CMPT307: Order Statistics

Week 6-1

Xian Qiu

Simon Fraser University

xianq@sfu.ca

Order Statistics

ith order statistic $=$ ith smallest element

minimum/maximum is the first/last order statistic

median is the "halfway point" of the set, *i.e.*, ith order with

- (lower) median: $i = \lfloor (n+1)/2 \rfloor$
- upper median: $i = \lceil (n+1)/2 \rceil$

Selection Problem

- input: a set A of n distinct numbers and index i, with $1 \leq i \leq n$
- output: ith order statistic of A

Min and Max

minimum (or maximum): $n-1$ comparisons

MINIMUM(A)

```
1 min = A[1];
2 for i = 2 to n do
3     if min > A[i] then
4         min = A[i];
```

simultaneous minimum and maximum

- ▷ run MINIMUM and MAXIMUM independently, performing $2(n-1)$ comparisons in total
- ▷ can we do better?

Min and Max

- record (current) $\min$ and $\max$ at the same time
- observation: for a_i and a_j with $a_i < a_j$, only need to compare a_i with $\min$ and a_j with $\max$ respectively

MIN-AND-MAX

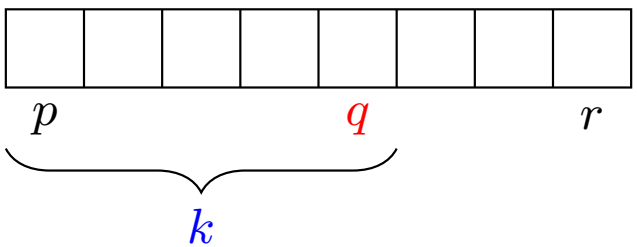
```
1 divide elements into ⌈n/2⌉ pairs: {a_{2i-1}, a_{2i}} for i = 1, ..., ⌈n/2⌉;
2 for i = 1 to ⌈n/2⌉ do
3     compare min with minimal of {a_{2i-1}, a_{2i}};
4     compare max with maximal of {a_{2i-1}, a_{2i}};
5 return min, max
```

- total #comparisons $\leq 3\lceil \frac{n}{2} \rceil$

Selection Problem

find the ith order statistic of $A[p..r]$

divide and conquer: divide by RANDOMIZED-PARTITION



- $A[q]$ is the $k := (q - p + 1)$th order statistic
- if $i = k$, done; else find such q recursively
- $i < k$: find the ith statistic of $A[p..q - 1]$;
- $i > k$: find the $(i - k)$th statistic of $A[q + 1..r]$

Pseudocode

RANDOMIZED-SELECT(A, p, r, i)

```
1  if p == r then
2      return A[p];
3  q = RANDOMIZED-PARTITION(A, p, r);
4  k = q − p + 1;
5  if i == k then
6      return A[q];
7  else if i < k then
8      return RANDOMIZED-SELECT(A, p, q − 1, i);
9  else
10     return RANDOMIZED-SELECT(A, q + 1, r, i − k);
```

Average-case Analysis

$$X_k = \mathrm{I}\{A[p..q] \text{ has exactly } k \text{ elements}\}, \quad \forall 1 \le k \le n \quad \Rightarrow \mathbb{E}[X_k] = \frac{1}{n}$$

$$T(n) \le \sum_{k=1}^{n} X_k[T(\max\{k-1, n-k\}) + O(n)]$$

$$\mathbb{E}[T(n)] \le \sum_{k=1}^{n} \mathbb{E}[X_k] \cdot \mathbb{E}[T(\max\{k-1, n-k\})] + O(n)$$

$$= \sum_{k=1}^{n} \frac{1}{n} \mathbb{E}[T(\max\{k-1, n-k\})] + O(n)$$

$$\max\{k-1, n-k\} = \begin{cases} k-1, & \text{if } k > \lceil n/2 \rceil \\ n-k, & \text{if } k \le \lceil n/2 \rceil \end{cases}$$

$$\mathbb{E}[T(n)] \le \frac{2}{n} \sum_{k=\lfloor n/2 \rfloor}^{n-1} \mathbb{E}[T(k)] + O(n)$$

Average-case Analysis

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

show $\mathbb{E}[T(n)] = O(n)$ by induction, *i.e.* prove $\mathbb{E}[T(n)] \leq cn$

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

Worst-case

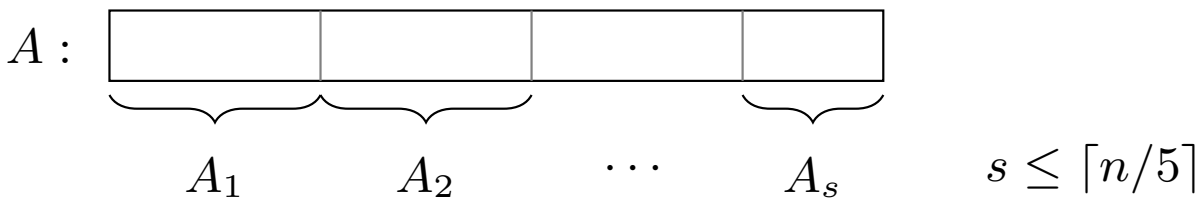
- what is the worst case running time?

$$T(n) = T(n-1) + \Theta(n)$$

- to achieve $O(n)$ in worst-case, want "equal" size partition
- choose median as pivot in $\textsc{Modified-Partition}(A, p, r, x)$
  but how to find the median?
- $x =$ approximate median and let
  $q = \textsc{Modified-Partition}(A, p, r, x) \quad k = p - q + 1$
- if $k = i$, done
- $i < k$: find the ith statistic of $A[p..q-1]$;
- $i > k$: find the $(i-k)$th statistic of $A[q+1..r]$

Approximating the Median

- divide A into subarrays of size 5 each (expect for the last one)



- sort A_i and find the median x_i for all i
- apply the same procedure for $A := \{x_1, \ldots, x_s\}$ until $|A| \leq 5$

Pseudocode

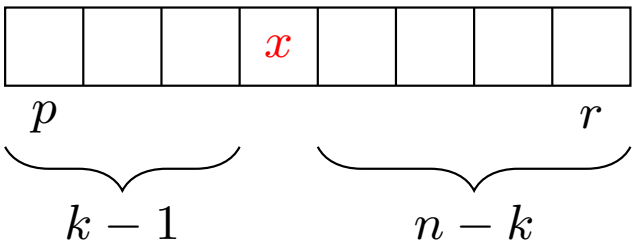
SELECT(A, p, r, i)

```
1  divide A into s = ⌈A.length/5⌉ groups A_1, ..., A_s;
2  for j = 1 to s do
3      sort A_j by insertion sort;
4      B[j] = (lower) median of A_j;
5  x = SELECT(B, 1, s, ⌊s/2⌋) ;                    // find median
6  q = MODIFIED-PARTITION(A, p, r, x);             // use x as pivot
7  k = q − p + 1;                                  // kth order
8  if i == k then
9      return x;
10 else if i < k then
11     return SELECT(A, p, q − 1, i);
12 else
13     return SELECT(A, q + 1, r, i − k);
```

$$T(n) = T(\lceil n/5 \rceil) + \max\{T(k-1), T(n-k)\} + O(n)$$

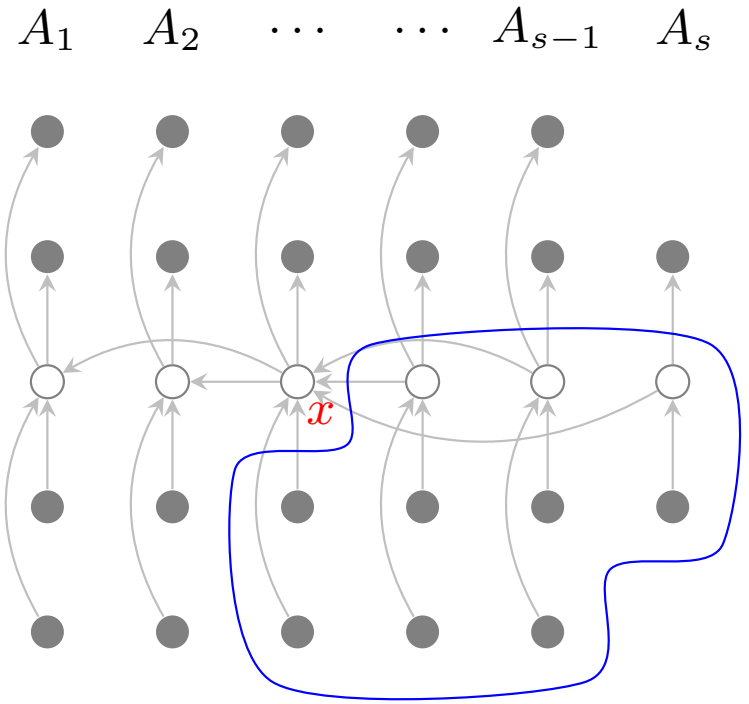
Analysis

$$T(n) = T(\lceil n/5 \rceil) + \max\{T(k-1), T(n-k)\} + O(n)$$



$n - k =$?

$\lceil \frac{s}{2} \rceil$ groups $\geq x$



Analysis

$\lceil \frac{s}{2} \rceil - 2$ groups contributes 3 elements ($\geq x$)

$$n - k \geq 3\left(\left\lceil \frac{1}{2}\left\lceil \frac{n}{5} \right\rceil \right\rceil - 2\right) \geq \frac{3n}{10} - 6$$

$$k - 1 \leq \frac{7n}{10} + 5$$

similarly, $\lceil \frac{s}{2} \rceil - 2$ groups contributes 3 elements ($\leq x$)

$$k - 1 \geq \frac{3n}{10} - 6 \quad \Rightarrow n - k \leq \frac{7n}{10} + 5$$

summarizing, $\max\{n-1, n-k\} \leq \frac{7n}{10} + 5$

$$T(n) \leq T(\lceil n/5 \rceil) + T(7n/10 + 5) + O(n)$$
$$= O(n) \qquad \text{by substitution method}$$