Kleinberg-Tardos 13.5

Randomized divide & conquer: median finding and Quicksort

Given $S = \{a_1, a_2, \dots, a_n\}$ a set of distinct numbers

Definition of median of S

- when $n = 2k+1$ for some integer k then the median of S is the $k+1$'st element in the sorted order

  $a_{i_1} < a_{i_2} < \dots a_{i_k} < \underline{a_{i_{k+1}}} < a_{i_{k+2}} < \dots < a_{i_n}$ of S

- when $n = 2k$ is even we define the k'th element a_{i_k} to be the median

> **Median Problem**
> Given a set $S = \{a_1, a_2, \dots a_n\}$ consisting of n distinct numbers
> Find the median of S

Easy if we may use sorting. This requires $\Omega(n \log n)$ (as you will see in DM553)

Can we do it faster?

Recall the pivoting idea in quicksort (QS)

- Chose an element $a \in S$ (in the standard QS this is the last element of S)
- construct the following partition of S

$$\underbrace{\boxed{<a}}_{S^-} \quad a \quad \underbrace{\boxed{>a}}_{S^+}$$

- Sort S^- and S^+ recursively

Try the same idea for the median (call it m)

assume $n = 2k+1$

$$\underbrace{\boxed{<a}}_{S^-} \quad a \quad \underbrace{\boxed{>a}}_{S^+}$$

if $|S| = k$ then a is the median m

if $|S^-| > k$ then m is the $k+1$'st element of S^-

of $|S^-| < k$ then m is the $k+1-(1+|S|)$'st element of $|S^+|$

<u>Conclusion</u>: we need to solve the more general

Select(S,k)

Given $S=\{a_1,a_2 \dots a_n\}$ all distinct numbers and

$k \in \{1,2 \dots n\}$

Find the element a_{i_k} (k'th the element in the sorted order of S.)

So median of S is the return of Select$(S,\frac{n}{2})$ when n is even and Select$(S,\frac{n+1}{2})$ when n is odd

<u>Select(S,k)</u>:

Choose a splitter $a_i \in S$

$S^- := \{a_j \in S \mid a_j < a_i\}$;

$S^+ := \{a_j \in S \mid a_j > a_i\}$

If $|S^-| = k-1$ then return a_i

Else if $|S^-| \geq k$ then return Select(S^-,k)

Else return Select$(S^+, k-(1+|S^-|))$

Clearly Select(S,k) returns the k'th element of S

what do we want from a splitter?

- we always make at most one recursive call (to either S^- or S^+). So we reduce the problem size by the size of the discarded set each time
- Ideal situation $|S^-| \sim |S^+|$ which would (almost) half the problem size in each step
- Suppose we can guarantee $\min\{|S^-|, |S^+|\} \geq \varepsilon n$ for some $0 < \varepsilon < 1$. Then the running time $T(n)$ on input of size n satisfies

$$T(n) \leq T((1-\varepsilon)n) + cn$$

work besides recursive calls

- $T(n) \leq cn + (1-\varepsilon)cn + (1-\varepsilon)^2 cn + \ldots$

$$= cn\left[1 + (1-\varepsilon) + (1-\varepsilon)^2 + \ldots\right]$$

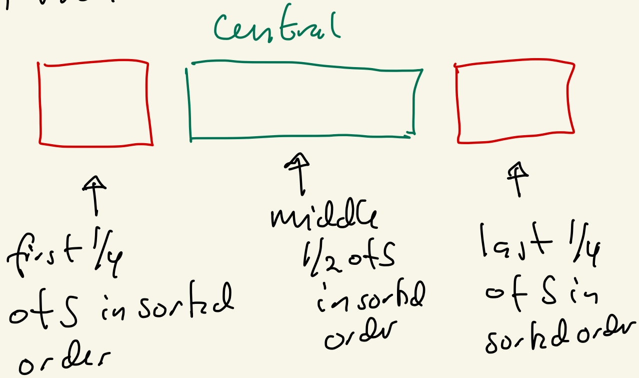
$$< cn \sum_{i=0}^{\infty} (1-\varepsilon)^i = cn \frac{1}{1-(1-\varepsilon)} = \frac{cn}{\varepsilon} = \frac{c}{\varepsilon} \cdot n$$

This is linear time!

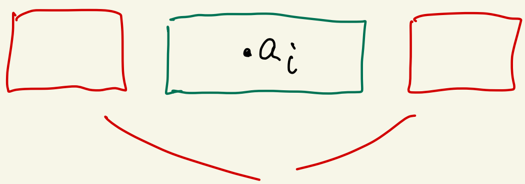
RandomSelect(S,k): choose the splitter a_i uniformly at random from S

RandomSelect(S,k): choose the splitter a_i uniformly at random from S

We expect to reduce $|S|$ by a constant fraction in each round



- Half of the elements of S are central (in the middle half
- If a_i is central we eliminate $\geq 1/4$ of the elements (central = belongs to middle set)



one of these is eliminated

Phases

phase j: actual set S' satisfies $\left(\frac{3}{4}\right)^{j+1} \cdot n < |S'| \le \left(\frac{3}{4}\right)^{j} \cdot n$

We start in phase $j=0$

If we choose a central splitter while in phase j the phase ends as $S' \to S''$ with

$|S''| \le \frac{3}{4}|S'| \le \left(\frac{3}{4}\right)^{j+1} \cdot n$ as $|S'| \le \left(\frac{3}{4}\right)^{j} \cdot n$

(note that we could skip several phases and go to phase $j+s$ for some $s \ge 1$)

The probability that we choose a central splitter is $\frac{1}{2}$ so the expected number of splitters we pick in phase j is $\frac{1}{1/2} = 2$ (if we ever arrived in that phase)

Define random variables $X_0, X_1, \dots X_i, \dots$ so that X_i is the #steps done by the algorithm in phase i. Then

$X = X_0 + X_1 + X_2 + \dots$ is the #steps in the algorithm

In phase j the input S_j has size at most $\left(\frac{3}{4}\right)^{j} \cdot n$ so at most $c \cdot \left(\frac{3}{4}\right)^{j} n$ steps per iteration (recursive call) in phase j

We expect (at most) 2 iterations in phase j so

$$E(X_j) \leq 2cn\left(\tfrac{3}{4}\right)^j$$

Hence

$$\begin{aligned}
E(X) &= E\left(\sum_j X_j\right) \\
&= \sum_j E(X_j) \\
&\leq 2cn \sum_j \left(\tfrac{3}{4}\right)^j \\
&\leq 2cn \sum_{k=0}^{\infty} \left(\tfrac{3}{4}\right)^k = 2cn\left(\frac{1}{1-\frac{3}{4}}\right) = 8cn
\end{aligned}$$

Conclusion: the expected running time of RandomSelect(n,k) is $O(n)$

RandQS(S) = quicksort on S when pivot is a random element from S

Same analysis as above

- small change: if the chosen a_i is <u>not</u> central then pick a new random a_i

(✱) The expected # of repetitions before we have a central splitter is 2 (as probability of good splitter = 1/2)

This gives us

(13.19) The expected running time on S excluding recursive calls is $O(n)$

Divide subproblems into types

Type j : the actual set S' satisfies $n\cdot\left(\frac{3}{4}\right)^{j+1} < |S'| \le n\cdot\left(\frac{3}{4}\right)^{j}$

By (✱) the expected work on S' minus recursive calls is $O\left(n\left(\frac{3}{4}\right)^{j}\right)$ (□)

$\boxed{\text{We need to bound \#subproblems of type } j}$

<u>Observation</u>: all subproblems of type j are <u>disjoint</u>!

since they came from subproblems of types j' when j' < j

As subproblems of type j are disjoint and each have size at least $\left(\frac{3}{4}\right)^{j+1} \cdot n$, there are at most

$$\frac{n}{n \cdot \left(\frac{3}{4}\right)^{j+1}} = \left(\frac{4}{3}\right)^{j+1} \text{ subproblems of type } j$$

By ($\square$) the expected time spend on each subproblem of type j is $O\left(n\left(\frac{3}{4}\right)^{j}\right)$

There are at most $\left(\frac{4}{3}\right)^{j+1}$ subproblems of type j so expected time on all of these is

$$O\left(n \cdot \left(\frac{3}{4}\right)^{j} \cdot \left(\frac{4}{3}\right)^{j+1}\right) = O\left(\frac{4}{3}n\right) = O(n)$$

There are at most $\log_{\frac{4}{3}} n = O(\log n)$ different types so

(13.21) The expected running time for RandQS on a set of n distinct numbers is $O(n \log n)$

Cormen 7.4.2

Let X be the number of comparisons made by RandQS on a set S of n distinct numbers

Then X is a random variable

Let $z_1 < z_2 < \cdots < z_n$ be the sorted order of the elements in S and let the random variable $X_{ij} = 1$ if z_i and z_j are compared in the algorithm and $X_{ij} = 0$ otherwise

Note: z_i and z_j can only be compared if one of them is the pivot (all comparisons are with the pivot)

So $X = \sum_{1 \le i < j \le n} X_{ij}$

Let $Z_{ij} = \{z_i, z_{i+1}, \ldots, z_j\}$

Then z_i and z_j are compared precisely if z_i or z_j is the first pivot chosen from Z_{ij}

The probability of this is $\frac{2}{|Z_{ij}|} = \frac{2}{j-i+1}$

Now we see that $p(X_{ij}=1) = \frac{2}{j-i+1}$

So $E(X_{ij}) = \frac{2}{j-i+1}$ and hence

$$E(X) = E\left[\sum_{i=1}^{n-1}\sum_{j=i}^{n} X_{ij}\right]$$

$$= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} E(X_{ij})$$

$$= \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \frac{2}{j-i+1}$$ take $k = j-i$

$$= \sum_{i=1}^{n-1}\sum_{k=1}^{n-i} \frac{2}{k+1}$$

$$< 2\sum_{i=1}^{n-1}\sum_{k=1}^{n} \frac{1}{k}$$

$$< 2\sum_{i=1}^{n-1} H(n)$$

$$< 2n\,H(n) = O(n\log n)$$

Conclusion: The expected # comparisons in RandQS is $O(n\log n)$