CIS 5020 Analysis of Algorithms | Fall 2024

Assignment 1

Due: September 11

Name: Yujie Xu, Richard Zhang

PennID: 51233809, 19331985

1 Question 1

1. (a) **Sub-question 1:**

Because $\mathcal{H}$ is a 2-wise independent hash function family, every hash value in $\{0,1\}^k$ should be equally likely for any input x from $\{0,1\}^n$.

We also know $\mathbf{0} \in \{0,1\}^k$, so we deduce this string 0 is merely one of all the possible outputs of the hash function.

Because $\mathcal{H}$ is a 2-wise independent hash function family, it's also 1-wise independent. Based on the definition of 1-wise independent hash function, for every $x \in \{0,1\}^n$ and for every $y \in \{0,1\}^k$, each possible output value $y \in \{0,1\}^k$ is equally likely for any given input x:

$$\Pr_{h\in\mathcal{H}}[h(x) = y] = \frac{1}{\textit{the number of all the possible outputs of the hash function}}$$

We know this pairwise independent hash function family maps $\{0,1\}^n$ to $\{0,1\}^k$, so the size of inputs is 2^n, and the size of outputs is 2^k. Thus:

$$\Pr_{h\in\mathcal{H}}[h(x) = 0] = \frac{1}{2^k}$$

Sub-question 2:

Since $\mathcal{H}$ is a pairwise independent hash function family, the values of $h(x)$ and $h(x')$ are independent (we assume $x \neq x'$).

We have already known: $\Pr_{h\in\mathcal{H}}[h(x) = 0] = \frac{1}{2^k}$, so we can deduce the probability of $h(x) = 0$ and $h(x') = 0$ is $\frac{1}{2^k}$ and $\frac{1}{2^k}$, respectively, as the following equations show:

$$\Pr_{h\in\mathcal{H}}[h(x) = 0] = \frac{1}{2^k}$$

Similarly,

$$\Pr_{h\in\mathcal{H}}[h(x') = 0] = \frac{1}{2^k}$$

Then, it's easy to deduce that the probability that both $h(x)$ and $h(x')$ equal the 0 string is the product of these two individual probabilities. Therefore:

$$\Pr_{h\in\mathcal{H}}[h(x) = h(x') = 0] = \Pr_{h\in\mathcal{H}}[h(x) = 0 \wedge h(x') = 0]$$

$$= \Pr_{h\in\mathcal{H}}[h(x) = 0] \cdot \Pr_{h\in\mathcal{H}}[h(x') = 0] = \frac{1}{2^k} \cdot \frac{1}{2^k} = \frac{1}{2^{2k}}$$

(b) **Sub-question 1:**

Based on the previous question and the definition of this random variable $\mathcal{E}_{x,h}$, we know $\Pr[h(x) = 0] = \frac{1}{2^k}$, so it's easy to compute $\mathbb{E}[\mathcal{E}_{x,h}]$:

$$\mathbb{E}[\mathcal{E}_{x,h}] = \Pr[h(x) = 0] = \frac{1}{2^k}$$

We also know:

$$\mathcal{E} = \sum_{x\in A} \mathcal{E}_{x,h}$$

So,

$$\mathbb{E}[\mathcal{E}] = \mathbb{E}[\sum_{x\in A} \mathcal{E}_{x,h}] = \sum_{x\in A} \mathbb{E}[\mathcal{E}_{x,h}] = |A| \cdot \mathbb{E}[\mathcal{E}_{x,h}]$$

$$|A| \cdot \mathbb{E}[\mathcal{E}_{x,h}] = p \cdot 2^k \cdot \frac{1}{2^k} = p$$

Thus, $\mathbb{E}[\mathcal{E}]$ is equal to p.

Sub-question 2:

$$\mathbb{E}[\mathcal{E}^2] = \mathbb{E}[(\sum_{x\in A} \mathcal{E}_{x,h})^2]$$

There would be two scenarios of $(\sum_{x\in A} \mathcal{E}_{x,h})^2$:

The first one is: $\sum_{x\in A} \mathcal{E}^2_{x,h}$, while the second one is $\sum_{x\neq x'} \mathcal{E}_{x,h} \cdot \mathcal{E}_{x',h}$.

Thus, we can rewrite the equation to the following form:

$$\mathbb{E}[(\sum_{x\in A} \mathcal{E}_{x,h})^2] = \mathbb{E}[\sum_{x\in A} \mathcal{E}^2_{x,h}] + \mathbb{E}[\sum_{x\neq x'} \mathcal{E}_{x,h}\mathcal{E}_{x',h}]$$

Simplify it a bit:

$$= \sum_{x\in A} [\mathbb{E}[\mathcal{E}^2_{x,h}] + \sum_{x\in A} [\mathbb{E}[\mathcal{E}_{x,h}\mathcal{E}_{x',h}]$$

Because $\mathcal{E}_{x,h}$ is either 0 or 1, $(\mathcal{E}_{x,h})^2$ is either 0 or 1. This implies $\mathcal{E}^2_{x,h} = \mathcal{E}_{x,h}$.

So, we can deduce:

$$\sum_{x\in A}[\mathbb{E}[\mathcal{E}_{x,h}^2] = \sum_{x\in A}[\mathbb{E}[\mathcal{E}_{x,h}] = |A| \cdot \mathbb{E}[\mathcal{E}_{x,h}] = p \cdot 2^k \cdot \frac{1}{2^k} = p$$

The second term is a bit tricky:

$$\sum_{x\in A}[\mathbb{E}[\mathcal{E}_{x,h}\mathcal{E}_{x',h}] = \sum_{x\in A} \Pr_{h\in\mathcal{H}}[h(x) = h(x') = 0] = \Pr_{h\in\mathcal{H}}[h(x) = 0 \wedge h(x') = 0]$$

(we assume $x \neq x'$)

The result of this equation is calculated in the previous question. Therefore:

$$\sum_{x\in A}[\mathbb{E}[\mathcal{E}_{x,h}\mathcal{E}_{x',h}] = \sum_{x\in A}[\Pr_{h\in\mathcal{H}}[h(x) = 0] \cdot \Pr_{h\in\mathcal{H}}[h(x') = 0] = \sum_{x\in A}[\frac{1}{2^k} \cdot \frac{1}{2^k}] = \sum_{x\in A}[\frac{1}{2^{2k}}]$$

$$\sum_{x\neq x'} \mathbb{E}[\mathcal{E}_{x,h}\mathcal{E}_{x',h}] = 2 \cdot \binom{|A|}{2} \cdot \frac{1}{2^{2k}} = \frac{|A|(|A|-1)}{2} \cdot \frac{1}{2^{2k}}$$

$$= 2 \cdot \frac{p \cdot 2^k(p \cdot 2^k - 1)}{2} \cdot \frac{1}{2^{2k}}$$

$$= 2 \cdot \frac{p \cdot (p \cdot 2^k - 1)}{2 \cdot 2^k}$$

$$= 2 \cdot \frac{p^2 \cdot 2^k - p}{2 \cdot 2^k}$$

$$= p^2 - \frac{p}{2^k}$$

Thus, the final result is:

$$\mathbb{E}[\mathcal{E}^2] = p + p^2 - \frac{p}{2^k}$$

(c) **Sub-question 1:**

$$(\mathcal{E} - 1)^2 = \mathcal{E}^2 - 2\mathcal{E} + 1$$

$$\mathbb{E}[(\mathcal{E} - 1)^2] = \mathbb{E}[(\mathcal{E}^2 - 2 \cdot \mathcal{E} + 1)]$$

$$= \mathbb{E}[\mathcal{E}^2] - 2 \cdot \mathbb{E}[\mathcal{E}] + 1$$

Based on the previous question, we know $\mathbb{E}[\mathcal{E}^2] = p + p^2 - \frac{p}{2^k}$, and $\mathbb{E}[\mathcal{E}] = p$. Thus:

$$\mathbb{E}[(\mathcal{E} - 1)^2] = \mathbb{E}[\mathcal{E}^2] - 2 \cdot \mathbb{E}[\mathcal{E}] + 1 = p + p^2 - \frac{p}{2^k} - 2 \cdot p + 1$$

$$= p^2 - \frac{p}{2^k} - p + 1$$

Sub-question 2:

Since k should be much larger than p, we can deduce that:

$$\mathbb{E}[(\mathcal{E}-1)^2] = p^2 - \frac{p}{2^k} - p + 1 \leq p^2 - p + 1$$

Based on Markov's inequality, we know that for any non-negative random variable $\mathcal{X}$ and a such that $a > 0$,

$$\Pr(\mathcal{X} \geq a) \leq \frac{\mathbb{E}[\mathcal{X}]}{a}$$

By plugging in $\mathcal{X} = (\mathcal{E}-1)^2$, we can get the following equation:

$$\Pr[(\mathcal{E}-1)^2 \geq a] \leq \frac{\mathbb{E}[(\mathcal{E}-1)^2]}{a}$$

Let a = 1, we can rewrite the equation to:

$$\Pr[(\mathcal{E}-1)^2 \geq 1] \leq \mathbb{E}[(\mathcal{E}-1)^2]$$

Since $\mathbb{E}[(\mathcal{E}-1)^2] \leq p^2 - p + 1$, we know:

$$\Pr[(\mathcal{E}-1)^2 \geq 1] \leq p^2 - p + 1$$

$(\mathcal{E}-1)^2 \geq 1$ is only possible when $\mathcal{E} \neq 1$, so we can deduce:

$$\Pr[(\mathcal{E}-1)^2 \geq 1] = \Pr[(\mathcal{E} \neq 1)] \leq p^2 - p + 1$$

Since we know $\frac{1}{4} \leq p \leq \frac{1}{2}$, we can plug in $p = \frac{1}{4}$ and get this equation:

$$\Pr[(\mathcal{E} \neq 1)] \leq p^2 - p + 1 = \frac{13}{16}$$

Thus,

$$\Pr[(\mathcal{E} = 1)] = 1 - \Pr[(\mathcal{E} \neq 1)] \geq 1 - p^2 - p + 1 = 1 - \frac{13}{16} = \frac{3}{16}$$

$$\Pr[(\mathcal{E} = 1)] \geq \frac{3}{16} \geq \frac{1}{16}$$

Therefore,

$$\Pr[(\mathcal{E} = 1)] \geq \frac{1}{16}$$

2 Question 2

The statement that with probability at least $1-1/n$, a sequence of n independent tosses will contain a run of length at least $\log_2 n - 2\log_2\log_2 n$ has the equivalent meaning with the following statement:

Theorem 1. Given a sequence of n independent tosses $\mathcal{S}$, the longest run in $\mathcal{S}$ has a length of r. The event $r \geq \log_2 n - 2\log_2\log_2 n$ is named A. Then $Pr(A) \geq 1 - \frac{1}{n}$.

This is equivalent to prove that:

Theorem 2. Given a sequence of n independent tosses $\mathcal{S}$, the longest run in $\mathcal{S}$ has a length of r. The event $r < \log_2 n - 2\log_2\log_2 n$ is named A. Then $Pr(A) < \frac{1}{n}$.

Given a block of sequence $\mathcal{S}'$ with a size of t, the probability that there are all *Tails* in this block is:

$$\Pr = (\frac{2}{3})^t$$

Therefore, the probability that there are at least one *Head* in $\mathcal{S}'$ is:

$$\Pr = 1 - (\frac{2}{3})^t$$

Therefore, the probability that no such block in k disjoint blocks $\{\mathcal{S}'\}$ with at least the size of t has all *Tails* in the whole sequence $\mathcal{S}$:

$$\Pr_1 = (1 - (\frac{2}{3})^t)^k$$

Similarly, the probability that no such block in k disjoint blocks $\{\mathcal{S}'\}$ with at least the size of t has all *Heads* in the whole sequence $\mathcal{S}$:

$$\Pr_2 = (1 - (\frac{1}{3})^t)^k$$

let $t = \lfloor \frac{1}{2}\log_2 n - \log_2\log_2 n \rfloor$, split the whole sequence $\mathcal{S}$ into at least $k = \frac{2n}{\log_2 n}$ disjoint blocks $\{\mathcal{S}'\}$. Note that:

$$\frac{2n}{\log_2 n}\lfloor\frac{1}{2}\log_2 n - \log_2\log_2 n\rfloor \leq \frac{2n}{\log_2 n}(\frac{1}{2}\log_2 n - \log_2\log_2 n) \leq n$$

After splitting $\mathcal{S}$ into continuous yet disjoint blocks there are some tosses left out, let them be A. Then $A \subset \mathcal{S}, \{\mathcal{S}'\} \cap A = \mathcal{S}$. It is obvious that:

$$\Pr(r<2t \text{ happens in } \mathcal{S}) \leq \Pr(r<2t \text{ happens in} \{\mathcal{S}'\}) = \Pr_1 + \Pr_2$$

Since the previous event is conditioned on the latter one.

Lemma 1. If there are two continuous yet disjoint blocks with the size of t that both has more than one run inside, i.e. does not have all *Tails* or *Heads* inside, then the longest run in these two blocks as a whole has a size less than $2t$.

Proof of Lemma 1. Suppose the longest run in these two blocks as a whole has a size of $2t$. Then this sequence have to be all *Tails* or *Heads*. There is a contradiction. Therefore the length of the longest run has to be less than $2t$.

From the *Lemma 1* we prove that if all the blocks in disjoint blocks $\{\mathcal{S}'\}$ have more than one run inside, then the length of the longest run $r < 2t$ in $\{\mathcal{S}'\}$. Note that:

$$(\frac{1}{4})^t \geq (\frac{1}{4})^{\frac{1}{2}\log_2 n - \log_2\log_2 n} = \frac{\log_2^2 n}{n}$$

From the inequality $1 - x \leq 2^{-x}, x \geq 0$

$$1 - \frac{\log_2^2 n}{n} \leq 2^{-\frac{\log_2^2 n}{n}}$$
$$(1 - \frac{\log_2^2 n}{n})^{\frac{2n}{\log_2 n}} \leq 2^{-2log_2 n} = \frac{1}{n^2}$$

Note that, when $n > 2$:

$$(\frac{2}{3})^t \geq (\frac{1}{4})^t \geq \frac{\log_2^2 n}{n},\ (\frac{1}{3})^t \geq (\frac{1}{4})^t \geq \frac{\log_2^2 n}{n}$$
$$(1 - (\frac{2}{3})^t)^{\frac{2n}{\log_2 n}} \leq (1 - \frac{\log_2^2 n}{n})^{\frac{2n}{\log_2 n}},\ (1 - (\frac{1}{3})^t)^{\frac{2n}{\log_2 n}} \leq (1 - \frac{\log_2^2 n}{n})^{\frac{2n}{\log_2 n}}$$
$$\Pr_1 + \Pr_2 \leq 2(1 - \frac{\log_2^2 n}{n})^{\frac{2n}{\log_2 n}} = \frac{2}{n^2} < \frac{1}{n}$$

When $n = 1, 2$, the proof is trivial. Therefore the *Theorem 2* is proved, and thus the *Theorem 1* is proved.

3 Question 3

From a set of n elements, we select k elements $1, i_1, i_2, \ldots, i_k$. The probability that its permutation $\sigma(i_1) < \sigma(i_2) < \cdots < \sigma(i_k)$ is:

$$\Pr = \frac{C(n,k)}{P(n,k)} = \frac{1}{k!}$$

Let X be a random variable, denoting the number of increasing sub-sequences of length k in a random permutation $\sigma : [n] \to [n]$. Then we can deduce that the expected value of X is:

$$\mathbb{E}(X) = \sum x \cdot \Pr(X = x) = \binom{n}{k} \frac{1}{k!} \leq (\frac{e^2 n}{k^2})^k$$

Let $k = C\sqrt{n}$, then the expectation of X under this k is:

$$\mathbb{E}(X_k) \leq (\frac{e}{c})^{2C \cdot \sqrt{n}}$$

Using Markov's inequality, we can get the following equation:

$$\Pr[X \geq t \cdot \mathbb{E}(X)] \leq \frac{1}{t}$$

Thus:

$$\Pr[X_k \geq t \cdot (\frac{e}{c})^{2C \cdot \sqrt{n}}] \leq \frac{1}{t}$$

Set $t = (\frac{e}{C})^{-2C \cdot \sqrt{n}}$. Then:

$$\Pr[X_k \geq 1] \leq \frac{1}{t}$$

This is equivalent to say that with probability less and equal to $\frac{1}{t}$, there is at least one increasing sub-sequence of length k in permutation. i.e.

$$\Pr[\text{LIS}(\sigma) \geq k] \leq \frac{1}{t}$$

To prove $\exists C' \in \mathbb{R}, \Pr(\text{LIS}(\sigma) < C'\sqrt{n}) > \frac{9}{10}$, is equivalent to prove that:

$$\Pr[\text{LIS}(\sigma) \geq C'\sqrt{n}] \leq \frac{1}{10}$$

Set $c = 2e$, then $t = (\frac{1}{2})^{-4e\sqrt{n}}$. Based on this, we can deduce:

$$\Pr[\text{LIS}(\sigma) \geq 2e\sqrt{n}] \leq (\frac{1}{2})^{4e\sqrt{n}}$$

Since $n \geq 1$, we can conclude:

$$\Pr[\text{LIS}(\sigma) \geq 2e\sqrt{n}] \leq (\frac{1}{2})^{4e\sqrt{n}} \leq (\frac{1}{2})^{4e} \leq (\frac{1}{2})^4 = \frac{1}{16} \leq \frac{1}{10}$$

Thus,

$$\exists C' = 2e \in \mathbb{R}, \Pr(\text{LIS}(\sigma) < C'\sqrt{n}) > \frac{9}{10}$$

Thus,

$$\text{LIS}(\sigma) = O(\sqrt{n})$$