Median and Order Statistics

Andres Mendez-Vazquez

September 30, 2018

Contents

1 Introduction

One of the most important tasks when analyzing a collection of numbers is to find the ith smallest element. Examples of these elements are

- The minimum, i=1.
- The maximum, $i = n$.
- The Median for n odd, $i = \frac{n+1}{2}$.

Then, it is necessary to find fast solutions to find this statistics.

2 Selection Problem

There is a clever way to see the finding of the ith statistics as a selection problem (Look at the slides).

- Input: A set A of n (distinct) numbers and an integer i, with $1 \leq i \leq n$.
- Output: The element $x \in A$ that is larger than exactly $i-1$ other elements of A.

Here, a classic way to solve the problem is to sort the elements with merge sort. Then, find the statistics. However, we can do much better.

3 On the way to a better solution

In order to find the minimum alone, using simple comparisons in a sequence of n numbers, the naive algorithm can find that element in n. However, if we try to find the maximum and minimum at the same time, we ca do much better:

- Take two elements at the same time.
- Compare them between them to get the min and max in the tuple.
- Compare the smallest with the actual minimum do the same with the biggest one.

This algorithm takes 3 comparisons per sets of two elements the we are bounded by $3\left\lfloor \frac{n}{2} \right\rfloor$. For example, if n is even we have one initial comparison followed by

$$3\left(\frac{n-2}{2}\right)+1$$

This is giving us that the total number of comparisons is at most$3\left\lfloor \frac{n}{2} \right\rfloor$

4 Selection in expected linear time

The analysis of the Randomized-selection can be done assuming the following:

- $X_k = I\{\text{the subarray A[p .. q] has exactly } k \text{elements}\}$ with $E[X_k] = \frac{1}{n}$ (Assuming that the elements are distinct)

Now, we need to bound the recursive function $T(n)$ describing the Randomized-select algorithm. This can be done, if we assume that

- The ith element is always in the largest partition size.

Therefore,

$$
\begin{aligned}
T(n) &\leq \sum_{k=1}^{n} X_k \times (T(\max(k-1, n-k)) + O(n)) \\
&= \sum_{k=1}^{n} X_k \times (T(\max(k-1, n-k)) + O(n)
\end{aligned}
$$

Then, we take the expected value:

$$
\begin{aligned}
E[T(n)] &\leq E\left[\sum_{k=1}^{n} X_k \times (T(\max(k-1, n-k)) + O(n)\right] \\
&= \sum_{k=1}^{n} E\left[X_k \times (T(\max(k-1, n-k))\right] + O(n) \\
&= \sum_{k=1}^{n} E\left[X_k\right] E\left[T(\max(k-1, n-k))\right] + O(n) \\
&= \sum_{k=1}^{n} \frac{1}{n} \times E\left[T(\max(k-1, n-k))\right] + O(n)
\end{aligned}
$$

Now if we take in account that

$$
max(k-1, n-k) = \begin{cases} k-1 & \text{if } k > \left\lceil \frac{n}{2} \right\rceil \\ n-k & \text{if } k \leq \left\lceil \frac{n}{2} \right\rceil \end{cases}
$$

This means that

- if n is even each term $T\left(\left\lceil \frac{n}{2} \right\rceil\right)$ to $T(n-1)$ appears exactly twice.
- if n is odd each term appears twice and $T\left(\left\lfloor \frac{n}{2} \right\rfloor\right)$ appears once.

Then, we have the following recursion:

$$E\left[T\left(n\right)\right] \leq \frac{2}{n}\sum_{k=\left\lfloor \frac{n}{2}\right\rfloor}^{n-1} E\left[T\left(k\right)\right] + O(n).$$

Thus, if we assume that $E\left[T(n)\right] \leq cn$

$$\begin{aligned}
E\left[T(n)\right] &\leq \frac{2}{n}\sum_{k=\left\lfloor \frac{n}{2}\right\rfloor}^{n-1} ck + an \\
&= \frac{2c}{n}\left(\sum_{k=1}^{n-1} k - \sum_{k=1}^{\left\lfloor \frac{n}{2}\right\rfloor -1} k\right) + an \\
&= \frac{2c}{n}\left(\frac{(n-1)\,n}{2} - \frac{\left(\left(\left\lfloor \frac{n}{2}\right\rfloor - 1\right)\left\lfloor \frac{n}{2}\right\rfloor\right)}{2}\right) + an \\
&\leq \frac{2c}{n}\left(\frac{(n-1)\,n}{2} - \frac{\left(\left(\frac{n}{2} - 1\right)\frac{n}{2}\right)}{2}\right) + an \\
&= c\left(\frac{3n}{4} + \frac{1}{2} - \frac{2}{n}\right) + an \\
&\leq \frac{3cn}{4} + \frac{c}{2} + an \\
&= cn - \left(\frac{cn}{4} - \frac{c}{2} - an\right)
\end{aligned}$$

Now, we need that $\frac{cn}{4} - \frac{c}{2} - an \geq 0$ or $n\left(\frac{c}{4} - a\right) \geq \frac{c}{2} > 0$. Thus, $\frac{c}{4} - a > 0$ i.e. w choose $c > 4a$. Therefore, $n \geq \frac{\frac{c}{2}}{\frac{c}{4} - a} = \frac{2c}{c-4a}$. Thus, if we assume that $T(n) = O(1)$, we have that $E\left[T(n)\right] = O(n)$.

5 Analysis of the Worst Case Selection

For this, we look at the following image

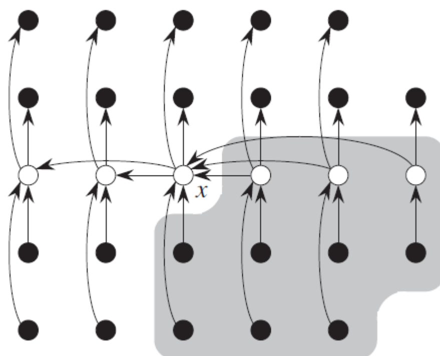

Figure 1: Look at this

1. At least half of the medians are greater than or equal to the median-of-medians x.
2. At least half of the $\left\lceil \frac{n}{5} \right\rceil$ groups contribute with at least 3 elements greater than x, except for two groups: one that has less than 5 elements and the one containing x.
3. Then, the number of elements greater than x is at least $3\left(\left\lceil \frac{1}{2}\left\lceil \frac{n}{2} \right\rceil \right\rceil - 2\right) \geq \frac{3n}{10} - 6$.
4. Similarly, at least $\frac{3n}{10} - 6$ elements are less than x.
5. We can then in the worst case we have select is called recursively in at most $\frac{3n}{10} - 6 + \frac{3n}{10} - 6 < \frac{7n}{10} + 6$.
6. Steps 1, 2 and 4 take $O(n)$ time.

Then, if we assume that

- $a < b \Rightarrow T(a) < T(b)$ (Monotonically increasing).

Thus, we have the following recurrence:

$$T(n) = \begin{cases} O(1) & \text{if } n < 140 \\ T\left(\left\lceil \frac{n}{5} \right\rceil\right) + T\left(\frac{7n}{10} + 6\right) + O(n) & \text{if } n \geq 140 \end{cases}.$$