r})]\ge E[L(\mathbf{r^1})]$$

又由于 $L(\mathbf{r})<L(\mathbf{r^1})+L(\mathbf{r^2})$, 得出

$$E[L(\mathbf{r})]\le 2E[L(\mathbf{r^1})]$$

因此 $\frac{1}{2}E[L(\mathbf{r})] \le E[L(\mathbf{r^1})] \le E[L(\mathbf{r})]$.
由于 $E[L(\mathbf{r})] = \Theta\left(\frac{\log n}{\log\log n}\right)$, 得出 $E[L(\mathbf{r})] = \Theta\left(\frac{\log n}{\log\log n}\right)$. □

考虑 $3n$ 个在 $[n]$ 中均匀分布的 mutually independent 的随机变量表示在此规则下这个过程所需要的所有随机变量. 对于任意一个 outcome:$\mathbf{x} = \mathbf{x^1} + \mathbf{x^2} + \mathbf{x^3}$(其中 $\mathbf{x^1}$ 生成 one choice 的部分, $\mathbf{x^2}$,$\mathbf{x^3}$ 生成 two choice 的部分, 下同), 三个规则都满足如下性质:

1. $$L(\mathbf{x}) \ge L(\mathbf{x^1})$$

   (对于每个 two choice(x_i^2, x_i^3), 不做任何操作即得到右式的 max load)

2. $$L(\mathbf{x}) \le L(\mathbf{x^1}) + L(\mathbf{x^2}) + L(\mathbf{x^3})$$

   (对于每个 two choice(x_i^2, x_i^3), 在 x_i^2 和 x_i^3 上各放一个 ball, 不等式由归纳可证)

由此可得 $L(\mathbf{x^1}) \le L(\mathbf{x}) \le L(\mathbf{x^1}) + L(\mathbf{x^2}) + L(\mathbf{x^3})$,
求期望得

$$E[L(\mathbf{x^1})] = \Theta\left(\frac{\log n}{\log\log n}\right)$$

ACKNOWLEDGEMENTS

第一题我原本并没有想到利用每两个字符是 00,11 类型还是 10,01 类型这一信息, 所以只推出了 $f(p) = pq + \frac{1}{2}\left(p^2 + q^2\right) f\left(\frac{p^2}{p^2+q^2}\right)$ 这一式子, 这个构造方法是在高宇学长的解法中看到的.
用信息熵推出 $E(K)$ 的上界来自 *Elements of Information Theory*, 习题 2-17 的 hint.
感谢游宇榕提供了第二题中 $Pr\{a_n \ge i\} = Pr\{X_n \ge i\} + Pr\{X_n \ge i+1\}$ 这一重要的基于对称性的式子.
感谢刘志健同学提供了第三题的函数形式.