Homework 6

姓名: 方嘉聪 学号: 2200017849

Problem 1. 令 $X \sim \text{Exp}(\lambda), \lambda > 0$. 本题中, 我们将对 $a > 1$ 给出 $\mathbb{P}(X \geq a/\lambda)$ 的上界.

(1) 使用 Markov 不等式, 给出 $\mathbb{P}(X \geq a/\lambda)$ 的上界.

(2) 使用 Chebyshev 不等式, 证明

$$\mathbb{P}(X \geq a/\lambda) \leq \frac{1}{(a-1)^2}$$

(3) 使用 Chernoff 不等式, 证明

$$\mathbb{P}(X \geq a/\lambda) \leq a \cdot e^{-a+1}$$

(4) 计算 $\mathbb{P}(X \geq a/\lambda)$ 的准确值.

◀

Solution. 由于 $X \sim \text{Exp}(\lambda)$, 故 $\mathbb{E}(X) = \lambda^{-1}, \mathbb{E}(X^2) = 2\lambda^{-2}, \sigma^2 = \lambda^{-2}$.

(1) 由 Markov 不等式, 有

$$\mathbb{P}\left(X \geq \frac{a}{\lambda}\right) \leq \frac{1}{a}.$$

(2) 由 Chebyshev 不等式, 有

$$\mathbb{P}\left(X \geq \frac{a}{\lambda}\right) = \mathbb{P}\left(X - \mathbb{E}(X) \geq (a-1) \cdot \sigma\right) \leq \frac{1}{(a-1)^2}.$$

(3) 我们先来计算矩生成函数 $M_X(t)$:

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_0^\infty \lambda e^{(t-\lambda)x} \mathrm{d}x = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

那么由 Chernoff Bound 有

$$\mathbb{P}\left(X \geq \frac{a}{\lambda}\right) = \mathbb{P}\left(e^{tX} \geq e^{ta/\lambda}\right) \leq \min_{t<\lambda} \left\{e^{-ta/\lambda} \cdot \frac{\lambda}{\lambda - t}\right\}$$

而

$$g(t) := e^{-ta/\lambda} \cdot \frac{\lambda}{\lambda - t}, \quad g'(t) = 0 \implies t^* = \lambda - \frac{\lambda}{a}$$

故有

$$\mathbb{P}\left(X \geq \frac{a}{\lambda}\right) \leq g(t^*) = a \cdot e^{-a+1}.$$

(4) 直接求积分

$$\mathbb{P}\left(X \geq \frac{a}{\lambda}\right) = \int_{a/\lambda}^\infty \lambda e^{-\lambda x} \mathrm{d}x = e^{-a}.$$

◁

Problem 2. 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. 若对于任意 $\varepsilon > 0$, 有

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

则称 $\{X_n\}$ *依概率收敛* 于 X, 记作 $X_n \xrightarrow{P} X$. 在本题, 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. 若 $\mathbb{P}(\lim_{n\to\infty} X_n \to X) = 1$, 也即对于任意 $\varepsilon > 0$, 有

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

则称 $\{X_n\}$ *几乎必然收敛* 于 X, 记作 $X_n \xrightarrow{a.s.} X$.

(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, 且 $X_n \sim B(1, p)$. 令 $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. 证明 $Y_n \xrightarrow{a.s.} p$.

◀

Solution. (1) 若 $\varepsilon > 1$, 那么 $\mathbb{P}(|X_n| < \varepsilon) = 1$, 符合依概率收敛的定义. 考虑 $\varepsilon \in (0, 1]$, 那么

$$\mathbb{P}(|X_n| < \varepsilon) = 1 - \frac{1}{n} \xrightarrow{n\to\infty} 1.$$

故 $X_n \xrightarrow{P} 0$. 任意 $\varepsilon > 0$, 记事件 $A_n(\varepsilon) := \{|X_n - X| \geq \varepsilon\}$. 那么

$$\mathbb{P}(A_n(\varepsilon)) = \mathbb{P}(X_n = 1) = \frac{1}{n} \implies \mathbb{P}\left(\bigcup_{m=n}^{\infty} A_m(\varepsilon)\right) = 1 - \prod_{m=n}^{\infty}\left(1 - \frac{1}{m}\right) \xrightarrow{n\to\infty} 1$$

故 X_n 不几乎必然收敛于 0.

(2) 类似的, 记 $B_n(\varepsilon) := \{|Y_n - p| \geq \varepsilon\}$. 那么只需验证

$$\mathbb{P}\left(\bigcup_{m=n}^{\infty} B_m(\varepsilon)\right) \xrightarrow{n\to\infty} 0.$$

我们先估计 $(\forall m \geq n)$, 由 Markov 不等式, 记 $S_m = \sum_{i=1}^{m}(X_i - p)$, 考虑四阶矩有:

$$\mathbb{P}(B_m(\varepsilon)) = \mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}(X_i - p)\right| \geq \varepsilon\right) = \mathbb{P}\left(\left|\frac{S_m}{m}\right| \geq \varepsilon\right) \leq \frac{\mathbb{E}(S_m^4)}{m^4\varepsilon^4}$$

在课上我们已经计算了, 若 $X \sim B(n, p)$, 那么

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}(X))^4] &= np(1-p)^4 + n(1-p)p^4 + 3n(n-1)p^2(1-p)^2 \\ &\leq n^2\left(p(1-p)^4 + (1-p)p^4 + 3p^2(1-p)^2\right)\end{aligned}$$

记 $C(p) = p(1-p)^4(1-p)p^4 + 3p^2(1-p)^2$ 为与 n 无关的常数.
而 $X_i \overset{\text{i.i.d}}{\sim} B(1, p) \implies \sum_{i=1}^{m} X_i \sim B(m, p)$, 故

$$\mathbb{P}(B_m(\varepsilon)) \leq \frac{\mathbb{E}(S_m^4)}{m^4\varepsilon^4} \leq \frac{m^2 \cdot C(p)}{m^4\varepsilon^4} = \frac{C(p)}{m^2\varepsilon^4}$$

故

$$\mathbb{P}\left(\bigcup_{m=n}^{\infty} B_m(\varepsilon)\right) \le \sum_{m=n}^{\infty} \mathbb{P}(B_m(\varepsilon)) \le \sum_{m=n}^{\infty} \frac{C(p)}{m^2\varepsilon^4} = \frac{C(p)}{\varepsilon^4}\sum_{m=n}^{\infty}\frac{1}{m^2} \xrightarrow{n\to\infty} 0$$

即 $Y_n \xrightarrow{a.s.} p$. 证毕.

注: 这题也可以用 Union Bound + Chernoff Bound 来证明.

◁

Problem 3. 某个不使用随机性的计算机程序 A, 为了输出正确的结果, 该程序需要对另一计算机程序 B 进行 T 次调用, 每次调用使用可能不同的输入, 且每次调用使用的输入依赖于之前对程序 B 的调用返回的结果. 程序 A 使用对程序 B 的 T 次调用返回的结果以输出最终结果 θ. 具体来说, 假设对程序 B 进行 T 次调用返回的结果为 $\omega_1, \omega_2, \cdots, \omega_T$, 在正确得到 $\omega_1, \omega_2, \cdots, \omega_T$ 的前提下, 程序 A 总是能输出正确的结果 θ.

现有计算机程序 B'. 在同样的输入下, 程序 B' 以 $2/3$ 的概率返回与程序 B 相同的结果, 以 $1/3$ 的概率返回不同的结果. 现在, 没有程序 B, 仅有程序 A, B' 的情况下, 设计一个方案, 以 $1-\delta$ 的概率输出正确的结果 θ. 该方案对程序 A, B' 的调用次数应与 T 和 $\log(1/\delta)$ 为多项式关系. ◀

Solution. 我们先证明一个引理:

Lemma 1. （伯努利不等式）若 $x > -1, n \in \mathbb{N}^+$, 则有 $(1+x)^n \ge 1 + nx$.

证明. 用数学归纳法. 容易验证 $n = 1, n = 2$ 时均成立. 假设对于 $n = k$ 成立, 考虑 $n = k + 2$ 时:

$$\begin{aligned}(1+x)^{k+2} &= (1+x)^k(1+x)^2 \ge (1+kx)(1+2x+x^2) \\ &= 1 + (k+2)x + kx^2(k+2) + x^2 \ge 1 + (k+2)x.\end{aligned}$$

引理证毕. □

回到本题, 考虑一个如下的算法 (即重复运行 T 轮, 每轮调用 B' 共 t_0 次, 选择出现次数最多的结果):

Algorithm 1 $1-\delta$ **Algorithm**

Require: 输入 x, 程序 A, 程序 B', 参数 T, δ. 待定参数 $t_0 > 0$.

Ensure: 输出 θ

```
已经得到的返回结果为 s = ∅
for i = 1 to T do
    根据 s 和 x, 重复调用程序 B' 共 t_0 次.
    记返回结果为 y_1, y_2, ··· , y_{t_0}.
    y = argmax_{y_i} Σ_{j=1}^{t_0} 𝟙(y_i = y_j)
    s.append(y)
将 s 作为输入, 调用程序 A.
return A(s)
```

对于 $i \in [T]$, 记事件 E_i 为前 $i-1$ 轮调用结果与 B 相同的条件下, 第 i 轮调用 B' 输出正确结果. 记 $X_{i,j} (\forall j \in [t_0])$ 为前 $i-1$ 轮调用结果与 B 相同的条件下, 第 i 轮调用 B' 输出错误结果. 那么

$$Y_i := \sum_{j=1}^{t_0} \mathbb{1}_{X_{i,j}} \sim B(t_0, 1/3).$$

在课上已经证明了, 在一轮调用中出现频率最高的结果为错误结果的上界为:

$$\mathbb{P}\left(Y_i \geq \frac{t_0}{2}\right) \leq e^{-t_0/18}$$

即 $\mathbb{P}(E_i) \geq 1 - e^{-t_0/18}$. 那么上述算法输出正确结果的概率为

$$\begin{aligned}
\mathbb{P}(A \text{ 输出正确结果}) &= \prod_{i=1}^{T}\left(1 - e^{-t_0/18}\right) = (1 - e^{-t_0/18})^T \\
&\geq 1 - T \cdot e^{-t_0/18} \quad \text{（伯努利不等式）} \\
&\geq 1 - \delta. \quad \text{取 } t_0 = 18\log(T/\delta)
\end{aligned}$$

B' 总调用次数为 $T \cdot t_0 = O(T\log(T/\delta)) = \mathrm{poly}(T, \log(1/\delta))$. 且算法正确性概率为 $1-\delta$. 证毕. ◁

Problem 4. 在课上, 我们用 Chernoff Bound 证明了下属不等式: 若 $X \sim B(n,p)$, 则

$$\begin{aligned}
\mathbb{P}(X \geq \mathbb{E}(X) + n\varepsilon) \leq e^{-2n\varepsilon^2}. \\
\mathbb{P}(X \leq \mathbb{E}(X) - n\varepsilon) \leq e^{-2n\varepsilon^2}.
\end{aligned}$$

在本题中, 我们将对二项分布证明另一版本的 Chernoff Bound.

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

(2) 证明对于任意 $\varepsilon > 0$, 有

$$\mathbb{P}(X \geq (1+\varepsilon)\mathbb{E}(X)) \leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\mathbb{E}(X)}.$$

对于任意 $0 < \varepsilon < 1$, 证明

$$\mathbb{P}(X \leq (1-\varepsilon)\mathbb{E}(X)) \leq \left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{\mathbb{E}(X)}.$$

提示: 参考作业二第六题.

(3) 利用 (2) 中的结论, 重新证明作业二第二题 (3). 即, 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, $Y = \max\{X_1, X_2, \cdots, X_n\}$. 证明

$$\mathbb{P}(Y \geq 4\log_2 n) \leq \frac{1}{n}$$

◀

Solution. (1) 对于二项分布 $X \sim B(n,p)$, 有

$$M_X(t) = \left(1 - p + pe^t\right)^n \leq e^{n(-p+pe^t)} = e^{(e^t-1)\mathbb{E}(X)}.$$

(2) 对于任意 $\varepsilon > 0$, 有 (对于任意 $t > 0$, 类似 Chernoff Bound 的证明)

$$\begin{aligned}
\mathbb{P}(X \geq (1+\varepsilon)\mathbb{E}(X)) &= \mathbb{P}(e^{tX} \geq e^{t(1+\varepsilon)\mathbb{E}(X)}) \\
&\leq \min_{t>0}\left\{e^{-t(1+\varepsilon)\mathbb{E}(X)}M_X(t)\right\} \quad \text{(Markov inequality)} \\
&\leq \min_{t>0}\left\{\exp\left(\mathbb{E}(X)[e^t - 1 - (1+\varepsilon)t]\right)\right\}
\end{aligned}$$

记 $g(t) = e^t - 1 - (1+\varepsilon)t$, 那么 $g'(t) = e^t - \varepsilon - 1 = 0 \implies t^* = \log(1+\varepsilon) > 0$. 故

$$\mathbb{P}(X \geq (1+\varepsilon)\mathbb{E}(X)) \leq e^{\mathbb{E}(X)\cdot g(t^*)} = \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\mathbb{E}(X)}.$$

类似的, 对于 $0 < \varepsilon < 1$, 考虑 $t < 0$, 有

$$\begin{aligned}
\mathbb{P}(X \leq (1-\varepsilon)\mathbb{E}(X)) &= \mathbb{P}(e^{tX} \geq e^{t(1-\varepsilon)\mathbb{E}(X)}) \\
&\leq \min_{t<0}\left\{e^{-t(1-\varepsilon)\mathbb{E}(X)}M_X(t)\right\} \quad \text{(Markov inequality)} \\
&\leq \min_{t<0}\left\{\exp\left(\mathbb{E}(X)[e^t - 1 - (1-\varepsilon)t]\right)\right\}
\end{aligned}$$

记 $h(t) = e^t - 1 - (1-\varepsilon)t$, 那么 $h'(t) = e^t + \varepsilon - 1 = 0 \implies t^* = \log(1-\varepsilon) < 0$. 故

$$\mathbb{P}(X \leq (1-\varepsilon)\mathbb{E}(X)) \leq e^{\mathbb{E}(X)\cdot h(t^*)} = \left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{\mathbb{E}(X)}.$$

综上, 证毕.

(3) 我们来证明 $\mathbb{P}(X_i \geq 4\log_2 n) \leq 1/n^2$.

注意到 $X_i \sim B(n, 1/n), \mathbb{E}(X_i) = 1$. $n = 1$ 时显然成立. 考虑 $n \geq 2$, 取 $\varepsilon = 4\log_2 n - 1 > 0$. 那么

$$\begin{aligned}
\mathbb{P}(X_i \geq 4\log_2 n) = \mathbb{P}(X_i \geq (1+\varepsilon)\mathbb{E}(X_i)) &\leq \left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{\mathbb{E}(X_i)} \\
&= \frac{e^{4\log_2 n - 1}}{(4\log_2 n)^{4\log_2 n}} = \frac{1}{e}\cdot\frac{1}{n^{8-4/\ln 2}}\cdot\frac{1}{(\log_2 n)^{4\log_2 n}} \\
&\leq \frac{1}{n^2}. \quad (\text{注意到 } 8 - 4/\ln 2 > 2)
\end{aligned}$$

那么由 Union Bound, 有

$$\mathbb{P}(Y \geq 4\log_2 n) = \mathbb{P}\left(\bigcup_{i=1}^{n}\{X_i \geq 4\log_2 n\}\right) \leq \sum_{i=1}^{n}\mathbb{P}(X_i \geq 4\log_2 n) \leq \frac{1}{n}.$$

证毕.

◁

Problem 5. 在课上, 我们证明了下述结论: 对于任意向量 $x_1, x_2, \cdots, x_n \in \mathbb{R}^d$, 令 $A \in \mathbb{R}^{k\times d}$ 为随机矩阵, A 的不同元素独立同分布于 $\mathcal{N}(0,1)$, $k = O(\log n/\varepsilon^2)$, 则以至少 $1/2$ 的概率, 对于任意 $1 \leq i, j \leq n$, 有

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

也即令 $F(x)=\frac{1}{\sqrt{k}}Ax$ 为一随机线性变换, 则至少以 1/2 的概率, $F(x)$ 保持了每一对 x_i,x_j 之间的距离. 证明该结论的核心工具使下述引理: 对于任意 $x\in\mathbb{R}^d$, 有

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

为证明原结论, 对所有可能的 $x=x_i-x_j$ 使用上述引理, 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,x_j 之间的距离, 还可以保持每一对 x_i,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 a 和 b 的点积.

(1) 考虑向量 $y_1,y_2,\cdots,y_n\in\mathbb{R}^d$, 对于任意 $1\le i\le n$ 满足 $\|y_i\|=1$. 令 $A\in\mathbb{R}^{k\times d}$ 为随机矩阵, 且不同元素独立同分布于 $\mathcal{N}(0,1)$, $k=O(\log n/\varepsilon^2)$. 证明以至少 1/2 的概率, 下述事件同时成立:

- 对于任意 $1\le i\le n$, 有

$$\left(1-\frac{\varepsilon}{4}\right)\|y_i\|^2\le\left\|\frac{1}{\sqrt{k}}Ay_i\right\|^2\le\left(1+\frac{\varepsilon}{4}\right)\|y_i\|^2. \tag{2}$$

- 对于任意 $1\le i,j\le n$ 且 $i\ne j$, 有

$$\left(1-\frac{\varepsilon}{4}\right)\|y_i+y_j\|^2\le\left\|\frac{1}{\sqrt{k}}A(y_i+y_j)\right\|^2\le\left(1+\frac{\varepsilon}{4}\right)\|y_i+y_j\|^2. \tag{3}$$

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

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

(3) 考虑向量 $x_1,x_2,\cdots,x_n\in\mathbb{R}^d$. 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 $1\le i,j\le n$, 有

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

◀

Solution. (1) 令 $k=512\log n/\varepsilon^2=O(\log n/\varepsilon^2)$. 对于任意给定的 $y_i,y_j,(i\ne j)$ 由引理(1)有

$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{k}}Ay_i\right\|^2\notin(1\pm\varepsilon/4)\|y_i\|^2\right)\le 2e^{-k\varepsilon^2/128}.$$
$$\mathbb{P}\left(\left\|\frac{1}{\sqrt{k}}A(y_i+y_j)\right\|^2\notin(1\pm\varepsilon/4)\|y_i+y_j\|^2\right)\le 2e^{-k\varepsilon^2/128}.$$

记事件 A 表示对于任意 $1\le i\le n$, 都有(2)成立. 由 Union Bound, 有

$$\mathbb{P}(\bar{A})\le\sum_{i=1}^n\mathbb{P}\left(\left\|\frac{1}{\sqrt{k}}Ay_i\right\|^2\notin(1\pm\varepsilon/4)\|y_i\|^2\right)\le 2ne^{-k\varepsilon^2/128}$$

那么

$$\mathbb{P}(\bar{A})\le 2n\cdot\frac{1}{n^4}\le\frac{1}{4}.$$

记事件 B 表示对于任意 $1 \le i, j \le n$ 且 $i \ne j$, 都有(3)成立. 由 Union Bound, 有

$$\begin{aligned}\mathbb{P}(\bar{B}) &\le \sum_{i\ne j} \mathbb{P}\left(\left\| \frac{1}{\sqrt{k}} A(y_i + y_j) \right\|^2 \notin (1 \pm \varepsilon/4) \|y_i + y_j\|^2 \right) \\ &\le 2e^{-k\varepsilon^2/128} \cdot \frac{n(n-1)}{2} \le n^2 e^{-k\varepsilon^2/128}.\end{aligned}$$

那么

$$\mathbb{P}(\bar{B}) \le n^2 \cdot \frac{1}{n^4} \le \frac{1}{4}.$$

由 Union Bound, 有

$$\mathbb{P}(\bar{A} \cup \bar{B}) \le \mathbb{P}(\bar{A}) + \mathbb{P}(\bar{B}) \le \frac{1}{2}.$$

故以至少 $1/2$ 的概率, 事件 A 和 B 同时成立. 证毕.

(2) 类似 (1) 中结论的证明, 可得在相同的条件下 (即 $k = O(\log n/\varepsilon^2)$), 以至少 $1/2$ 的概率下述事件同时成立:

- 对于任意 $1 \le i, j \le n$ 且 $i \ne j$, 有

$$(1-\varepsilon) \|y_i + y_j\|^2 \le \left\| \frac{1}{\sqrt{k}} A(y_i + y_j) \right\|^2 \le (1+\varepsilon) \|y_i + y_j\|^2. \tag{4}$$

- 对于任意 $1 \le i, j \le n$ 且 $i \ne j$, 有

$$(1-\varepsilon) \|y_i - y_j\|^2 \le \left\| \frac{1}{\sqrt{k}} A(y_i - y_j) \right\|^2 \le (1+\varepsilon) \|y_i - y_j\|^2. \tag{5}$$

那么以至少 $1/2$ 的概率, 对于 $1 \le i, j \le n$ 且 $i \ne j$, 有

$$\begin{aligned}(1-\varepsilon) \|y_i + y_j\|^2 - (1+\varepsilon) \|y_i - y_j\|^2 &\le \left\| \frac{1}{\sqrt{k}} A(y_i + y_j) \right\|^2 - \left\| \frac{1}{\sqrt{k}} A(y_i - y_j) \right\|^2 \\ &\le (1+\varepsilon) \|y_i + y_j\|^2 - (1-\varepsilon) \|y_i - y_j\|^2. \\ \iff 4\langle y_i, y_j\rangle - \varepsilon \cdot (2\|y_i\|^2 + 2\|y_j\|^2) &\le 4 \cdot \left\langle \frac{1}{\sqrt{k}} A(y_i + y_j), \frac{1}{\sqrt{k}} A(y_i - y_j) \right\rangle \\ &\le 4\langle y_i, y_j\rangle + \varepsilon \cdot (2\|y_i\|^2 + 2\|y_j\|^2).\end{aligned}$$

注意到 $\|y_i\| = 1$, 那么等价于

$$\begin{aligned}-\varepsilon \le \left\langle \frac{1}{\sqrt{k}} A(y_i + y_j), \frac{1}{\sqrt{k}} A(y_i - y_j) \right\rangle - \langle y_i, y_j\rangle \le \varepsilon. \\ \iff \left| \left\langle \frac{1}{\sqrt{k}} A y_i, \frac{1}{\sqrt{k}} A y_j \right\rangle - \langle y_i, y_j\rangle \right| \le \varepsilon.\end{aligned}$$

证毕.

(3) 对于任意 x_i, 若 $\|x_i\| = 0$, 那么有 $x_i = 0$, 此时注意到对于任意的 $x_j, j \ne i$, 有

$$\left| \left\langle \frac{1}{\sqrt{k}} A x_i, \frac{1}{\sqrt{k}} A x_j \right\rangle - \langle x_i, x_j\rangle \right| = 0 \le \varepsilon \|x_i\| \cdot \|x_j\|.$$

故不妨设 $\|x_i\| \neq 0, \forall i$. 那么对于任意 i, 令 $y_i = x_i/\|x_i\|$, 那么由 (2) 的结论, 以至少 1/2 的概率, 对于任意 $1 \leq i, j \leq n$, 有

$$\left|\left\langle \frac{1}{\sqrt{k}}Ay_i, \frac{1}{\sqrt{k}}Ay_j \right\rangle - \langle y_i, y_j\rangle\right| \leq \varepsilon. \iff \left|\left\langle \frac{1}{\sqrt{k}}A\frac{x_i}{\|x_i\|}, \frac{1}{\sqrt{k}}A\frac{x_j}{\|x_j\|} \right\rangle - \left\langle \frac{x_i}{\|x_i\|}, \frac{x_j}{\|x_j\|}\right\rangle\right| \leq \varepsilon.$$
$$\iff \left|\left\langle \frac{1}{\sqrt{k}}Ax_i, \frac{1}{\sqrt{k}}Ax_j \right\rangle - \langle x_i, x_j\rangle\right| \leq \varepsilon \|x_i\| \cdot \|x_j\|.$$

这里用到了点积的线性性质, 即 $\left\langle \alpha\vec{a}, \beta\vec{b}\right\rangle = \alpha\beta\left\langle \vec{a}, \vec{b}\right\rangle$. 证毕.

$\triangleleft$

Problem 6(Bonus 30%). 在课上, 我们证明了对于任意 $S_1, S_2, \cdots, S_m \subseteq \{1, 2, \cdots, n\}$, 存在 $\chi : \{1, 2, \cdots, n\} \to \{-1, +1\}$, 使得对于任意 $1 \leq i \leq m$, 有

$$\mathrm{disc}_\chi(S_i) = \left|\sum_{j\in S_i} \chi(j)\right| \leq O(\sqrt{n \log m}).$$

在本题中, 我们将证明存在 $S_1, S_2, \cdots, S_m \subseteq \{1, 2, \cdots, n\}$, 对于任意 $\chi : \{1, 2, \cdots, n\} \to \{-1, +1\}$, 存在 $1 \leq i \leq m$, 使得

$$\mathrm{disc}_\chi(S_i) = \left|\sum_{j\in S_i} \chi(j)\right| \geq \Omega(\sqrt{n}).$$

也即可上给出的上界 $O(\sqrt{n \log m})$ 几乎是最优的.

(1) 证明下述反集中不等式: $X \sim B(n, 1/2)$, 存在常数 $c_1, c_2 > 0$, 使得

$$\mathbb{P}(X \geq n/2 + c_1 \cdot \sqrt{n}) \geq c_2.$$

提示: 该不等式有多种证明方法. 一种可能的思路是首先使用定量化的中心极限定理 (课上提到的 Berry-Esseen 定理) 建立二项分布与标准正态分布的联系，之后对标准正态分布证明反集中不等式。

(2) 令 S 为 $\{1, 2, \cdots, n\}$ 的子集, 对于每个 $j \in \{1, 2, \cdots, n\}$, $\mathbb{P}(j \in S) = 1/2$, 且不同 j 是否被包含在 S 中相互独立. 利用 (1) 中的结论, 证明存在常数 $c_3, c_4 > 0$, 对于任意 $\chi : \{1, 2, \cdots, n\} \to \{-1, +1\}$, 使得

$$\mathbb{P}\left(\left|\sum_{j\in S} \chi(j)\right| \geq c_3\sqrt{n}\right) \geq c_4.$$

(3) 证明存在 $m = O(n)$(也即对于某个常数 C, $m \leq Cn$) 个集合 $S_1, S_2, \cdots, S_m \subseteq \{1, 2, \cdots, n\}$ 和常数 $c > 0$, 使得对于任意 $\chi : \{1, 2, \cdots, n\} \to \{-1, +1\}$, 存在 $1 \leq i \leq m$, 使得

$$\left|\sum_{j\in S_i} \chi(j)\right| \geq c\sqrt{n}.. \tag{6}$$

提示: 考虑使用概率证法, 将 $S_1, S_2, \cdots, S_m$ 取为 $\{1, 2, \cdots, n\}$ 独立同分布的随机子集, 并扩展 (2) 中的分析.

最后编译时间: 2024 年 12 月 13 日

(4) 证明当 $m=n$ 时, (3) 中的结论同样成立.

◀

Solution. (1) 由于 $X \sim B(n,1/2)$, 记标准化后为 $\widetilde{X}$ 为

$$\widetilde{X} = \frac{X-n/2}{\sqrt{n}/2}$$

那么由 Berry-Esseen 定理, 存在常数 $t_0 < 0.4748$[1]

$$\left|\mathbb{P}(\widetilde{X} \geq x) - \mathbb{P}(Z \geq x)\right| \leq \frac{t_0}{\sqrt{n}}.$$

令 $x = 2c_1$, 则

$$\begin{aligned}
\mathbb{P}(X \geq n/2 + c_1 \cdot \sqrt{n}) = \mathbb{P}(\widetilde{X} \geq 2c_1) &\geq \mathbb{P}(Z \geq 2c_1) - \frac{t_0}{\sqrt{n}} \\
&= \frac{1}{\sqrt{2\pi}} \int_{2c_1}^{+\infty} e^{-x^2/2}\mathrm{d}x - \frac{t_0}{\sqrt{n}} \\
&= \frac{1}{2} - \int_0^{2c_1} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\mathrm{d}x - \frac{t_0}{\sqrt{n}} \\
&\geq \frac{1}{2} - \frac{2c_1}{\sqrt{2\pi}} - \frac{t_0}{\sqrt{n}}, \quad (t_0 \text{ has a upper bound}) \\
&> \frac{1}{2} - \frac{2c_1}{\sqrt{2\pi}} - \frac{0.4748}{\sqrt{n}} := \hat{c}.
\end{aligned}$$

取 $c_1 = 10^{-5}$, 那么 $n \geq 4$ 时, 有

$$\hat{c} = \frac{1}{2} - \frac{2\cdot 10^{-5}}{\sqrt{2\pi}} - \frac{0.4748}{\sqrt{n}} \geq \frac{1}{2} - \frac{2\cdot 10^{-5}}{\sqrt{2\pi}} - \frac{0.4748}{2} \geq \frac{1}{4}.$$

而对于 $n < 4$ 的情况, 可以直接验证 $\mathbb{P}(X \geq n/2 + c_1 \cdot \sqrt{n}) \geq 1/4$. 故取 $c_2 = 1/4$, 证毕.

(2) 对于任意的 χ, 记

$$\begin{aligned}
T_+ &= \{j \in \{1,2,\cdots,n\} : \chi(j) = 1\}, \\
T_- &= \{j \in \{1,2,\cdots,n\} : \chi(j) = -1\}.
\end{aligned}$$

不妨 $|T_+| \geq |T_-|$, 那么有

$$\sum_{j\in S}\chi(j) = \sum_{j\in T_+} \mathbb{1}_{j\in S} - \sum_{j\in T_-}(1 - \mathbb{1}_{j\notin S}) = \sum_{j\in T_+} \mathbb{1}_{j\in S} + \sum_{j\in T_-} \mathbb{1}_{j\notin S} - \#\{T_- \cap S\}.$$

由于对任意的 j, $\mathbb{P}(j \in S) = 1/2$, 且不同的 j 是否在 S 中相互独立, 那么有

$$X := \sum_{j\in T_+} \mathbb{1}_{j\in S} + \sum_{j\in T_-} \mathbb{1}_{j\notin S} \sim B(n,1/2).$$

记 $k := \#\{T_- \cap S\} \leq n/2$, 那么有

$$\begin{aligned}
\mathbb{P}\left(\left|\sum_{j\in S}\chi(j)\right| \geq c_3\sqrt{n}\right) &= \mathbb{P}(|X-k| \geq c_3\sqrt{n}) = \mathbb{P}(X \geq c_3\sqrt{n} + k \cup X \leq k - c_3\sqrt{n}) \\
&\geq \mathbb{P}(X \geq n/2 + c_3\sqrt{n} \cup X \geq n/2 - c_3\sqrt{n}) = \mathbb{P}\left(|X-n/2| \geq c_3\sqrt{n}\right)
\end{aligned}$$

[1]来自 wiki 中提到的上界 https://en.wikipedia.org/wiki/Berry–Esseen_theorem

令 $c_3=c_1$, 由 (1) 的结论, 对称的有

$$\mathbb{P}(X\geq n/2+c_1\cdot\sqrt{n})\geq c_2.\quad \mathbb{P}(X\leq n/2-c_1\cdot\sqrt{n})\geq c_2 \implies \mathbb{P}\left(|X-n/2|\geq c_1\sqrt{n}\right)\geq 2c_2.$$

那么令 $c_4=2c_2$, 有

$$\mathbb{P}\left(\left|\sum_{j\in S}\chi(j)\right|\geq c_3\sqrt{n}\right)\geq 2c_2=c_4=\frac{1}{2}$$

证毕.

(3) 我们希望证明存在一组集合 $S_1,S_2,\cdots,S_m\subseteq\{1,2,\cdots,n\}$ 和常数 c, 使得对于任意的 χ, 至少存在一个 S_i, 使得 (6)成立.

我们考虑对于一组 $S_1,S_2,\cdots,S_m\subseteq\{1,2,\cdots,n\}$ 和常数 c, 存在一个 χ, 使得对于任意的 S_i, 有

$$\left|\sum_{j\in S_i}\chi(j)\right|<c\sqrt{n}. \tag{7}$$

成立的概率.

对于 $\{1,2,\cdots,n\}$ 的任意一个随机子集 S_i, 由 (2) 的结论, 对于任意的 χ, 有

$$\mathbb{P}\left(\left|\sum_{j\in S_i}\chi(j)\right|\geq c_3\sqrt{n}\right)\geq c_4=\frac{1}{2}.\implies \mathbb{P}\left(\left|\sum_{j\in S_i}\chi(j)\right|<c_3\sqrt{n}\right)<\frac{1}{2}.$$

取独立同分布的 m 个随机子集 $S_1,S_2,\cdots,S_m$, 其中设 $m=Cn$, C 为某个常数. 同时令 $c=c_3$. 对于任意的映射 χ, 记事件 $A_{i,\chi}$ 为

$$A_{i,\chi}=\left\{\left|\sum_{j\in S_i}\chi(j)\right|<c_3\sqrt{n}\right\}.$$

由于 $S_1,S_2,\cdots,S_m$ 独立同分布, 那么 $A_{1,\chi},A_{2,\chi},\cdots,A_{m,\chi}$ 相互独立. 故有

$$\mathbb{P}\left(\bigcap_{i=1}^m A_{i,\chi}\right)=\prod_{i=1}^m\mathbb{P}(A_{i,\chi})<\left(\frac{1}{2}\right)^{Cn}.$$

对于任意的 χ, 注意到有 $\#\{\chi|\chi:\{1,2,\cdots,n\}\to\{\pm1\}\}=2^n$, 那么有

$$\mathbb{P}\left(\bigcup_{\chi}\left\{\bigcap_{i=1}^m A_{i,\chi}\right\}\right)<2^n\left(\frac{1}{2}\right)^{Cn}=\left(\frac{1}{2}\right)^{Cn-n}.$$

取 $C=1$, 那么有

$$\mathbb{P}\left(\bigcup_{\chi}\left\{\bigcap_{i=1}^m\left|\sum_{j\in S_i}\chi(j)\right|\leq c_3\sqrt{n}\right\}\right)<2^n\cdot\left(\frac{1}{2}\right)^n=1.$$

即考虑取 $S_1,S_2,\cdots,S_m$ 和常数 $c=c_3$, 存在一个 χ, 使得对于任意的 S_i, 有(7)成立的概率小于 1. 那么我们有

$$\mathbb{P}\left(\bigcap_{\chi}\left\{\bigcup_{i=1}^m\left|\sum_{j\in S_i}\chi(j)\right|\geq c_3\sqrt{n}\right\}\right)>0.$$

即取集合 $S_1, S_2, \cdots, S_m$ 和常数 $c = c_3$, 对于任意的 χ, 存在 $1 \le i \le m$, 使得

$$\left|\sum_{j \in S_i} \chi(j)\right| \ge c\sqrt{n}.$$

成立. 且 $m = n = O(n)$, 证毕.

注: 这样我们就顺便证明了当 $m = n$ 时, (3) 中的结论同样成立.

$\triangleleft$