信息学中的概率统计：作业六

截止日期：2025 年 12 月 10 日（周三）下午 3 点前。请务必通过教学网提交电子版。可下课前同时提交纸质版。

与本次作业相关的事宜,请发邮件给我(ruosongwang@pku.edu.cn),抄送研究生助教叶昊洋(yhyfhgs@gmail.com),以及负责本次作业的本科生助教王楷斐（wkf5094@stu.pku.edu.cn）。

注意：本次作业第六题第四问为**附加问**，正确解决该题目可以得到额外 10% 的分数。

第一题

令 $X \sim \text{Exp}(1)$。本题中，我们将对 $a > 1$ 给出 $P(X \geq a)$ 的上界。

(1) 使用马尔可夫不等式，给出 $P(X \geq a)$ 的上界。

(2) 使用切比雪夫不等式，证明 $P(X \geq a) \leq \frac{1}{(a-1)^2}$。

(3) 对于任意正整数 k，计算 $E(X^k)$，并证明 $P(X \geq a) \leq \frac{k!}{a^k}$。

(4) 使用 Chernoff Bound，证明 $P(X \geq a) \leq a \cdot e^{-a+1}$。

(5) 计算 $P(X \geq a)$ 的准确值。

(6) 比较第四问中 Chernoff Bound 给出的上界与第三问中（选取最佳 k 时）给出的上界，指出哪一种方法能提供更紧的上界。

提示：对固定的 $a > 1$ 和 $t > 0$，令 $K \sim \pi(ta)$，$f(k) = E(X^k)/a^k$，并计算 $E(f(K))$。

第二题

在课上，我们介绍了随机变量的收敛性。设 $\{X_n\}$ 为一列随机变量，X 为另一随机变量。如果对于任意 $\epsilon > 0$，有

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1,$$

则称 $\{X_n\}$ 依概率收敛于 X，写作 $X_n \xrightarrow{P} X$。在本题中，我们将介绍随机变量的另一种收敛性。

设 $\{X_n\}$ 为一列随机变量，X 为另一随机变量。如果对于任意 $\epsilon > 0$，

$$\lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} |X_m - X| \geq \epsilon\right) = 0,$$

则称 $\{X_n\}$ 几乎必然收敛于 X，写作 $X_n \xrightarrow{a.s.} X$。

(1) 令 $\{X_n\}$ 为一列相互独立的随机变量，且 $X_n \sim B(1, 1/n)$。证明 $\{X_n\}$ 依概率收敛于 0，但 $\{X_n\}$ 不几乎必然收敛于 0。

(2) 令 $\{X_n\}$ 为一列独立同分布的随机变量，$X_n \sim B(1, p)$。令 $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$。证明 $Y_n \xrightarrow{a.s.} p$。

第三题

有两个用于判断给定输入正整数是否为质数的程序 A 和 B，时间复杂度均为 T。两个程序均使用随机性，但程序使用的随机性不会影响其时间复杂度。

程序 A 和 B 满足

- 若输入正整数为质数，程序 A 总输出 “质数”，而程序 B 以 1/2 的概率输出 “质数”，以 1/2 的概率输出 “合数”；
- 若输入正整数为合数，程序 A 以 1/2 的概率输出 “合数”，以 1/2 的概率输出 “质数”，而程序 B 总输出 “合数”。

通过调用程序 A 和 B，构造程序 C，使得 C 总能正确判断给定输入正整数是否为质数，而程序 C 运行时间的数学期望为 $O(T)$。

第四题

在课上，我们用 Chernoff bound 证明了下述不等式：若 $X \sim B(n,p)$，则

$$P(X \geq E(X) + n\epsilon) \leq e^{-2n\epsilon^2},$$

$$P(X \leq E(X) - n\epsilon) \leq e^{-2n\epsilon^2}。$$

在本题中，我们将对二项分布证明另一版本的 Chernoff bound。

(1) 证明 $M_X(t) \leq e^{(e^t-1)\cdot E(X)}$。提示：使用不等式 $1 + x \leq e^x$。

(2) 证明对于任意 $\epsilon > 0$，

$$P(X \geq (1+\epsilon)E(X)) \leq \left(\frac{e^{\epsilon}}{(1+\epsilon)^{1+\epsilon}}\right)^{E(X)};$$

对于任意 $0 < \epsilon < 1$，

$$P(X \leq (1-\epsilon)E(X)) \leq \left(\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right)^{E(X)}。$$

提示：参考课上关于泊松分布的 Chernoff bound。

(3) 有 n 个球，每个球都等可能被放到 $m = n$ 个桶中的任一个。令 X_i 表示第 i 个桶中球的数量，$Y = \max\{X_1, X_2, \ldots, X_n\}$。证明存在常数 $c_1 > 0$，$P(Y \geq c_1 \ln n / \ln\ln n) \leq 1/n$。这里只需要对充分大的正整数 n 证明结论。

(4) 若 n 为充分大的正整数，证明存在常数 $c_2 > 0$，$E(Y) \leq c_2 \ln n / \ln\ln n$。

第五题

在课上，我们证明了下述结论：对于任意向量 $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$, 令 $A \in \mathbb{R}^{k\times d}$ 为随机矩阵，A 的不同元素独立同分布且均服从 $N(0,1)$，$k = O(\log n/\epsilon^2)$，则以至少 1/2 的概率，对于任意 $1 \leq i, j \leq n$，

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left\|\frac{1}{\sqrt{k}}A(x_i - x_j)\right\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2,$$

也即令 $F(x) = \frac{1}{\sqrt{k}}Ax$ 为一随机线性变换，则以至少 1/2 的概率，$F(x)$ 保持了每一对 x_i 和 x_j 之间的距离。

证明该结论的核心工具是下述引理：对于任意 $x \in \mathbb{R}^d$，

$$P\left((1-\epsilon)\|x\|^2 \le \left\|\frac{1}{\sqrt{k}}Ax\right\|^2 \le (1+\epsilon)\|x\|^2\right) \ge 1 - 2e^{-k\epsilon^2/8}。 \tag{1}$$

为了证明原结论，对所有可能的 $x = x_i - x_j$ 使用上述结论，并使用 Union bound。

在本题中，我们将证明随机线性变换 $F(x) = \frac{1}{\sqrt{k}}Ax$ 不仅可以保持每一对 x_i 和 x_j 之间的距离，还可以保持每一对 x_i 和 x_j 之间的点积。在本题中，对于向量 $a, b \in \mathbb{R}^d$，$\langle a, b\rangle = a^\top b$ 为向量 a 与 b 的点积。

(1) 考虑向量 $y_1, y_2, \ldots, y_n \in \mathbb{R}^d$，对于全部 $1 \le i \le n$，满足 $\|y_i\| = 1$。令 $A \in \mathbb{R}^{k\times d}$ 为随机矩阵，A 的不同元素独立同分布且均服从 $N(0,1)$，$k = O(\log n/\epsilon^2)$ 。证明以至少 1/2 的概率，下述事件同时成立：

- 对于任意 $1 \le i \le n$，$(1-\epsilon/4)\|y_i\|^2 \le \left\|\frac{1}{\sqrt{k}}Ay_i\right\|^2 \le (1+\epsilon/4)\|y_i\|^2$；
- 对于任意 $1 \le i, j \le n$ 且 $i \ne j$，$(1-\epsilon/4)\|y_i+y_j\|^2 \le \left\|\frac{1}{\sqrt{k}}A(y_i+y_j)\right\|^2 \le (1+\epsilon/4)\|y_i+y_j\|^2$

(2) 在（1）中结论的基础上，证明以至少 1/2 的概率，对于任意 $1 \le i, j \le n$，

$$\left|\left\langle \frac{1}{\sqrt{k}}Ay_i, \frac{1}{\sqrt{k}}Ay_j\right\rangle - \langle y_i, y_j\rangle\right| \le \epsilon。$$

(3) 考虑向量 $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$。注意 x_i 不一定满足 $\|x_i\| = 1$。证明以至少 1/2 的概率，对于任意 $1 \le i, j \le n$，

$$\left|\left\langle \frac{1}{\sqrt{k}}Ax_i, \frac{1}{\sqrt{k}}Ax_j\right\rangle - \langle x_i, x_j\rangle\right| \le \epsilon\|x_i\|\|x_j\|。$$

(4) 使用上一问中的结论，证明对于任意正整数 n，存在矩阵 $A \in \mathbb{R}^{n\times n}$ 满足 $\mathrm{rank}(A) \le c\log n$，这里 c 为与 n 无关的常数，且对于 $n \times n$ 的单位矩阵 I，满足对于任意 $1 \le i, j \le n$，$|I_{i,j} - A_{i,j}| \le 0.1$，也即 $I - A$ 任意一个元素的绝对值不超过 0.1。

第六题

$X_1, X_2, \ldots, X_n$ 为独立同分布的随机变量，X_i 服从柯西分布，其概率密度函数满足对于任意实数 x，$f(x) = \frac{1}{\pi(x^2+1)}$。令 $Y_1 = |X_1|, Y_2 = |X_2|, \ldots, Y_n = |X_n|$。

提示：在本题中，请**避免**先将分布函数化为含 arctan 的形式并据此进行分析；请**直接从密度函数出发**进行分析。

(1) 给出 Y_i 的边际密度函数。

(2) 证明存在常数 $c_1 > 0$，

$$P\left(\sum_{i=1}^n Y_i \le c_1 n^2\right) \ge 2/3。$$

(3) 证明存在常数 $c_2 > 0$，

$$P\left(\sum_{i=1}^n Y_i \ge c_2 n\right) \ge 2/3。$$

(4) **附加问**。给出函数 $f(n)$，使得对于充分大的正整数 n，存在与 n 无关的常数 $c_3, c_4 > 0$，

$$P\left(\sum_{i=1}^n Y_i \le c_3 f(n)\right) \ge 2/3,$$

$$P\left(\sum_{i=1}^{n} Y_i \geq c_4 f(n)\right) \geq 2/3。$$

请注意，只有写出正确的 $f(n)$ 并给出**完全正确**的证明才可以获得额外的分数（10%）。