Order Statistics

Given: $A[1 \dots n]$ of n distinct numbers and i such that $1 \leq i \leq n$,

find $x \in A$ that is larger than exactly $i-1$ other elements in A.

Obvious solution: Sort A in $O(n \log n)$ time then output $A[i]$

But, we can do better.

Note: we define median as corresponding to $i = \left\lfloor \frac{n+1}{2} \right\rfloor$

n odd:

median

$n = 7$

$$\left\lfloor \frac{n+1}{2} \right\rfloor$$

median

n even:

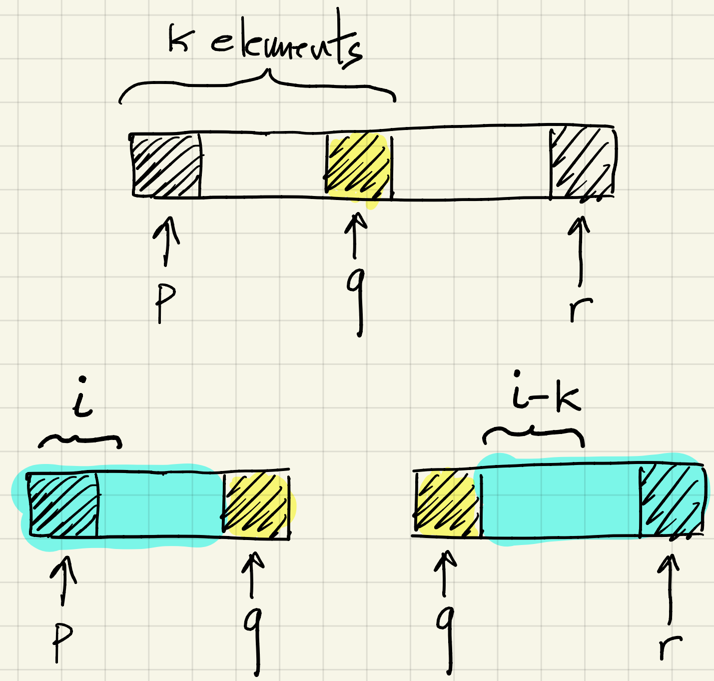
$n = 8$

$$\left\lfloor \frac{n+1}{2} \right\rfloor$$

A divide-and-conquer randomized algorithm.

Randomized-select (A, p, r, i)

    if p = r
        then return A[p]
    q ← randomized-partition (A, p, r)
    k ← q − p + 1   ▷ rank of pivot in A[p...r]
    if i = k
        then return A[q]
    else if i < k
        then return randomized-select (A, p, q−1, i)
    else return randomized-select (A, q+1, r, i−k)



Analysis: Similar to Randomized Quicksort

$$T(n) \leqslant \frac{1}{n} \sum_{k=0}^{n-1} T(\max(k, n-k-1)) + \Theta(n)$$

Assuming all partitions equally likely

$$T(n) \leqslant \frac{2}{n} \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n-1} T(k) + \Theta(n) \qquad \text{Guess } T(n) \leq dn$$

(Really $T(n)$ here is $E[T(n)]$)

$$T(n) \leqslant \frac{2}{n} \sum_{k=\lfloor \frac{n}{2} \rfloor}^{n-1} dk + cn = \frac{2d}{n}\left[\frac{n(n-1)}{2} - \frac{\lfloor \frac{n}{2} \rfloor (\lfloor \frac{n}{2} \rfloor - 1)}{2}\right] + cn$$

$$\leqslant d\left[(n-1) - \frac{(\frac{n}{2}-2)(\frac{n}{2}-1)}{n}\right] + cn \leqslant \frac{3dn}{4} + \frac{d}{2} + cn$$

$$= dn - \left(\frac{dn}{4} - \frac{d}{2} - cn\right)$$ [choose d to make blue $\geqslant 0$]

$n = 11$

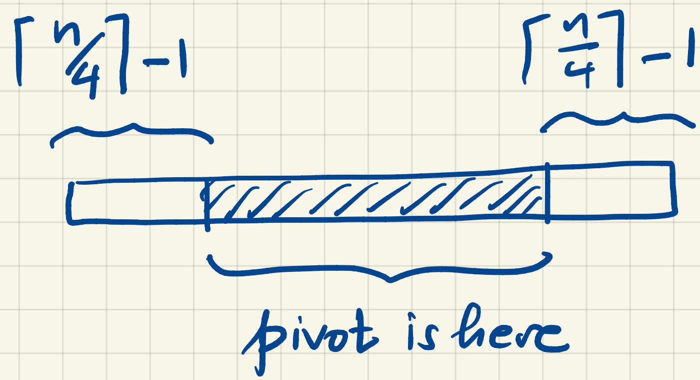
0	10	
1	9	
2	8	
3	7	
4	6	
5	5	$\lfloor \frac{n}{2} \rfloor$
6	4	
7	3	
8	2	
9	1	
10	0	

$n = 10$

0	9	
1	8	
2	7	
3	6	
4	5	$\lfloor \frac{n}{2} \rfloor$
5	4	
6	3	
7	2	
8	1	
9	0	

Another way:

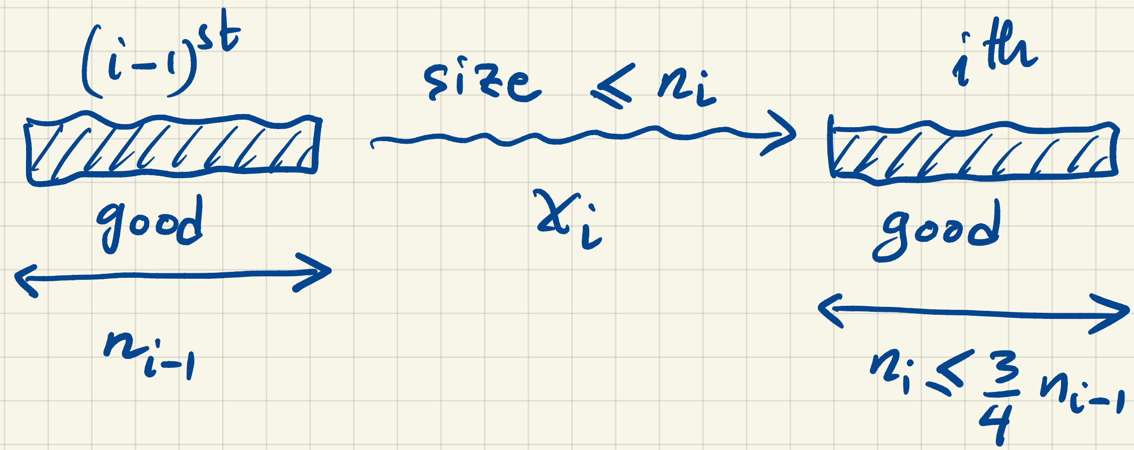
What is the probability that the largest partition $\leq \frac{3}{4}n$?



probability of "good" is : $\dfrac{n - 2\left(\lceil \frac{n}{4} \rceil - 1\right)}{n}$

$\lceil \frac{n}{4} \rceil - 1 \leq \frac{n}{4}$, so probability above is $\geq \dfrac{n - 2n/4}{n} = \dfrac{1}{2}$

Let x_i be the # times we partition from the $(i-1)^{st}$ good partition until the i^{th} good partition



$$\text{Then running time} \leq c \sum_{i=0}^{?} x_i \left(\frac{3}{4}\right)^i n$$

$$E[x_i] \leq 2 \Rightarrow \quad \leq 2cn \sum_{i=0}^{\infty} \left(\frac{3}{4}\right)^i = 8cn$$

We can do even better!

A deterministic algorithm that runs in $\Theta(n)$

Idea: Guarantee a good split.

Find a good pivot

1) Divide elements into groups of 5: get $\lceil \frac{n}{5} \rceil$ groups
2) Find median of each group in $O(1)$
   - Run insertion sort on each group, then find median
3) Find the median x of the $\lceil \frac{n}{5} \rceil$ medians (Recursively)

Partition

4) Partion elements using x as Pivot $O(n)$

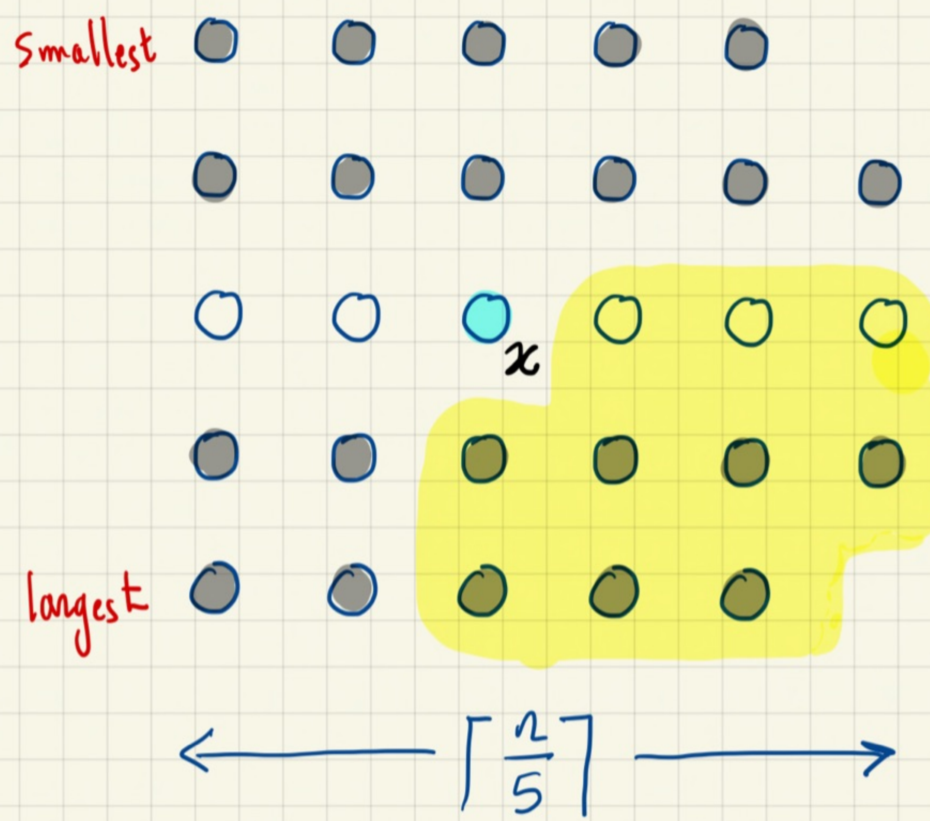
$k-1$	x	$n-k$

As before

5) Continue recursively as usual using either left or right
   - $i=k$: return x
   - $i<k$: recourse left with i
   - $i>k$: recourse right with $i-k$

$$T(n) = T\left(\left\lceil \frac{n}{5} \right\rceil\right) + T\left(\max(k-1,\ n-k)\right) + \Theta(n)$$

We can find an upper bound on $\underbrace{(k-1)}_{\#\,\text{elem.} < x}$ and $\underbrace{(n-k)}_{\#\,\text{elem.} > x}$



$$\geq 3\left(\left\lceil \frac{1}{2}\left\lceil \frac{n}{5} \right\rceil \right\rceil - 2\right)$$

$$\geq \frac{3n}{10} - 6$$

How many elements $> x$?

Similarly we can show at least $\frac{3n}{10} - 6$ elements $< x$

If # elem. $> x$ is at least $\frac{3n}{10} - 6$

then # elem $< x$ is at most

$$n - \left(\frac{3n}{10} - 6 \right) = \frac{7n}{10} + 6$$

same claim for # elem $> x$

$$T(n) \leq \begin{cases} T(\lceil \frac{n}{5} \rceil) + T(\frac{7n}{10} + 6) + \Theta(n) & n \geq n_0 \\ \Theta(1) & \text{otherwise} \end{cases}$$

Guess $T(n) \leq dn$

$$\begin{aligned} T(n) &\leq d\lceil n/5 \rceil + d(7n/10 + 6) + cn \\ &\leq dn/5 + d + 7dn/10 + 6d + cn \\ &= \frac{9dn}{10} + 7d + cn \\ &= dn - \frac{dn}{10} + 7d + cn \end{aligned}$$

We need $d \geq \dfrac{10cn}{n-70} = 20c$ if $n_0 = 140$.

choose $d \geq 20c$

What is the mystery behind groups of 5?

Why not groups of k for instance for some k?

We need $\frac{n}{k} + n - \lceil \frac{k}{2} \rceil \times \frac{1}{2} \times \frac{n}{k} < n$

$\underbrace{\lceil \tfrac{k}{2} \rceil}_{\#\text{ elem in group}/2}$ $\quad \underbrace{\tfrac{n}{k}}_{\#\text{groups}}$

e.g $k=3$ does not work.

Gives $T(n) = T\left(\frac{n}{3}\right) + T\left(\frac{2n}{3}\right) + \Theta(n)$

$T(n) = \Theta(n \log n)$

What about Quicksort with a partition around the median (i.e. pivot is median)

- Theoretically, it's $\Theta(n\log n)$
- Practically, not better than randomized Quicksort (constant term of above alg. in O notation is rather large.