CSE 632: Analysis of Algorithms II: Randomized Algorithms **Spring 2024**

Lecture 5: Median

Lecturer: Zongchen Chen

1 Median and QuickSelect

Median problem: Given an unsorted list $S = [s_1, \ldots, s_n]$ of integers, find the median of S (i.e., the $(n/2)$th smallest element).

The more general version is the following problem.

Selection problem: Given an unsorted list $S = [s_1, \ldots, s_n]$ of integers and an integer $k \in [n]$, find the kth smallest element.

One can find the median by sorting S which takes $O(n \log n)$ time. The median of medians algorithm is a deterministic algorithm which finds the median in $O(n)$ time.

QuickSelect is a simple randomized algorithm with expected running time $O(n)$. The approach is described as follow.

1. Find a good "pivot" $p \in S$;
2. Partition S into $S_{\le p} = \{x \in S : x \le p\}$ and $S_{>p} = \{x \in S : x > p\}$;
3. The median is in one of $S_{\le p}$ and $S_{>p}$, and then recurse.

A pivot $p \in S$ is said to be *good* if both $|S_{\le p}| \le 3n/4$ and $|S_{>p}| \le 3n/4$. Assuming we always get a good pivot, the running time of QuickSelect then satisfies the recursion:

$$T(n) = T\left(\frac{3n}{4}\right) + O(n).$$

Solving the recursion gives $T(n) = O(n)$.

We still need a strategy to find good pivots. This can be done by picking random elements from S. Notice that a pivot $p \in S$ is good if it is in between the $(n/4)$th smallest element and the $(3n/4)$th smallest. Hence, there are (at least) $n/2$ good pivots, and with probability $1/2$ a random element $p \in S$ is a good pivot. We can thus use the following procedure to find a good pivot: Choose a random pivot $p \in S$ and check if p is good; if not, repeat. The number of rounds needed is $O(1)$ in expectation (it has geometric distribution). Therefore, the overall running time of QuickSelect is $O(n)$ in expectation.

2 A Randomized Algorithm for Median

Here we give a simple randomized algorithm that has $O(n)$ running time with high probability. More precisely, there exists a constant $c > 0$ such that, with probability at least $1 - n^{-c}$ our algorithm successfully finds the median with running time $O(n)$. In other words, the running time of our algorithm is $O(n)$ while the success probability is at least $1 - n^{-c}$. Note that the expected running time being $O(n)$ (like QuickSelect) does not imply such high success probability.

The ideas of our randomized algorithm are as follows.

1. Find ℓ and u in S such that

   (1) $\ell \le m$;

(2) $u \geq m$;

(3) $C = \{x \in S : \ell \leq x \leq u\}$ is small.

"ℓ and u are lower and upper bounds on m which are very close to m, so that m is in a small center set C."

2. Find m in C by sorting C.

For this algorithm to run in $O(n)$ time we need $|C| = o(n)$ since we need to sort C. The major question is how to find such "good" ℓ and u. We achieve this by inspecting a random subset of S. More precisely, let R be a set of $n^{3/4}$ random elements of S chosen with replacement (for simplicity); note that R may be a multiset. Such R serves as a sketch or approximation of the original set S:

- R is much smaller in size (e.g., sorting R takes $o(n)$ time);
- R preserves or approximates important information of S (e.g., the median of R is "close" to m).

We now state the algorithm in full details.

Algorithm 1 A randomized algorithm for median

Input: $S = [s_1, \ldots, s_n]$
Output: median m of S

$R \leftarrow$ multiset of $n^{3/4}$ random elements of S (with replacement);
Sort R;
$\ell \leftarrow (\frac{1}{2}n^{3/4} - \sqrt{n})$th smallest of R;
$u \leftarrow (\frac{1}{2}n^{3/4} + \sqrt{n})$th smallest of R;
$C \leftarrow \{x \in S : \ell \leq x \leq u\}$;
$S_{<\ell} \leftarrow \{x \in S : x < \ell\}$;
$S_{>u} \leftarrow \{x \in S : x > u\}$;
if $|S_{<\ell}| \geq n/2$ ($\Leftrightarrow \ell > m$) **then** ▷ Bad event $\mathcal{E}_1$
  return FAIL
end if
if $|S_{>u}| \geq n/2$ ($\Leftrightarrow u < m$) **then** ▷ Bad event $\mathcal{E}_2$
  return FAIL
end if
if $|C| > 4n^{3/4}$ **then** ▷ Bad event $\mathcal{E}_3$
  return FAIL
end if
▷ Otherwise, $\ell \leq m \leq u$ and $|C| \leq 4n^{3/4}$
Sort C;
  return $(n/2 - |S_{<\ell}|)$th smallest of C

Running time. It is not hard to see that the running time of Algorithm 1 is $O(n)$.

Success probability. Observe that if Algorithm 1 does not return FAIL, then it successfully finds the median m. Furthermore, the algorithm fails iff at least one of the following three bad events happens:

- $\mathcal{E}_1 = \left\{|\{r \in R : r \leq m\}| < \frac{1}{2}n^{3/4} - \sqrt{n}\right\}$;
- $\mathcal{E}_2 = \left\{|\{r \in R : r \geq m\}| < \frac{1}{2}n^{3/4} - \sqrt{n}\right\}$;

- $\mathcal{E}_3 = \left\{|\{x \in S : \ell \le x \le u\}| > 4n^{3/4}\right\}$.

By the union bound,

$$\Pr(\text{FAIL}) = \Pr(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3) \le \Pr(\mathcal{E}_1) + \Pr(\mathcal{E}_2) + \Pr(\mathcal{E}_3).$$

We shall upper bound the probability of each of these bad events in the next three lemmas respectively. Together they imply that Algorithm 1 succeeds with probability at least $1 - n^{-1/4}$.

Lemma 1. $\Pr(\mathcal{E}_1) \le \frac{1}{4}n^{-1/4}$.

Proof. Let $r_1, \ldots, r_{n^{3/4}}$ be elements of R. For each i define

$$X_i = \begin{cases} 1, & \text{if } r_i \le m; \\ 0, & \text{o/w.} \end{cases}$$

Note that $\mathbb{E}X_i = \Pr(X_i = 1) = \Pr(r_i \le m) = 1/2$. Further, let

$$Y = \sum_{i=1}^{n^{3/4}} X_i.$$

Observe that $Y = |\{r \in R : r \le m\}|$ and hence

$$\mathcal{E}_1 = \left\{|\{r \in R : r \le m\}| < \frac{1}{2}n^{3/4} - \sqrt{n}\right\} = \left\{Y < \frac{1}{2}n^{3/4} - \sqrt{n}\right\}.$$

The random variable Y has binomial distribution, and its expectation and variance are given by

$$\mathbb{E}Y = \frac{1}{2}n^{3/4} \quad \text{and} \quad \mathrm{Var}(Y) = \frac{1}{4}n^{3/4}.$$

We shall apply the Chebyshev's inequality.

Theorem 2 (Chebyshev's Inequality)**.** *For any $a > 0$, it holds*

$$\Pr\left(|Y - \mathbb{E}Y| \ge a\right) \le \frac{\mathrm{Var}(Y)}{a^2}.$$

By Theorem 2, we have

$$\begin{aligned} \Pr(\mathcal{E}_1) &= \Pr\left(Y < \frac{1}{2}n^{3/4} - \sqrt{n}\right) \\ &= \Pr\left(Y < \mathbb{E}Y - \sqrt{n}\right) \\ &\le \Pr\left(|Y - \mathbb{E}Y| \ge \sqrt{n}\right) \\ &\le \frac{\mathrm{Var}(Y)}{n} \\ &\le \frac{1}{4}n^{-1/4}, \end{aligned}$$

as claimed. ∎

Lemma 3. $\Pr(\mathcal{E}_2) \le \frac{1}{4}n^{-1/4}$.

Proof. The proof is the same as Lemma 1. ∎

Lemma 4. $\Pr(\mathcal{E}_3) \le \frac{1}{2}n^{-1/4}$.

Proof. Let a be the $(n/2 - 2n^{3/4})$th smallest of S, and b be the $(n/2 + 2n^{3/4})$th smallest of S. Define two bad events:

$$\mathcal{F}_1 = \{\ell \le a\} = \left\{ |\{r \in R : r \le a\}| \ge \frac{1}{2}n^{3/4} - \sqrt{n} \right\};$$
$$\mathcal{F}_2 = \{u \ge b\} = \left\{ |\{r \in R : r \ge b\}| \ge \frac{1}{2}n^{3/4} - \sqrt{n} \right\}.$$

A key observation here is that $\mathcal{E}_3 \subseteq \mathcal{F}_1 \cup \mathcal{F}_2$, and hence an application of the union bound yields

$$\Pr(\mathcal{E}_3) \le \Pr(\mathcal{F}_1) + \Pr(\mathcal{F}_2).$$

Let us upper bound $\Pr(\mathcal{F}_2)$ as an example. For each i define

$$X_i = \begin{cases} 1, & \text{if } r_i \ge b \\ 0, & \text{o/w.} \end{cases}$$

Note that

$$\mathbb{E}X_i = \Pr(X_i = 1) = \Pr(r_i \ge b) = \frac{n/2 - 2n^{3/4}}{n} = \frac{1}{2} - \frac{2}{n^{1/4}}.$$

Further, let

$$Y = \sum_{i=1}^{n^{3/4}} X_i.$$

Observe that $Y = |\{r \in R : r \ge b\}|$ and hence

$$\mathcal{F}_2 = \left\{ Y \ge \frac{1}{2}n^{3/4} - \sqrt{n} \right\}.$$

The random variable Y has binomial distribution, and its expectation and variance are given by

$$\mathbb{E}Y = n^{3/4}\left(\frac{1}{2} - \frac{2}{n^{1/4}}\right) = \frac{1}{2}n^{3/4} - 2\sqrt{n}$$
$$\text{and} \quad \operatorname{Var}(Y) = n^{3/4}\left(\frac{1}{2} - \frac{2}{n^{1/4}}\right)\left(\frac{1}{2} + \frac{2}{n^{1/4}}\right) \le \frac{1}{4}n^{3/4}.$$

By Theorem 2, we have

$$\begin{aligned} \Pr(\mathcal{F}_2) &= \Pr\left(Y \ge \frac{1}{2}n^{3/4} - \sqrt{n}\right) \\ &= \Pr\left(Y \ge \mathbb{E}Y + \sqrt{n}\right) \\ &\le \Pr\left(|Y - \mathbb{E}Y| \ge \sqrt{n}\right) \\ &\le \frac{\operatorname{Var}(Y)}{n} \\ &\le \frac{1}{4}n^{-1/4}. \end{aligned}$$

Similarly, $\Pr(\mathcal{F}_1) \le \frac{1}{4}n^{-1/4}$ and by the union bound we get $\Pr(\mathcal{E}_3) \le \frac{1}{2}n^{-1/4}$. $\square$