Medians and Order Statistics

CLRS Chapter 9

What Are Order Statistics?
The \( k \)-th order statistic is the \( k \)-th smallest element of an array.

\[
\begin{array}{cccccccc}
3 & 4 & 13 & 14 & 23 & 27 & 41 & 54 & 65 & 75
\end{array}
\]

8th order statistic

The lower median is the \( \left\lfloor \frac{n}{2} \right\rfloor \)-th order statistic

The upper median is the \( \left\lceil \frac{n}{2} \right\rceil \)-th order statistic

If \( n \) is odd, lower and upper median are the same

What are Order Statistics?

Selecting \( i \)-th ranked item from a collection.

- First: \( i = 1 \)
- Last: \( i = n \)
- Median(s): \( i = \left\lfloor \frac{n}{2} \right\rfloor \) or \( \left\lceil \frac{n}{2} \right\rceil \)

Order Statistics Overview

- Assume collection is unordered, otherwise trivial.
  \[ \text{find } i \text{-th order stat} = A[i] \]
- Can sort first \(- \Theta(n \lg n)\), but can do better \(- \Theta(n)\).
- I can find max and min in \( \Theta(n) \) time (obvious)
- Can we find any order statistic in linear time? (not obvious!)

Using the Pivot Idea

- Randomized-Select\((A[p..r],i)\) looking for \( i \)-th o.s.
  \[
  \begin{align*}
  &\text{if } p = r \\
  &\quad \text{return } A[p] \\
  &q \leftarrow \text{Randomized-Partition}\(A,p,r\) \\
  &k \leftarrow q \cdot p + 1 \quad \text{the size of the left partition} \\
  &\text{if } i = k \quad \text{then the pivot value is the answer} \\
  &\quad \text{return } A[q] \\
  &\text{else if } i < k \quad \text{then the answer is in the front} \\
  &\quad \text{return Randomized-Select}\(A,p,q-1,i\) \\
  &\text{else} \quad \text{then the answer is in the back half} \\
  &\quad \text{return Randomized-Select}\(A,q+1,r,i-k\)
  \end{align*}
\]
Randomized Selection

• Analyzing RandomizedSelect()
  − Worst case: partition always 0:n-1
    \[ T(n) = T(n-1) + O(n) \]
    \[ = O(n^2) \]
  − No better than sorting!
  − "Best" case: suppose a 9:1 partition
    \[ T(n) = T(9n/10) + O(n) \]
    \[ = O(n) \] (Master Theorem, case 3)
  − Better than sorting!
  − Average case: \( O(n) \) remember from quicksort

Worst-Case Linear-Time Selection

• Randomized algorithm works well in practice
• What follows is a worst-case linear time algorithm, really of theoretical interest only
• Basic idea:
  − Guarantee a good partitioning element
  − Guarantee worst-case linear time selection
• Warning: Non-obvious & unintuitive algorithm ahead!
• Blum, Floyd, Pratt, Rivest, Tarjan (1973)

Order Statistics: Algorithm

Select(A, n, i):
  Divide input into \([n/5]\) groups of size 5.
  /* Partition on median-of-medians */
  medians = array of each group's median.
  pivot = Select(medians, \([n/5]\), \([n/10]\))
  Left Array L and Right Array G = partition(A, pivot)
  /* Find i\textsuperscript{th} element in L, pivot, or G */
  k = \(|L| + 1\)
  If i=k, return pivot
  If i<k, return Select(L, k-1, i)
  If i>k, return Select(G, n-k, i-k)

Order Statistics: Analysis

\[ T(n) = T\left(\frac{n}{5}\right) + T\left(\max(k-1, n-k)\right) + O(n) \]

Only one done.
Order Statistics: Analysis

All groups of 5 elements. (And at most one smaller group.)

Order Statistics: Analysis 1

\[
\left\lceil \frac{n}{5} \right\rceil \text{ full groups of 5} \\
\left\lceil \frac{n}{5} \right\rceil \text{ partial groups of 2}
\]

Order Statistics: Analysis

Definitely Lesser Elements

Definitely Greater Elements

Order Statistics: Analysis 1

Must recur on all elements outside one of these boxes. How many?

Order Statistics: Analysis

\[
T(n) = T\left(\left\lfloor \frac{n}{5} \right\rfloor\right) + T\left(\frac{7n}{10} + 6\right) + O(n)
\]

A very unusual recurrence. How to solve?

Order Statistics: Analysis

Substitution: Prove \( T(n) \leq c \times n \).

\[
T(n) \leq c \times \left\lceil \frac{n}{5} \right\rceil + c \times \left(\frac{7n}{10} + 6\right) + d \times n
\]

Overestimate ceiling

\[
= \frac{9}{10} c \times n + 7c + d \times n
\]

Algebra

\[
= c \times n - (c \times \frac{n}{10} - 7c - d \times n)
\]

Algebra

\[
\leq c \times n
\]

when choose \( c,d \) such that \( 0 \leq c \times \frac{n}{10} - 7c - d \times n \)
Order Statistics

Why groups of 5?

Sum of two recurrence sizes must be < 1.
Grouping by 5 is smallest size that works.