LECTURE 9

Last time
• Chebyshev’s inequality
• Variance of Binomial and Geometric RVs

Today
• Computing the median of an array
Randomized Algorithm for the Median of an Array

- Given elements $a_1 \leq a_2 \leq \cdots \leq a_n$, their median is $a_{\lfloor n/2 \rfloor}$.

**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 $\ell$ 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 $(\ell, u)$ is small.
Randomized Median Algorithm

**Input:** array $A$ of elements $a_1, ..., a_n$

**Output:** median of $A$

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

<table>
<thead>
<tr>
<th><strong>Theorem 1</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized Median Algorithm (RMA) terminates in $O(n)$ time. It outputs either <strong>fail</strong> or the median.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>Theorem 2</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>RMA outputs <strong>fail</strong> with probability at most $n^{-1/4}$.</td>
</tr>
</tbody>
</table>

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

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

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

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

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

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

- RMA fails iff $\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3$ occurs
Lemma 1. \( \Pr[\mathcal{E}_1] \leq \frac{1}{4} \cdot \frac{1}{n^{1/4}} \)

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

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

\[
p = \Pr[X_1 = 1] = 
\]

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

\[
\mathbb{E}[Y_1] = 
\]

\[
\text{Var}[Y_1] = 
\]

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

\[ \text{Pr}[\mathcal{E}_2] \leq \frac{1}{4} \cdot \frac{1}{n^{1/4}} \]

**Proof:** The same as proof for Lemma 1.

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

\[ \mathcal{E}_2: \quad Y_2 < \frac{n^{3/4}}{2} - \sqrt{n} \]
Lemma 3. \( \Pr[\mathcal{E}_3] \leq \frac{1}{2} \cdot \frac{1}{n^{1/4}} \)

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

Proof: Define events

\[ \mathcal{E}_{3,1}: \quad \geq 2n^{3/4} \text{ elements of } C \text{ are greater than the median } m \]

\[ \mathcal{E}_{3,2}: \quad \geq 2n^{3/4} \text{ elements of } C \text{ are smaller than the median } m \]

By a union bound, \( \Pr[\mathcal{E}_3] \leq \Pr[\mathcal{E}_{3,1}] + \Pr[\mathcal{E}_{3,2}] = 2 \Pr[\mathcal{E}_{3,1}] \)

\( \mathcal{E}_{3,1} \) holds \( \iff \) 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 \( u \)

\( \geq \frac{n^{3/4}}{2} - \sqrt{n} \) samples in \( R \) are among \( C \) largest in \( A \)
**Analysis**

**Lemma 3.** \( \Pr[\mathcal{E}_3] \leq \frac{1}{2} \cdot \frac{1}{n^{1/4}} \)

**\( \mathcal{E}_3: \)** \(|C| > 4n^{3/4} \)

**Proof:**

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

\( \mathcal{E}_{3,1} \) holds \( \iff \geq \frac{n^{3/4}}{2} - \sqrt{n} \) 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

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

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