Randomness in Computing

LECTURE 8

CS 537

Last time

- Markov's inequality
- Variance, covariance
- Variance of binomial and geometric RVs

Today

- Chebyshev's inequality
- Finding the median of an array

Recall: variance

- The variance of a random variable X with expectation $\mathbb{E}[X] = \mu$ is
$$\mathrm{Var}[X] = \mathbb{E}[(X - \mu)^2].$$
- Equivalently, $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mu^2$.

Variance: additional facts

- Theorem. For $a, b \in \mathbb{R}$ and a random variable X,
$$\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X].$$
- Theorem. If $X_1, \dots, X_n$ are pairwise independent random variables, then
$$\mathrm{Var}[X_1 + \cdots + X_n] = \mathrm{Var}[X_1] + \cdots + \mathrm{Var}[X_n].$$

Chebyshev's Inequality

Theorem (Chebyshev's Inequality)

Let X be a random variable. For all $a > 0$,

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Pafnuty Chebyshev
[1821-1894]

Proof: $\Pr[|X - \mathbb{E}[X]| \geq a] = \Pr[\underbrace{(X - \mathbb{E}[X])^2}_{Y \geq 0} \geq a^2]$

$$\leq \frac{\mathbb{E}[Y]}{a^2} \quad \text{(by Markov)}$$

$$= \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2}$$

$$= \frac{\mathrm{Var}[X]}{a^2}$$

image source https://www.britannica.com/biography/Pafnuty-Lvovich-Chebyshev

Chebyshev's Inequality: Corollary

Theorem (Chebyshev's Inequality)

Let X be a random variable. For all $a > 0$,

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}(X)}{a^2}.$$

Alternative form: for all $t > 1$,

$$\Pr[|X - \mathbb{E}[X]| \geq t \cdot \sigma[X]] \leq \frac{1}{t^2}.$$

Pafnuty Chebyshev
[1821-1894]

image source https://www.britannica.com/biography/Pafnuty-Lvovich-Chebyshev

Chebyshev's Inequality: Example 1

$X \sim \text{Bin}(n,1/2)$. Bound $\Pr\left[X \geq \frac{3n}{4}\right]$ using Markov and Chebyshev.

Solution: $\mathbb{E}[X] =$; $\text{Var}[X] =$

Markov: $\Pr\left[X \geq \frac{3n}{4}\right] \leq$

Chebyshev: $\Pr[|X - \mathbb{E}[X]| \geq \quad] \leq$

$\Pr\left[X \geq \frac{3n}{4}\right] =$

For $n > 3$, Chebyshev is much better!

Chebyshev's Inequality: Example 2

Coupon Collector Problem. $X = \#$ of trials to collect n coupons.

Bound $\Pr[X > 2nH_n]$ using Markov and Chebyshev.

Solution: $\mathbb{E}[X] =$

Markov: $\Pr[X > 2nH_n] \leq$

Chebyshev's Inequality: Example 2

Solution with Chebyshev: Recall that $X = \sum_{i\in[n]} X_i$, where

- X_i is the number of trials to collect the i^{th} coupon
- $X_i \sim Geom\left(\frac{n-i+1}{n}\right)$
- Random variables X_i are independent

By independence

$$\mathrm{Var}[X] = \mathrm{Var}\left[\sum_{i\in[n]} X_i\right] = \sum_{i\in[n]} \mathrm{Var}[X_i]$$

$$\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$$

$$\mathrm{Var}[X_i] = \frac{1-p}{p^2} \le \frac{1}{p^2} = \frac{n^2}{(n-i+1)^2}$$

$$\mathrm{Var}[X] = \sum_{i\in[n]} \mathrm{Var}[X_i] \le \sum_{i\in[n]} \frac{n^2}{(n-i+1)^2} = n^2 \sum_{i\in[n]} \frac{1}{i^2} \le \frac{\pi^2}{6} n^2$$

Chebyshev: $\Pr[X > 2nH_n] \le \Pr[|X - nH_n| \ge nH_n]$

$$\le \frac{\mathrm{Var}[X]}{(nH_n)^2}$$

Union Bound: an Even Better Solution

Let F_i be the event that we failed to collect coupon of type i in $2n \ln n$ trials

By a union bound

By symmetry

$$\Pr[X > 2nH_n] = \Pr\left[\bigcup_{i \in [n]} F_i\right] \leq \sum_{i \in [n]} \Pr[F_i] = n \Pr[F_1]$$

$$\Pr[F_1] = \left(1 - \frac{1}{n}\right)^{2n \ln n} \leq$$

Randomized Algorithm for the Median of an Array

- Given elements $a_1 \leq a_2 \leq \cdots \leq a_n$, their *median* is $a_{\lceil \frac{n}{2} \rceil}$.

Task: Find the median of an array if the elements are not sorted.

- Deterministic algorithm (median of medians): $O(n)$ time.
- Today: simple randomized algorithm: $O(n)$ time.

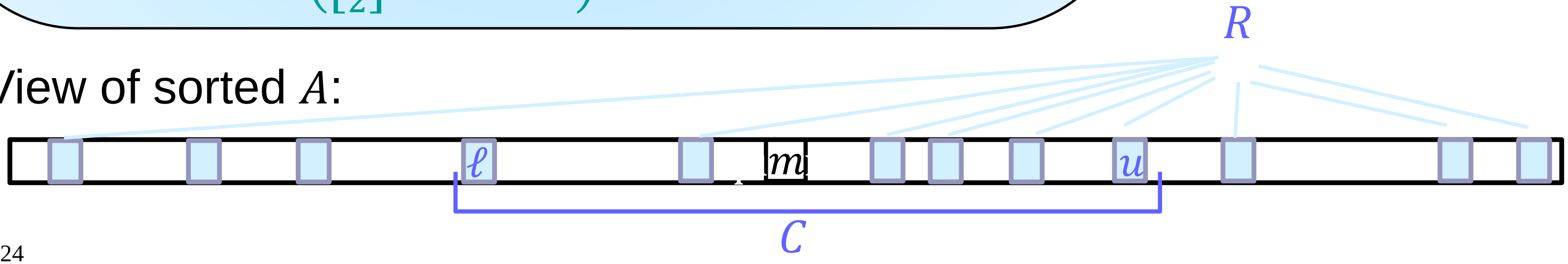
- Simplifying assumptions:
  - all elements are distinct;
  - n is odd;
  - we can sample from the array in constant time.
- Idea: Sample to find elements ℓ and u such that
  1. $\ell \leq m \leq u$, where m denotes the median.
  2. The number of input elements that lie in the interval (ℓ, u) is small.

Randomized Median Algorithm

Input: array A of elements $a_1, \ldots, a_n$
Output: median of A

1. Let R be an array $r_1, \ldots, r_t$, where each r_i is chosen from A u.i.r. with replacement, where $t = \lceil n^{3/4} \rceil$.
2. Sort R.
3. Let ℓ be the $\left\lceil \frac{n^{3/4}}{2} - \sqrt{n} \right\rceil$-th smallest element in R.
4. Let u be the $\left\lceil \frac{n^{3/4}}{2} + \sqrt{n} \right\rceil$-th smallest element in R.
5. Use PARTITION from Quicksort to compute
$$C = \{a \in \mathrm{A} \mid \ell \leq a \leq u\},$$
$$n_\ell = |\{\, a \in A \mid a < \ell \,\}| \text{ and } n_u = |\{\, a \in A \mid a > u \,\}|$$
6. If $n_\ell > \left\lceil \frac{n}{2} \right\rceil$ or $n_u > \left\lceil \frac{n}{2} \right\rceil$ then **fail**.
7. If $|C| \leq 4n^{3/4}$ then sort C; otherwise **fail**.
8. Output the $\left(\left\lfloor \frac{n}{2} \right\rfloor - n_\ell + 1 \right)$-th smallest element in C.

View of sorted A:



Analysis

Theorem 1

Randomized Median Algorithm (RMA) terminates in $O(n)$ time.

It outputs either **fail** or the median.

Theorem 2

RMA outputs **fail** with probability at most $n^{-1/4}$.

Proof: Bad events $\mathcal{E}_1, \mathcal{E}_2$, and $\mathcal{E}_3$

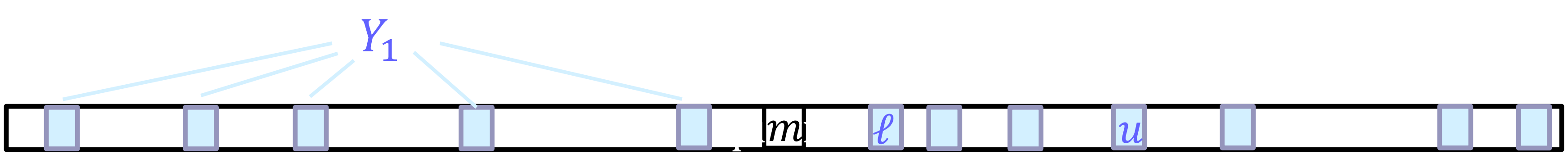
$$Y_1 = |\{i \in [t] : r_i \leq m\}|$$

$$Y_2 = |\{i \in [t] : r_i \geq m\}|$$

$$\mathcal{E}_1: \quad Y_1 < \frac{n^{3/4}}{2} - \sqrt{n} \qquad \mathcal{E}_2: \quad Y_2 < \frac{n^{3/4}}{2} - \sqrt{n}$$

$$\mathcal{E}_3: \quad |C| > 4n^{3/4}$$

- RMA fails iff $\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3$ occurs



Analysis

Lemma 1. $\Pr[\mathcal{E}_1] \leq \frac{1}{4} \cdot \frac{1}{n^{1/4}}$

$Y_1 = |\{i \in [t] : r_i \leq m\}|$

$\mathcal{E}_1$: $\quad Y_1 < \frac{n^{3/4}}{2} - \sqrt{n}$

Proof: Recall: $t = n^{3/4}$. For all $i \in [t]$, define

$$X_i = \begin{cases} 1 & \text{if } r_i \leq m \\ 0 & \text{otherwise} \end{cases}$$

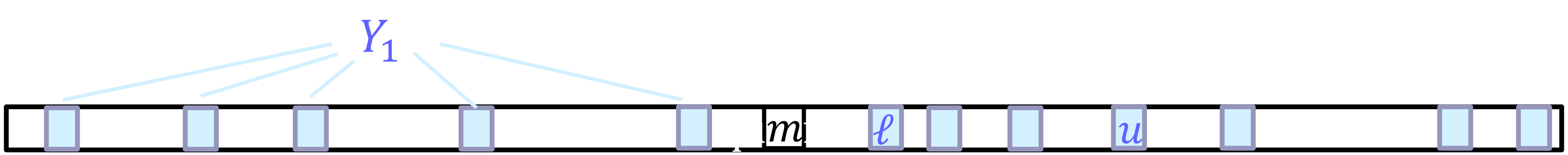
$p = \Pr[X_i = 1] =$

$$Y_1 = \sum_{i \in [t]} X_i$$

$\mathbb{E}[Y_1] =$

$\mathrm{Var}[Y_1] =$

By Chebyshev: $\Pr[\mathcal{E}_1] = \Pr\left[Y_1 < \frac{n^{3/4}}{2} - \sqrt{n}\right] \quad \leq \Pr[|Y_1 - \mathbb{E}[Y_1]| > \sqrt{n}]$



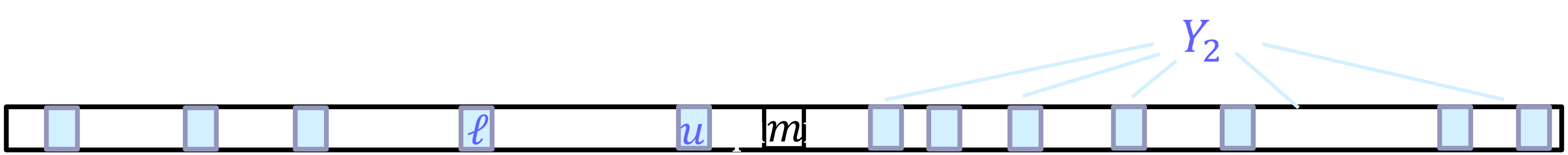
Analysis

Lemma 2. $\Pr[\mathcal{E}_2] \le \frac{1}{4} \cdot \frac{1}{n^{1/4}}$

$$Y_2 = |\{i \in [t] : r_i \ge m\}|$$

$$\mathcal{E}_2: \quad Y_2 < \frac{n^{3/4}}{2} - \sqrt{n}$$

Proof: The same as proof for Lemma 1.



Analysis

Lemma 3. $\Pr[\mathcal{E}_3] \le \frac{1}{2} \cdot \frac{1}{n^{1/4}}$

$\mathcal{E}_3$: $|C| > 4n^{3/4}$

Proof: Define events

$\mathcal{E}_{3,1}$: $\ge 2n^{3/4}$ elements of C are greater than the median m

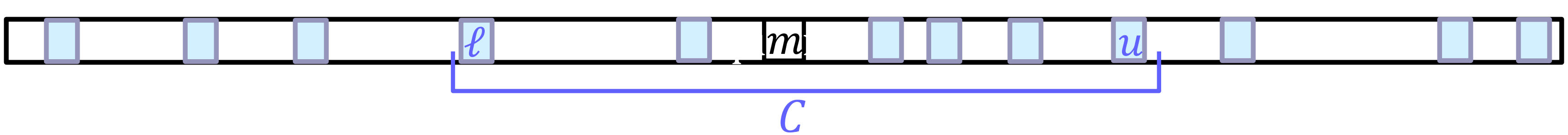
$\mathcal{E}_{3,2}$: $\ge 2n^{3/4}$ elements of C are smaller than the median m

By a union bound, $\Pr[\mathcal{E}_3] \le \Pr\left[\mathcal{E}_{3,1}\right] + \Pr\left[\mathcal{E}_{3,2}\right] = 2\Pr\left[\mathcal{E}_{3,1}\right]$

$\mathcal{E}_{3,1}$ holds $\Leftrightarrow$ rank of u in A is

but we threw out $\frac{n^{3/4}}{2} - \sqrt{n}$ samples in R with a larger value than u

$\ge \frac{n^{3/4}}{2} - \sqrt{n}$ samples in R are among largest in A



Analysis

Lemma 3. $\Pr[\mathcal{E}_3] \le \frac{1}{2} \cdot \frac{1}{n^{1/4}}$

$\mathcal{E}_3$: $|C| > 4n^{3/4}$

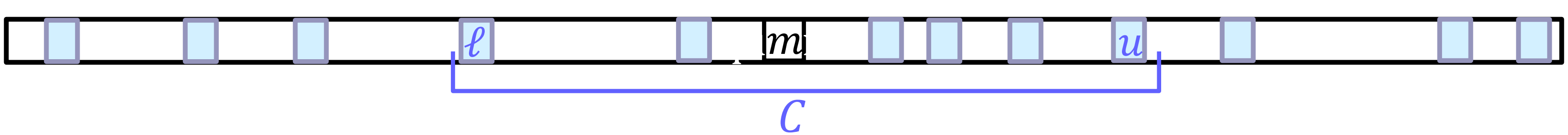
Proof: $\mathcal{E}_{3,1}$: $\ge 2n^{3/4}$ elements of C are greater than the median m

$\mathcal{E}_{3,1}$ holds $\Leftrightarrow \ge \frac{n^{3/4}}{2} - \sqrt{n}$ samples in R are among $\frac{n}{2} - 2n^{3/4}$ largest in A

Recall: $t = n^{3/4}$. For all $i \in [t]$, define

$$X_i = \begin{cases} 1 & \text{if } r_i \text{ is among } \frac{n}{2} - 2n^{3/4} \text{ largest elements in } A \\ 0 & \text{otherwise} \end{cases}$$

$$X = \sum_{i \in [t]} X_i$$



Monte Carlo vs. Las Vegas

- Monte Carlo: a randomized algorithm that may fail or produce an incorrect answer.
- Las Vegas: a randomized algorithm that always returns the right answer.

- We can get a Las Vegas algorithm from a Monte Carlo algorithm that may fail by repeating until it succeeds.