Homework #2

Due: 2025-4-3 23:59 | 7 Questions, 120 Pts

Name: 方嘉聪 ID: 2200017849

Note: The total points of this homework is $15 + 10 + 10 + 10 + 15 + 10 + 50 = 120$, with 50 points being the *Challenge Problem.*

Question 1 (15') (Primality Test with Square Root Oracle). Suppose you are given a black-box algorithm (known as "oracle") $\mathcal{S}(b, n)$ for computing square roots of b modulo n. In other words, the algorithm may return *one* a such that $a^2 \equiv b \pmod n$ in each invocation, or output $\perp$ if there is no root. Using this algorithm as a black box, design an RP algorithm (i.e., algorithm with one-side error) for compositeness, and analyze its error bound.

[Hint: You do not know the behavior of algorithm $\mathcal{S}$. For example, the solution to $a^2 \equiv 1 \pmod{12}$ is $a \equiv 1, 5, 7, 11 \pmod{12}$, but $\mathcal{S}(1, 12)$ may return 1 all the time. You can never expect $\mathcal{S}$ to return a root randomly; its output can be adversarial. Therefore, in this question you must use some randomness in your algorithm.] ◀

Answer. Consider the following algorithm (suppose n is odd here):

```
Algorithm 1 Primality Test with Square Root Oracle
Input: A number $n$ and the oracle $\mathcal{S}(b, n)$.
Output: Whether $n$ is prime or not.
Randomly sample $b \in \mathbb{Z}_n$. If $\gcd(b, n) \neq 1$, then return No immediately.
Loop to check if $n$ is a perfect power of prime $p \in [2, \log n]$.    ▷ Time complexity: $O(\log^2 n)$
$a \leftarrow \mathcal{S}(b^2, n)$.
if $a = \pm b$ then
    return Yes
else
    return No
```

If n is prime, then $a^2 \equiv b^2 \pmod n \implies (a-b)(a+b) \equiv 0 \pmod n \implies n|(a-b) \vee n|(a+b)$. Therefore, $a \equiv \pm b \pmod n$, and the algorithm will return **Yes** with probability 1. **Algorithm 1** is a RP algorithm.

If n is composite, when $n = p^k$ for some k and prime p, then the algorithm will return **No**. Now consider $n = pq$ where p, q are distinct odd prime. And analyse the solutions of Congruence Equation

$$x^2 \equiv b^2 \pmod{pq} \tag{1}$$

By the Chinese Remainder Theorem, we can decompose the equation into two equations:

$$x^2 \equiv b^2 \pmod p$$
$$x^2 \equiv b^2 \pmod q$$

whose solutions are:

$$\begin{aligned} x &\equiv b \pmod p, \quad x \equiv -b \pmod p \\ x &\equiv b \pmod q, \quad x \equiv -b \pmod q \end{aligned}$$

Furthermore, we can get 4 solutions of (1):

$$x \equiv \pm b \pmod{pq}, \quad x \equiv \pm(b(qq^{-1} - pp^{-1})) \pmod{pq}$$

where q^{-1} and p^{-1} are the inverses of q and p modulo p and q, respectively. We can verify that the four solutions are distinct.

For more general cases, i.e., $n = \prod_i p_i^{k_i}$ where p_i are distinct primes and $i \geq 2$. Using similar argument, the congruence equation $x^2 \equiv b^2 \pmod n$ has at least 4 distinct solutions. Therefore,

$$\Pr[\mathcal{S}(b,n) \not\equiv \pm b | n \text{ is composite}] \geq \frac{1}{2}. \implies \Pr[\text{Error}] \leq \frac{1}{2}$$

By repeating the algorithm(Line 3-7) k times, we can reduce the error probability to $1/2^k$. **Q.E.D.**$\triangleleft$

Question 2 (10') (Ramsey Number).

Define the non-diagonal Ramsey number $R(k,t)$ as the minimum number n such that, in every 2-coloring of K_n, there exists either a red k-clique, or a blue t-clique.

Prove that, if there is a real $p, 0 \le p \le 1$ such that

$$\binom{n}{k}p^{\binom{k}{2}} + \binom{n}{t}(1-p)^{\binom{t}{2}} < 1,$$

then $R(k,t) > n$. Using this, show that

$$R(4,t) = \Omega\left(\frac{t^{3/2}}{(\ln t)^{3/2}}\right).$$

[Hint: $\binom{n}{t} \le \frac{n^t}{t!}, t! \sim \sqrt{2\pi t}(\frac{t}{e})^t, (1-p)^x \le e^{-px}$. Note that the last inequality here will be frequently used in our lecture.] ◀

Answer. We will prove the first part of the question.

Randomly color each edge of K_n red with probability p and blue with probability $1-p$, independently. Let C be any k-clique in K_n, then $\Pr[C \text{ is red } k\text{-clique}] = p^{\binom{k}{2}}$. Let C' be any t-clique in K_n, then $\Pr[C' \text{ is blue } k\text{-clique}] = (1-p)^{\binom{t}{2}}$. Denote

$$M := \{\text{every 2-coloring, } K_n \text{ has a red } k\text{-clique or blue } t\text{-clique}\}$$

$$M_1 := \{\text{every 2-coloring, } K_n \text{ has a red } k\text{-clique}\},\ M_2 := \{\text{every 2-coloring,, } K_n \text{ has a blue } t\text{-clique}\}$$

then by union bound, we have

$$\begin{aligned}\Pr[M] &\le \Pr[M_1] + \Pr[M_2]\\ &\le \binom{n}{k}\cdot \Pr[\text{a given } k\text{-clique is red}] + \binom{n}{t}\cdot\Pr[\text{a given } t\text{-clique is blue}]\\ &\le \binom{n}{k}p^{\binom{k}{2}} + \binom{n}{t}(1-p)^{\binom{t}{2}} < 1\end{aligned}$$

Therefore, there exists a 2-coloring of K_n such that there is no red k-clique and no blue t-clique which implies $R(k,t) > n$.

Then we show that $R(4,t) = \Omega\left(\frac{t^{3/2}}{(\ln t)^{3/2}}\right)$ by proving that

$$p = \frac{\ln t}{t} \in (0,1), n = \left(\frac{t}{\ln t}\right)^{3/2} \implies \binom{n}{4}p^6 + \binom{n}{t}(1-p)^{\binom{t}{2}} < 1.$$

Using the hint(and another form of Stirling's formula), we have

$$\begin{aligned}\binom{n}{4}p^6 + \binom{n}{t}(1-p)^{\binom{t}{2}} &\le \frac{n^4}{4!}p^6 + \frac{n^t}{t!}e^{-\frac{pt(t-1)}{2}} = \frac{n^4p^6}{24} + \frac{n^t}{e^{\theta_t/12t}\sqrt{2\pi t}\left(\frac{t}{e}\right)^t}\cdot e^{-\frac{pt(t-1)}{2}}\\ &\le \frac{1}{24} + \frac{n^te^t}{\sqrt{2\pi t}\cdot t^t}\cdot t^{-\frac{1}{2}(t-1)} = \frac{1}{24} + \frac{n^te^t}{\sqrt{2\pi}\cdot t^{3t/2}}\\ &= \frac{1}{24} + \frac{e^t}{\sqrt{2\pi}\cdot(\ln t)^{3t/2}} =: f(t)\end{aligned}$$

We will show that $f(t) < 1$ for sufficiently large t.

$$f'(t) = \frac{e^t}{\sqrt{2\pi}\,(\ln t)^{3t/2}} \left(1 - \frac{3}{2}\ln(\ln t) - \frac{3}{2\ln t}\right) < 0 \text{ for } t \in \mathbb{Z}_{\geq 2}$$

Thus, for $t \geq 6, f(t) \leq f(6) < 1$. Therefore, for sufficiently large t, we have

$$R(4,t) > n = \left(\frac{t}{\ln t}\right)^{3/2} = \Omega\left(\frac{t^{3/2}}{(\ln t)^{3/2}}\right).$$

Another Form of Stirling's Formula: For $n \in \mathbb{Z}^*$, we have

$$n! = e^{\frac{\theta_n}{12n}}\sqrt{2\pi n}\left(\frac{n}{e}\right)^n, \text{ where } \theta_n \in (0,1)$$

Q.E.D. $\triangleleft$

Question 3 (10') (Conditional Expectation).

Let $v_1, \cdots, v_n \in \mathbb{R}^n$, all with $\|v_i\|_2 = 1$.

a. (5') Prove that there exist $\varepsilon_1, \ldots, \varepsilon_n = \pm 1$ such that

$$\|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|_2 \le \sqrt{n}.$$

b. (5') Please provide a polynomial-time *deterministic* algorithm for finding an assignment of $\varepsilon_1, \ldots, \varepsilon_n$ in ± 1 such that $\|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|_2 \le \sqrt{n}$, and analyze its time complexity.

[Hint: To derandomize the algorithm, consider how you can leverage the result from a. *in an inductive manner to fix the values of $\varepsilon_1, \cdots, \varepsilon_n$ one by one.]*

◀

Answer. a). Randomly assign $\varepsilon_i = \pm 1$ for $i = 1, \ldots, n$ with equal probability independently. Define random variable $X = \|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|_2^2$. Then we have

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}\left[\varepsilon_i^2 \cdot \|v_i\|^2\right] + \sum_{i \neq j} \mathbb{E}[\varepsilon_i \varepsilon_j \cdot \langle v_i, v_j \rangle] = \sum_{i=1}^{n} \mathbb{E}\left[\varepsilon_i^2 \|v_i\|^2\right] = n$$

Note that for any $i \neq j$, $\mathbb{E}[\varepsilon_i \varepsilon_j] = 0$. Therefore, there exist $\varepsilon_1, \cdots, \varepsilon_n$ such that

$$\|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|_2 \le \sqrt{\mathbb{E}[X]} = \sqrt{n}.$$

b). Using Conditional Expectation, determinte $\varepsilon_1, \cdots, \varepsilon_n$ one by one. The detail algorithm is as follows:

Algorithm 2 Deterministic Algorithm for Finding $\varepsilon_1, \cdots, \varepsilon_n$

```
Input: $v_1, \cdots, v_n \in \mathbb{R}^n$ with $\|v_i\|_2 = 1$.
Output: $\varepsilon_1, \cdots, \varepsilon_n$ with deterministic values in $\pm 1$
$f(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n) := \mathbb{E}_{\varepsilon_i \leftarrow_R \{-1,1\}}\left[\|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|_2^2\right]$
for $i = 1$ to $n$ do
    $f_1 \leftarrow f\left(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \middle| \varepsilon_i = 1 \wedge \sum_{k=1}^{i-1} \varepsilon_k$ fixed already$\right)$
    $f_2 \leftarrow f\left(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \middle| \varepsilon_i = -1 \wedge \sum_{k=1}^{i-1} \varepsilon_k$ fixed already$\right)$
    if $f_1 \le f_2$ then
        $\varepsilon_i \leftarrow 1$
    else
        $\varepsilon_i \leftarrow -1$
return $\varepsilon_1, \cdots, \varepsilon_n$
```

Time Complexity Analysis. Each iteration of the for loop takes $O(n)$ time to compute f_1 and f_2.[1] Therefore, the total time complexity is $O(n^3)$.

Correctness. We can prove the correctness of the algorithm by induction on n.

- For $n = 1$, the algorithm is correct trivially.
- Suppose the algorithm is correct for $n = k - 1$.

[1] Naive method costs $O(n^2)$, and it can achieve $O(n)$ by computing f_1 and f_2 using the stored previous values.

- For $n = k$, $f_1 := f(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n | \varepsilon_1 = 1), f_2 := f(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n | \varepsilon_1 = -1)$. then

$$\text{for } j \in \{0, 1\}, f_j = \mathbb{E}\left[\left\|(-1)^j \cdot v_1 + \sum_{i=2}^{n} \varepsilon_i v_i\right\|_2^2\right] \leq 1 + \mathbb{E}\left[\left\|\sum_{i=2}^{n} \varepsilon_i^2\right\|_2^2\right] \leq n$$

The last inequality holds due to the induction hypothesis. Therefore, we have

$$\begin{aligned} f(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n) &= \Pr[\varepsilon_1 = 1] \cdot f_1 + \Pr[\varepsilon_1 = -1] \cdot f_2 \\ &\leq \min\{f_1, f_2\} \leq n. \end{aligned}$$

Inductively, we can prove that the algorithm is correct for all $n \in \mathbb{N}^+$.

Q.E.D. $\triangleleft$

Question 4 (10') (d-wise Independent Random Variables). Show how to construct 2^m m-bits random variables which are d-wise independent, using only dm random bits.

[Hint: Consider the finite field $\mathbb{F}_{2^m} \cong \mathbb{F}_2^m$.] ◀

Answer. Consider the finite field $\mathbb{F}_{2^m} \cong \mathbb{F}_2^m$, which means each element in $\mathbb{F}_{2^m}$ can be represented as a m-bits 0-1 vector. Then consider the polynomial

$$P(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{d-1} x^{d-1} \in \mathbb{F}_{2^m}[x]$$

where a_i is chosen uniformly at random from $\mathbb{F}_{2^m}$. Thus, the total random bits we need is dm.

For any element $x_i (i \in \{0, 1, \cdots, 2^m - 1\})$ in $\mathbb{F}_{2^m}$, let random variable $X_i = P(x_i)$. Then we will prove that $X_0, X_1, \cdots, X_{2^m-1}$ are d-wise independent.

For any d elements $x_{i_1}, x_{i_2}, \cdots, x_{i_d}$ in $\mathbb{F}_{2^m}$, in analyzing the probability

$$\Pr[X_{i_1} = t_1, X_{i_2} = t_2, \cdots, X_{i_d} = t_d]$$

we have the following equations:

$$\forall k \in \{1, 2, \cdots, d\}, \quad \sum_{j=1}^{d} a_j x_{i_k}^j = t_k$$

which has a unique solution in $\mathbb{F}_{2^m}$, because the degree of $P(x)$ is $d-1$ and we can rewrite the above equations as

$$V = \begin{bmatrix} 1 & x_{i_1} & x_{i_1}^2 & \cdots & x_{i_1}^{d-1} \\ 1 & x_{i_2} & x_{i_2}^2 & \cdots & x_{i_2}^{d-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{i_d} & x_{i_d}^2 & \cdots & x_{i_d}^{d-1} \end{bmatrix}, \quad T = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_d \end{bmatrix}, \quad \text{and } A = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_{d-1} \end{bmatrix}$$

Notice that V is a Vandermonde matrix with $x_i \neq x_j$ for $i \neq j$, which means V is invertible. Therefore, $VA = T$ has a unique solution $A = V^{-1}T$. Thus, since a_i is chosen uniformly at random from $\mathbb{F}_{2^m}$, we have

$$\Pr[X_{i_1} = t_1, X_{i_2} = t_2, \cdots, X_{i_d} = t_d] = \frac{1}{|\mathbb{F}_{2^m}|^d}$$

which means $X_0, X_1, \cdots, X_{2^m-1}$ are d-wise independent. **Q.E.D.** ◁

PS: I input the problem to test DeepSeek, the sketch answer is almost correct..

Question 5 (15') (Coupon Collector). Farmer John is deploying sprinklers in a square field with side length 1. Each sprinkler has a cover radius of r ($r < 1$), meaning that all crops within distance r of the sprinkler can be watered. The sprinklers are dropped independently and uniformly in the square field at random. Farmer John wants to know how many sprinklers are needed to ensure that the whole area of the field is covered with probability at least $(1 - \varepsilon)$. Let m be the minimum number of sprinklers required. Prove that,

$$m = \mathcal{O}\left(\frac{1}{r^2} \log \frac{1}{\varepsilon r^2}\right).$$

[Hint: To show the required bound you only need a weak coupon collector bound; the result can be made stronger. In your solution, prove the coupon collector bound you use.] ◀

Answer. We firstly divide the farm into a square grid with side length d. In order to cover single small grid when the sprinkler is located in any position within the grid, let $d = r/\sqrt{2}$ (the worest case is when the sprinkler is located in the corner of the grid). Therefore, when there is a sprinkler in each grid, the whole farm is covered.

Denote $M = \{\text{there is a sprinkler in each grid}\}$, and $N_0 = \{\text{a grid is not covered by any sprinkler}\}$

$$\Pr[M] = 1 - \sum_{\text{any grid}} \left(1 - \frac{r^2}{2}\right)^m = 1 - \frac{2}{r^2}\left(1 - \frac{r^2}{2}\right)^m$$

We hope that $\Pr[M] \geq 1 - \varepsilon$, i.e.,

$$\frac{2}{r^2}\left(1 - \frac{r^2}{2}\right)^m \leq \varepsilon \implies m \geq \frac{\log(\varepsilon r^2/2)}{\log(1 - r^2/2)} = \log\left(\frac{2}{\varepsilon r^2}\right) \cdot \frac{1}{\log(2/(2 - r^2))} \tag{2}$$

Let $k = r^2 \in (0, 1)$, now we try to prove that

$$\log\left(\frac{2}{2 - k}\right) \leq k \iff \frac{2}{2 - k} \leq e^k \iff 0 \leq (2 - k)e^k - 2 := f(k)$$

Notice that

$$f'(k) = (1 - k)e^k \geq 0 \implies f(k) > f(0) = 0.$$

Therefore,

$$\log\left(\frac{2}{2 - k}\right) \leq k \iff \frac{1}{\log(2/(2 - r^2))} \geq \frac{1}{r^2} \tag{3}$$

Combining (2) and (3), we have

$$m = \mathcal{O}\left(\frac{1}{r^2} \log \frac{1}{\varepsilon r^2}\right).$$

Q.E.D. ◁

Question 6 (10') (Median-of-Mean Trick). Let $X_i, i = 1, 2, \cdots, 2s+1, s > 0$ be i.i.d random variables with $E[X_i] = \mu$. Assume each X_i is concentrated well, i.e., $\Pr[|X_i - \mu| \geq \varepsilon\mu] \leq \frac{1}{4}$. Prove

$$\Pr[|\mathrm{MED}(X_1, \cdots, X_{2s+1}) - \mu| \geq \varepsilon\mu] \leq \exp\left(-\frac{s}{4}\right),$$

where $\mathrm{MED}(X_1, \cdots, X_{2s+1})$ is the median value of $X_1, \cdots, X_{2s+1}$.

[Hint: If the medium is biased, then more than half are biased.] ◀

Answer. Define T_i be the indicator random variable for X_i being biased, i.e.,

$$T_i = \mathbf{1}\{|X_i - \mu| \geq \varepsilon\mu\} = \begin{cases} 1, & \text{if } |X_i - \mu| \geq \varepsilon\mu \\ 0, & \text{otherwise} \end{cases}$$

Then $\Pr[T_i = 1] \leq 1/4$, and $\mathbb{E}[T_i] = \Pr[T_i = 1] \leq 1/4$. Motivated by the hint, we have

$$\Pr\left[|\mathrm{MED}(X_1, \cdots, X_{2s+1}) - \mu| \geq \varepsilon\mu\right] = \Pr\left[\sum_{i=1}^{2s+1} T_i \geq s+1\right]$$

Denote

$$T = \frac{1}{2s+1}\sum_{i=1}^{2s+1} T_i \implies \mathbb{E}[T] \leq \frac{1}{4}.$$

Then we can apply Chernoff bound to get

$$\begin{aligned} \Pr\left[\sum_{i=1}^{2s+1} T_i \geq s+1\right] &= \Pr\left[T - \frac{1}{4} \geq \frac{s+1}{2s+1} - \frac{1}{4}\right] \leq \Pr\left[T - \mathbb{E}[T] \geq \frac{2s+3}{4(2s+1)}\right] \\ \text{(Chernoff Bound)} \quad &\leq \exp\left(-2\cdot(2s+1)\cdot\frac{(2s+3)^2}{16(2s+1)^2}\right) = \exp\left(-\frac{(2s+3)^2}{8(2s+1)}\right) \\ &= \exp\left(-\frac{s}{4} - \frac{2s+5}{2(2s+1)}\right) \\ &\leq \exp\left(-\frac{s}{4}\right). \end{aligned}$$

Therefore, we have

$$\Pr\left[|\mathrm{MED}(X_1, \cdots, X_{2s+1}) - \mu| \geq \varepsilon\mu\right] \leq \exp\left(-\frac{s}{4}\right).$$

Q.E.D. ◁

Question 7 (50') (A Combinatorial Proof of Chernoff Bound). In the lecture, you learned how to prove Chernoff Bound by Generating Function. In this problem, you will learn another way to prove Chernoff Bound. Our goal is to prove the following theorem:

Theorem 1. *Suppose $X_1, \dots, X_n$ are i.i.d random variables from $\{-1, 1\}$, each w.p. $\frac{1}{2}$, and k be a positive integer with $k \le \sqrt{n}$. Then*

$$\Pr\left[\sum_{i=1}^{n} X_i \ge k\sqrt{n}\right] \le e^{-\Theta(k^2)}$$

You will achieve this goal step by step.

a. (10') Prove the following lemma:

Lemma 2. *(Poor Man Chernoff Bound) Suppose $X_1, \dots, X_n$ are i.i.d random variables from $\{-1, 1\}$ each w.p. $\frac{1}{2}$, and k be a positive integer. Then*

$$\Pr\left[\sum_{i=1}^{n} X_i \ge 2k\sqrt{n}\right] \le 2^{-k}$$

Hint: First, you can use Chebeyshev's Inequality to prove the following fact:

Fact 3. *Suppose $X_1, \dots, X_n$ are i.i.d random variables from $\{-1, 1\}$ each w.p. $\frac{1}{2}$. Then*

$$\Pr\left[\sum_{i=1}^{n} X_i \ge 2\sqrt{n}\right] \le \frac{1}{4}$$

Then, consider $S_i = \sum_{j=1}^{i} X_j$, let p be the first point where $S_p \ge 2\sqrt{n}$, then $\Pr[S_n \ge 2k\sqrt{n}] = \Pr[p \text{ exists}] \cdot \Pr[S_n - S_p \ge 2(k-1)\sqrt{n} \mid p \text{ exists}]$. If you can prove $\Pr[p \text{ exists}] \le \frac{1}{2}$, you can prove the lemma by induction.

b. (10') Prove the following lemma:

Lemma 4. *(Chernoff Bound for Geometric Distribution)*
Suppose $X_1, \dots, X_n$ are i.i.d random variables, that $X_i \ge 0, \Pr[X_i \ge j] \le p^j, \forall j = 1, 2, \dots$ for a $p < \frac{1}{4}$. Then

$$\Pr\left[\sum_{i=1}^{n} X_i \ge 2n\right] \le (4p)^n$$

.

Hint: If we can prove $\Pr[\sum_{i=1}^{n} \lfloor X_i \rfloor \ge n] \le (4p)^n$, then it's easy to see the lemma will hold. Suppose $\sum_{i=1}^{n} \lfloor X_i \rfloor \ge n$, then there exist $Y_1, \dots, Y_n$, that $\forall 1 \le i \le n, X_i \ge Y_i$ and $\sum_{i=1}^{n} Y_i = n$. Fix the sequence $Y_1, \dots, Y_n$, calculate the probability of sequence X_i satisfies $\forall i, X_i \ge Y_i$. Then use union bound for all the posible sequence of Y.

c. (10') Prove the following lemma:

Lemma 5. *(Lowerbound for Chernoff Bound)*

Suppose $X_1, \ldots, X_n$ are i.i.d random variables from $\{-1, 1\}$ each w.p. $\frac{1}{2}$, and k be a positive integer and $k \leq \sqrt{n}$, Then

$$\Pr\left[\sum_{i=1}^{n} X_i \geq \frac{k}{2}\sqrt{n}\right] \geq \left(\frac{1}{4}\right)^{k^2}$$

.

You can use the fact:

Fact 6. *Suppose $X_1, \ldots, X_n$ are i.i.d random variables from $\{-1, 1\}$ each w.p. $\frac{1}{2}$. Then*

$$\Pr\left[\sum_{i=1}^{n} X_i \geq \frac{1}{2}\sqrt{n}\right] \geq \frac{1}{4}$$

.

Hint: Divide $X_1, \ldots, X_n$ into $m = k^2$ groups, use the Fact 6 on each group.

d. (20') Prove Theorem 1. ◀

Answer. a). We prove **Fact 3** first. Notice that

$$\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] = n, \quad \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = 0$$

Then by Chebeyshev's Inequality, we have

$$\Pr\left[\sum_{i=1}^{n} X_i \geq 2\sqrt{n}\right] \leq \Pr\left[\left|\sum_{i=1}^{n} X_i - 0\right| \geq 2\sqrt{n}\right] \leq \frac{\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right]}{4n} = \frac{1}{4}.$$

Now we prove that $\Pr[p \text{ exists}] \leq \frac{1}{2}$. Notice that, for any $i \in [n]$, by symmetry, we have

$$\Pr\left[S_n - S_i > 0\right] = \Pr\left[S_n - S_i < 0\right] \implies \Pr\left[S_n - S_i \geq 0\right] \geq \frac{1}{2}.$$

Therefore,

$$\begin{aligned}\Pr[p \text{ exists}] &= \Pr\left[\bigcup_{i=1}^{n}\left\{S_i \geq 2\sqrt{n}\right\}\right] \leq \sum_{i=1}^{n}\Pr\left[S_i \geq 2\sqrt{n}\right] \\ &\leq 2\sum_{i=1}^{n}\Pr\left[S_i \geq 2\sqrt{n}\right]\cdot\Pr[S_n - S_i \geq 0]\end{aligned}$$

And by union bound, we have

$$\begin{aligned}\frac{1}{4} \geq \Pr\left[S_n \geq 2\sqrt{n}\right] &= \Pr\left[\bigcup_{i=1}^{n}\left\{S_i \geq 2\sqrt{n}\right\} \wedge \left\{S_n - S_i \geq 0\right\}\right] \\ &\geq \sum_{i=1}^{n}\Pr\left[\left\{S_i \geq 2\sqrt{n}\right\} \wedge \left\{S_n - S_i \geq 0\right\}\right] \\ &= \sum_{i=1}^{n}\Pr\left[S_i \geq 2\sqrt{n}\right]\cdot\Pr[S_n - S_i \geq 0]\end{aligned}$$

Combining the two inequalities, we have

$$\frac{1}{4} \geq \sum_{i=1}^{n}\Pr\left[S_i \geq 2\sqrt{n}\right]\cdot\Pr[S_n - S_i \geq 0] \geq \frac{1}{2}\Pr[p \text{ exists}] \implies \Pr[p \text{ exists}] \leq \frac{1}{2}.$$

Then we prove the leamma by induction, for any $m \in [1, n]$:

- **Base case:** For $k = 1$, we have
$$\Pr\left[S_m \geq 2\sqrt{n}\right] \leq \frac{\mathrm{Var}[S_m]}{4n} \leq \frac{1}{4}.$$
- **Induction Hypothesis:** Assume for any $k \leq t (t \in \mathbb{Z}^*)$, we have
$$\Pr\left[S_m \geq 2k\sqrt{n}\right] \leq 2^{-k} \text{ for } k = 1, 2, \ldots, t.$$
- **Induction Step:** For $k = t + 1$, we have
$$\Pr\left[S_m \geq 2(t+1)\sqrt{n}\right] = \Pr\left[p \text{ exists}\right] \cdot \Pr\left[S_n - S_p \geq 2t\sqrt{n} \mid p \text{ exists}\right] \leq \frac{1}{2} \cdot 2^{-t} = 2^{-(t+1)}.$$
Here, $S_n - S_p$ are still sum of $n-p$ i.i.d random variables from $\{-1, 1\}$, and by the induction hypothesis, we have $\Pr\left[S_n - S_p \geq 2t\sqrt{n}\right] \leq 2^{-t}$.

Therefore, by induction, we have proved a stronger version of the lemma, i.e., for any $m \in [1, n]$ and $k \in \mathbb{Z}^*$, we have
$$\Pr\left[S_m \geq 2k\sqrt{n}\right] \leq 2^{-k} \implies \Pr\left[S_n \geq 2k\sqrt{n}\right] \leq 2^{-k}.$$

b). Notice that
$$A := \left\{\sum_{i=1}^{n} X_i \geq 2n\right\} \subseteq \left\{\sum_{i=1}^{n} \lfloor X_i \rfloor \geq n\right\} := B \implies \Pr[A] \leq \Pr[B].$$

Fixed the sequence $Y_1, \ldots, Y_n$ with $\sum_{i=1}^{n} Y_i = n$, we have
$$\Pr\left[\forall i, X_i \geq Y_i\right] = \prod_{i=1}^{n} \Pr\left[X_i \geq Y_i\right] = p^{\sum_{i=1}^{n} Y_i} = p^n.$$

Then by union bound, we have
$$\begin{aligned}\Pr\left[\sum_{i=1}^{m} \lfloor X_i \rfloor \geq n\right] &\leq \sum_{Y_i \in \mathbb{Z}_{\geq 0}, \sum Y_i = n} \Pr\left[\forall i, X_i \geq Y_i\right] \\ &= \binom{2n-1}{n-1} \cdot p^n \leq (4p)^n.\end{aligned}$$

Therefore,
$$\Pr\left[\sum_{i=1}^{n} X_i \geq 2n\right] \leq (4p)^n.$$

c). We first consider a special case when $k^2 | n$. Divide $X_1, \ldots, X_n$ into $m = k^2$ groups, each group contains $n/m = n/k^2$ random variables. And then apply **Fact 6** to each group, we have
$$\forall i \in \{1, \cdots, k^2\}, \quad \Pr\left[\sum_{j \in G_i} X_j \geq \frac{1}{2k}\sqrt{n}\right] \geq \frac{1}{4}$$

Notice that
$$\bigcap_{i=1}^{k^2} \left\{\sum_{j \in G_i} X_j \geq \frac{1}{2k}\sqrt{n}\right\} \subseteq \left\{\sum_{i=1}^{n} X_i \geq \frac{k}{2}\sqrt{n}\right\}$$

Therefore,

$$\Pr\left[\sum_{i=1}^{n} X_i \geq \frac{k}{2}\sqrt{n}\right] \geq \Pr\left[\bigcap_{i=1}^{k^2}\left\{\sum_{j\in G_i} X_j \geq \frac{1}{2k}\sqrt{n}\right\}\right] = \left(\frac{1}{4}\right)^{k^2} \tag{4}$$

For the general case, suppose $n = qk^2 + r$ where $0 < r < q$. Then we divide $X_1, \ldots, X_n$ into $m = k^2$ groups, where r groups contain $q+1$ random variables and $m - r$ groups contain q random variables. Similar to the above case, we hope that for $q = \frac{n-r}{k^2}$, we have

$$r \cdot \frac{1}{2}\sqrt{q+1} + (k^2 - r) \cdot \frac{1}{2}\sqrt{q} \approx \frac{k}{2}\sqrt{n} \iff r\sqrt{q+1} + (k^2 - r)\sqrt{q} \approx k\sqrt{n}.$$

If so we have equation (4) holds for $n = qk^2 + r$ (may introduce some error).

d). Follow the same idea as in part (c), we first divide $X_1, \ldots, X_n$ into $m = k^2$ groups with the number of elements in each group is $\{n_i\}_{i=1}^{k^2}$. Then apply **Lemma 2** to each group, we have

$$\forall i \in \{1, \cdots, k^2\}, \quad \Pr\left[\sum_{j\in G_i} X_j \geq 2k'\sqrt{n_i}\right] \leq 2^{-k'}$$

Now we try to construct a geometric random variable Y_i for each group G_i. Let $k' = ck$ here, where c is a constant. Then we have

$$\Pr\left[\sum_{j\in G_i} X_j \geq 2ck\sqrt{n_i}\right] \leq \left(\frac{1}{2^c}\right)^k$$

Select $c \geq 3$ then consider random variable Y_i such that

$$Y_i = \frac{\sum_{j\in G_i} X_j}{2c\sqrt{n_i}} \implies Y_i \text{ satisfies the requirement of } \textbf{Lemma 4}.$$

Therefore, we have

$$\Pr\left[\sum_{i=1}^{k^2} Y_i \geq 2k^2\right] \leq \left(\frac{1}{2^{c-2}}\right)^{k^2}$$

Note that ($c \geq 3$, and let n_i as close as possible which means $n_i = \lfloor n/k^2 \rfloor$ or $\lceil n/k^2 \rceil$)

$$\sum_{i=1}^{k^2} Y_i = \sum_{i=1}^{k^2} \frac{\sum_{j\in G_i} X_j}{2c\sqrt{n_i}} \approx \frac{\sum_{i=1}^{n} X_i}{2c\sqrt{n}} \tag{5}$$

Therefore, we have (select constnt $c \geq 3$)

$$\Pr\left[\sum_{i=1}^{n} X_i \geq 4ck\sqrt{n}\right] \leq \left(\frac{1}{2^{c-2}}\right)^{k^2} \leq e^{-\Theta(k^2)}$$

Q.E.D.

Note: The approximation in formula (5) may introduce some error, but I did not find a better way to prove it :<(. And I did not know if Lemma 5 is necessary for part (d). ◁