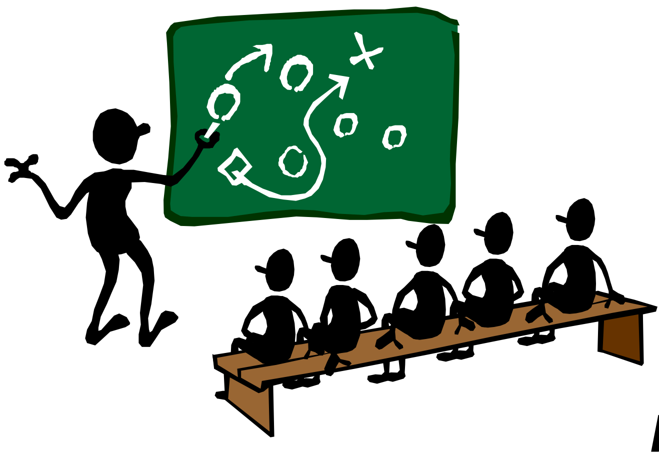
Algorithms – Chapter 9
Medians and Order Statistics



Juinn-Dar Huang

Professor

jdhuang@mail.nctu.edu.tw

August 2007

Rev. '08, '11, '12, '15, '16, '18, '19, '20, '21

Medians and Order Statistics

- The ith order statistic of a set of n elements is the ith smallest element
  - minimum: first order statistic
  - maximum: n-th order statistic
  - median: $(n+1)/2$-th order statistic
- The selection problem can be specified formally as follows
  - Input: A set of n (distinct) numbers and a number i, with $1 \le i \le n$
  - Output: the element $x \in A$ that is larger than exactly $i - 1$ other elements of A

Juinn-Dar Huang jdhuang@mail.nctu.edu.tw

Juinn-Dar Huang
jdhuang@mail.nctu.edu.tw

Find Minimum Or Maximum

```
MINIMUM(A)
min ← A[1]
for i ← 2 to length[A]
    do if min > A[i]
        then min ← A[i]        Expected number?
return min
```

Time complexity:

need n – 1 comparisons ➔ Θ(n)

It's optimal !

Find Minimum AND Maximum

- Find minimum and maximum simultaneously
  - run MINIMUM and MAXIMUM separately
  - need 2(n – 1) comparisons ➔ $\Theta(n)$
- Better way
  - compare a pair of input elements at a time first
  - compare the current maximum with the larger one
  - compare the current minimum with the smaller one
  - 3 comparisons for 2 elements ➔ $3\lfloor n/2 \rfloor$ ➔ $\Theta(n)$

Juinn-Dar Huang
jdhuang@mail.nctu.edu.tw

Find the i-th Order Statistic

- Think about Randomized Quick Sort…

RANDOMIZED_SELECT(A, p, r, i)

1 **if** $p = r$

2 **then return** A[p]

3 $q \leftarrow RANDOMIZED\_PARTITION(A, p, r)$ ▶ defined in Quick Sort

4 $k \leftarrow q - p + 1$

5 **if** $i = k$ ▶ the pivot value is the answer

6 **then return** A[q]

7 **elseif** $i < k$

8 **then return** RANDOMIZED_SELECT($A, p, q-1, i$) **in left part**

9 **else return** RANDOMIZED_SELECT($A, q+1, r, i-k$) **in right part**

Time Complexity (1/4)

- Worst case
  - $\Theta(n^2)$; same scenario as in Quick Sort
  - in case that always the biggest elements are chosen as pivots (though it's very unlikely to happen)
- Expected runtime
  - for k = 1, 2, …, n, we define indicator random variables X_k = I {the subarray A[$p..q$] has exactly k elements} ,
  - assume all elements are distinct ➔ $E[X_k] = 1/n$

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

choose the larger partition

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

Medians & Order Statistics

Juinn-Dar Huang jdhuang@mail.nctu.edu.tw

Time Complexity (2/4)

$$E[T(n)]$$

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

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

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

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

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

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

Juinn-Dar Huang jdhuang@mail.nctu.edu.tw

Medians & Order Statistics

Time Complexity (3/4)

Solve the recurrence:
➔ **assume E[T(n)] ≤ cn**

$$E[T(n)] \le \frac{2}{n}\sum_{k=\lfloor n/2 \rfloor}^{n-1} ck + an = \frac{2c}{n}\left(\sum_{k=1}^{n-1} k - \sum_{k=1}^{\lfloor n/2 \rfloor - 1} k\right) + an$$

$$= \frac{2c}{n}\left(\frac{(n-1)n}{2} - \frac{(\lfloor n/2 \rfloor - 1)\lfloor n/2 \rfloor}{2}\right) + an$$

$$\le \frac{2c}{n}\left(\frac{(n-1)n}{2} - \frac{(n/2-2)(n/2-1)}{2}\right) + an$$

$$= \frac{2c}{n}\left(\frac{n^2-n}{2} - \frac{n^2/4 - 3n/2 + 2}{2}\right) + an$$

$$= \frac{c}{n}\left(\frac{3n^2}{4} + \frac{n}{2} - 2\right) + an = c\left(\frac{3n}{4} + \frac{1}{2} - \frac{2}{n}\right) + an$$

$$\le \frac{3cn}{4} + \frac{c}{2} + an$$

$$= cn - \boxed{\left(\frac{cn}{4} - \frac{c}{2} - an\right) \quad \text{must be} \ge 0}$$

Juinn-Dar Huang jdhuang@mail.nctu.edu.tw

Time Complexity (4/4)

- Must: $cn/4 - an \geq c/2$
  - choose the constant c ➔ $c/4 - a > 0$ ➔ **$c > 4a$**
  - divide both sides by $c/4 - a$

$$n \geq \frac{c/2}{c/4 - a} = \frac{2c}{c - 4a}$$

➔ $E[T(n)] = O(n)$

- Any order statistic, including the median, can be determined in linear time on average

Juinn-Dar Huang jdhuang@mail.nctu.edu.tw