COMP7102P Fall 2021 | Homework 1

BA21011024 Qianyu Zhang 张钱宇 | due: Oct 7, 09:30

Problem 1 (30 points)**.** Prove the following claims and show your calculations.

(a) Prove that $\mathbb{E}[Y] \leq \mathbb{E}[Y^2]^{\frac{1}{2}}$ for any real random variable Y. Moreover, show this implies $\mathbb{E}[|X - \mathbb{E}[X]|] \leq \sqrt{\mathsf{Var}(X)}$.

Answer:

$$\mathsf{Var}(Y) = \mathbb{E}[(Y - E(Y))^2] = \mathbb{E}[Y^2 - 2Y\,\mathbb{E}(Y) + \mathbb{E}(Y)^2] = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2,$$

where $\mathsf{Var}(Y) \geq 0$, *i.e.*, $\mathbb{E}[Y^2] - \mathbb{E}[Y]^2 \geq 0$.
Thus, we can prove that $\mathbb{E}[Y] \leq \mathbb{E}[Y^2]^{\frac{1}{2}}$.

Then, according to $\mathsf{Var}(X) = \mathbb{E}[(X - E(X))^2]$,
we can obtain $\sqrt{\mathsf{Var}(X)} = \sqrt{\mathbb{E}[(X - E(X))^2]} = \mathbb{E}[(X - E(X))^2]^{\frac{1}{2}}$.
Let's assume that $Y = X - E(X)$, and then $\mathbb{E}[X - E(X)] \leq \mathbb{E}[(X - E(X))^2]^{\frac{1}{2}} = \sqrt{\mathsf{Var}(X)}$.
Thus, the above conclusion implies $\mathbb{E}[X - E(X)] \leq \sqrt{\mathsf{Var}(X)}$.

(b) For n independent variables $X_1, \cdots, X_n$, prove that $\mathsf{Var}(X_1 + X_2 + \cdots + X_n) = \mathsf{Var}(X_1) + \mathsf{Var}(X_2) + \cdots + \mathsf{Var}(X_n)$.

Answer:

$$\begin{aligned}
\mathsf{Var}(X_1 + X_2 + ... + X_n) &= \mathbb{E}[[(X_1 + X_2 + \cdots + X_n) - \mathbb{E}(X_1 + X_2 + \cdots + X_n)]^2] \\
&= \mathbb{E}[[(X_1 - \mathbb{E}(X_1)) + (X_2 - \mathbb{E}(X_2)) + ... + (X_n - \mathbb{E}(X_n))]^2] \\
&= \mathbb{E}[\sum_{i=1}^{n} (X_i - \mathbb{E}(X_i))^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} (X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))] \\
&= \mathbb{E}[\sum_{i=1}^{n} (X_i - \mathbb{E}(X_i))^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} (X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))] \\
&= \sum_{i=1}^{n} \mathbb{E}[(X_i - \mathbb{E}(X_i))^2] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))] \\
&= \sum_{i=1}^{n} \mathsf{Var}(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \mathsf{Cov}(X_i, X_j)
\end{aligned}$$

As we know, when $X_1, ..., X_n$ are independent variables, $\mathsf{Cov}(X_i, X_j) = 0$.
Thus, we can prove that $\mathsf{Var}(X_1 + X_2 + \cdots + X_n) = \mathsf{Var}(X_1) + \mathsf{Var}(X_2) + \cdots + \mathsf{Var}(X_n)$.

(c) Consider d independent random coins $Z_1, ..., Z_d \in \{\pm 1\}$ where each Z_i is 1 or -1 with probability 1/2 separately. We define $n = 2^d - 1$ random variables as follows: For each non-empty subset $S \subseteq [n]$, we define $X_s = \prod_{i \in S} Z_i$.
For example when $d = 3$, there are 7 random variables $X_1 = Z_1, X_2 = Z_2, X_3 = Z_3, X_{1,2} = Z_1 \cdot Z_2, X_{1,3} = Z_1 \cdot Z_3, X_{2,3} = Z_2 \cdot Z_3$ and $X_{1,2,3} = Z_1 \cdot Z_2 \cdot Z_3$.
Calculate $\mathsf{Var}(\sum_S X_S)$ and compare this with part (b).

Answer:

For $\forall i \in \{1, 2, \cdots, d\}, \mathbb{E}[Z_i] = 1 \times \frac{1}{2} + (-1) \times \frac{1}{2} = 0, \mathbb{E}[Z_i^2] = 1^2 \times \frac{1}{2} + (-1)^2 \times \frac{1}{2} = 1$,
For $\forall i \in S, S \subseteq [n]$, each coin Z_i is independent, then we can calculate $\mathbb{E}[X_S]$ and $\mathbb{E}[X_S^2]$:

$$\mathbb{E}[X_S] = \mathbb{E}[\prod_{i \in S} Z_i] = \prod_{i \in S} \mathbb{E}[Z_i] = 0$$
$$\mathbb{E}[X_S^2] = \mathbb{E}[\prod_{i \in S} Z_i^2] = \prod_{i \in S} \mathbb{E}[Z_i^2] = 1$$

Thus, we can calculate $\mathsf{Var}(\sum_S X_S)$ as follows:

$$\begin{aligned}
\mathsf{Var}(\sum_S X_S) &= \mathsf{Var}(X_1 + X_2 + \cdots + X_n) \\
&= \mathbb{E}[(X_1 + X_2 + \cdots + X_n)^2] - \mathbb{E}[X_1 + X_2 + \cdots + X_n]^2 \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i^2] + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}[X_i X_j] - \sum_{i=1}^{n} \mathbb{E}[X_i]^2 - 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{E}[X_i]\, \mathbb{E}[X_j] \\
&= n + 0 - 0 - 0 \\
&= n
\end{aligned}$$

In addition, we can calculate $\sum_S \mathsf{Var}(X_S)$ as follows:

$$\begin{aligned}
\sum_S \mathsf{Var}(X_S) &= \mathsf{Var}(X_1) + \mathsf{Var}(X_2) + \cdots + \mathsf{Var}(X_n) \\
&= (\mathbb{E}[X_1] - \mathbb{E}[X_1]^2) + (\mathbb{E}[X_2] - \mathbb{E}[X_2]^2) + \cdots + (\mathbb{E}[X_n] - \mathbb{E}[X_n]^2) \\
&= (1 - 0) + (1 - 0) + \cdots + (1 - 0) \\
&= n
\end{aligned}$$

Obviously, the above results are consistent with the results in problem I(b).

Problem 2 (20 points.)**.** Consider a random distribution D over those bins $[n] = \{1, 2, ..., n\}$. We define its collision probability to be

$$\mathbf{Pr}_{a\sim D, b\sim D}[a = b]$$

where a and b are drawn from D independently.
Prove that the uniform distribution has the smallest collision probability among all distributions. This is the reason why we only consider uniform distribution over n bins in hash functions.

Hint 1. *Define a random variable X (depends on D) such that the collision probability is equal.*

Answer: First, we define a random variable X with the random distribution D, *i.e.*, $P(X = i) = \frac{1}{n}, \forall i \in \{1, 2, ..., n\}$.

Second, we define a random variable $Y = \{Y_1, Y_2, \cdots, Y_n\}$ with the distribution Y, we assume that $P(Y = i) = p_i$, and we can obtain $\sum_{i=1}^{n} p_i = 1$.

We define the collision probability of Y as $\sum_{i=1}^{n} p_i^2$, which means the probability that two variables drawn independently from Y are the same.

According to the *Average Inequality*: $\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \geq \frac{\sum_{i=1}^{n} x_i}{n}$, we have $\sum_{i=1}^{n} p_i^2 \geq n \cdot (\frac{\sum_{i=1}^{n} p_i}{n})^2 = n \cdot (\frac{1}{n})^2 = \frac{1}{n}$. Only when $p_i = \frac{1}{n}$, $\sum_{i=1}^{n} p_i^2$ can achieve the minimum value $\frac{1}{n}$, so that the collision probability of Y (*i.e.*, $\sum_{i=1}^{n} p_i^2$) achieves the minimum value $\frac{1}{n}$.

We can also calculate the l_2 distance of distribution Y and D to further prove the conclusion of the above minimum value:

$$\begin{aligned}
||Y - D||_2 &= \sqrt{\sum_{i=1}^{n}(p_i - \frac{1}{n})^2} \\
&= \sqrt{\sum_{i=1}^{n}(p_i^2 - \frac{2p_i}{n} + \frac{1}{n^2})} \\
&= \sqrt{\sum_{i=1}^{n} p_i^2 - \frac{2}{n}\sum_{i=1}^{n} p_i + \sum_{i=1}^{n}\frac{1}{n^2}} \\
&= \sqrt{\sum_{i=1}^{n} p_i^2 - \frac{2}{n} + \frac{1}{n}} \\
&= \sqrt{\sum_{i=1}^{n} p_i^2 - \frac{1}{n}}
\end{aligned}$$

Because $||Y - D||_2 \geq 0 \Rightarrow \sum_{i=1}^{n} p_i^2 \geq \frac{1}{n}$, with equality holds only when $p_i = \frac{1}{n}$.

Finally, we prove that the uniform distribution has the smallest possible collision probability over all distributions.

Problem 3 (Birthday paradox 10 points.)**.** Suppose everybody's birthday is a uniform random number in $\{1, 2, \cdots, 365\}$ independently. Now we wanna ask m persons' birthday such that with probability more than $\frac{1}{2}$, two of them will have the same birthday.

Show the best estimation of m.

Hint 2. *Think of this as balls into bins with $n = 365$ bins.*

Answer:

We transform the problem to at most how many persons do not have the same birthday with a probability of no more than $\frac{1}{2}$? The probability is as follows:

$$\mathbf{Pr}[\forall i \neq j, b_i \neq b_j] = (1 - \frac{1}{n})(1 - \frac{2}{n}) \cdots (1 - \frac{m-1}{n}) \leq \frac{1}{2}$$

According to *Taylor Formula*, the polynomial expansion of exponential function is as follows:

$$\exp(x) = \sum_{k=0}^{+\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \frac{x^4}{24} + \cdots$$

When $x \geq 0$, $1 + x \leq e^x$, and we replace it in the above equation:

$$\begin{aligned}
\mathbf{Pr}[\forall i \neq j, b_i \neq b_j] &= (1 - \frac{1}{n})(1 - \frac{2}{n}) \cdots (1 - \frac{m-1}{n}) \\
&\leq e^{-\frac{1}{n}} e^{-\frac{2}{n}} \cdots e^{-\frac{m-1}{n}} \\
&= e^{-\frac{1}{n} - \frac{1}{n} \cdots - \frac{1}{n}} \\
&= e^{-\frac{m(m-1)}{2n}} \leq \frac{1}{2}
\end{aligned}$$

Thus, we have:

$$\begin{gathered}
e^{-\frac{m(m-1)}{2n}} \leq \frac{1}{2}, \\
-\frac{m(m-1)}{2n} \leq -ln2, \\
m(m-1) - 2nln2 \geq 0, \\
m^2 - m - 2nln2 \geq 0, \\
m \geq \frac{1 + \sqrt{1 + 8nln2}}{2} \approx 23.
\end{gathered}$$

When $n = 365$, we can obtain the best estimation of m is 23.

Problem 4. Consider the following game between Alice and Bob:
1. At the beginning of the game, Alice record a secret binary string $z \in \{0,1\}^n$ of length n.
2. Bob guesses m strings $w_1, w_2, \cdots, w_m$ of length n and sends these to Alice.
3. We define the agreement of two strings x and y to be the number of the same entries in x and y. Then Alice announce Bob's score as the largest agreement between z and one string of $w_1, w_2, \cdots, w_m$. In another word, Bob has a score

$$\max_{i \in [m]} \{\sum_{j=1}^{n} 1\{z(j) = w_i(j)\}\}$$

Please design a strategy for Bob such that he could score as high as possible.

(a) (5 points) When $m = 2^n$, come up with a strategy to score n.

Answer: For a '0-1' variable string with the length of n, there are 2^n possibilities, just satisfying $m = 2^n$. Thus, Bob can directly enumerate the 2^n kinds of string z.

Bob can use a simple scheme, that is, converting decimal numbers $0 - 2^n$ into binary strings in turn (if the length is less than n, use '0' to fill in).

We show a Python code example, where $n = 4$ and $m = 2^4 = 16$:

```
>>> n = 4
>>> [bin(i)[2:].zfill(n) for i in range(pow(2, n))]
['0000', '0001', '0010', '0011', '0100', '0101', '0110', '0111', '1000',
'1001', '1010', '1011', '1100', '1101', '1110', '1111']
```

(b) (5 points) When $m = 2$, come up with a strategy to score at least $\frac{n}{2}$.

Answer: Bob can construct a string with length n that is all composed of 0, *i.e.*, $w_1 =$ '000...000', and a string that is all composed of 1, *i.e.*, $w_2 =$ '111...111'.

Assume that if Alice records the binary string with x '0' and y '1', and $x + y = n$.
Case 1, if $x = y = \frac{n}{2}$, $score = \frac{n}{2}$;
Case 2, if $x > y$, $score = \sum_{j=1}^{n} 1\{z(j) = w_1(j)\} = x > \frac{n}{2}$;
Case 3, if $x < y$, $score = \sum_{j=1}^{n} 1\{z(j) = w_2(j)\} = y > \frac{n}{2}$.
In summary, the score of the above strategy is at least $\frac{n}{2}$.

(c) (Optional with 20 bonus points) Let us prove the following strategy of 2 guesses can score $\frac{n}{2} + 0.1 \cdot \sqrt{n}$ with probability at least 0.5: Bob sends a random string w and its flip (on every coordinate) $\overline{w}$ to Alice.

Hint 3. *omitted*

Answer: First, Say the agreement between w and z is X. Since each coordinate of w and $\overline{w}$ takes the opposite value, it is obvious that $X_i = \{0,1\}, \forall i \in [n]$. In addition, $P(X_i = 0) = P(X_i = 1) = \frac{1}{2}$, $\mathbb{E}(X_i) = \frac{1}{2}$. Thus, we have $\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \frac{n}{2}$.
Then, we calculate $\mathsf{Var}(X)$. $\mathsf{Var}(X_i) = \mathbb{E}[(X_i - \mathbb{E}[X_i])^2] = \frac{1}{2} \cdot [(0-\frac{1}{2})^2 + (1-\frac{1}{2})^2] = \frac{1}{4}$. As X_i is independent with each other, according to the conclusion in Problem I(b), we have $\mathsf{Var}(X) = \sum_{i=1}^{n} \mathsf{Var}(X_i) = \frac{n}{4}$.
On the one hand, according to the conclusion in Problem I(a), we have $\mathbb{E}[|X - \mathbb{E}[X]|] \leq \sqrt{\mathsf{Var}(X)} = \frac{\sqrt{n}}{2} = \Theta(\sqrt{n})$.
On the other hand, according to *Chevyshev's inequality*, we have $\mathbf{Pr}[|X - \mathbb{E}[X]| \geq \sqrt{n}] \leq \frac{\frac{n}{4}}{(\sqrt{n})^2} = \frac{1}{4}$, *i.e.*, $\mathbf{Pr}[|X - \mathbb{E}[X]| \leq \sqrt{n}] \geq 1 - \frac{1}{4} = \frac{3}{4}$.
Formally, suppose the probabilities that the score X around $\frac{n}{2}$ are the same, which is the uniform distribution:

$$\mathbf{Pr}[X = \frac{n}{2} - \sqrt{n}] = \mathbf{Pr}[X = \frac{n}{2} - \sqrt{n} + 1] = \cdots = \mathbf{Pr}[X = \frac{n}{2}] = \cdots = \mathbf{Pr}[X = \frac{n}{2} + \sqrt{n}]$$

Thus, we have:

$$\begin{aligned}
\mathbf{Pr}[|X - \mathbb{E}[X]| \geq 0.1\sqrt{n}] &> \mathbf{Pr}[0.1\sqrt{n} \leq |X - \mathbb{E}[X]| \leq \sqrt{n}] \\
&= \mathbf{Pr}[0.1\sqrt{n} \leq |X - \mathbb{E}[X]|] \cdot \mathbf{Pr}[|X - \mathbb{E}[X]| \leq \sqrt{n}] \\
&= \frac{1 - 0.1}{1} \cdot \frac{3}{4} > \frac{1}{2}
\end{aligned}$$

Finally, we prove that the above strategy of 2 guesses can score $\frac{n}{2} + 0.1 \cdot \sqrt{n}$ with probability at least 0.5.

Problem 5 (20 points)**.** Consider the following generalization of Power of 2 choices to $d > 2$ choices:
1. Prepare d perfectly random hash functions $h_1, \cdots, h_d : U -> [n]$.
2. For each ball A, allocate it into the bin among his d choices $\{h_1(A), \cdots, h_d(A)\}$ with the lightest load at this moment.

Suppose we are throwing n balls into n bins. Modify the proof sketch in lecture 2 to show that: With probability $1 - n^{-2}$, the max-load is $\log_d \log n + O(1)$ instead of $\log_2 \log n + O(1)$

Answer:

Base case: $m_3 \leq \frac{n}{3}$;

According to *Chernoff bounds*,

Bounding m_4: since $\mathbb{E}[m_4] \leq \frac{n}{3^d}$, it implies $m_4 \leq 1.1 \cdot \frac{n}{3^d}$ w.h.p;

Then, we can obtain:

Bounding m_5: $m_5 \leq 1.1n \cdot (\frac{m_4}{n})^d = 1.1^{d+1} \cdot \frac{n}{3^{d^2}}$;

$\cdots$

Bounding m_l: $m_l \leq 1.1^{d^{l-3}-1} \cdot \frac{n}{3^{d^{l-3}}} \leq \frac{n}{2^{d^{l-3}}}$.

We take logarithms on both sides of the equation:

$$
\begin{aligned}
&\frac{n}{2^{d^{l-3}}} \leq 1 \\
\Rightarrow & n \leq 2^{d^{l-3}} \\
\Rightarrow & \log_2 n \leq d^{l-3} \\
\Rightarrow & \log_d \log_2 n \leq l-3
\end{aligned}
$$

Finally, we can prove that with probability $1 - n^{-2}$, the max-load is $\log_d \log n + O(1)$.