Linear-Time Selection

Randomized Selection (Algorithm)

Design and Analysis of Algorithms I

Prerequisites

Watch this after:

- QuickSort - Partitioning around a pivot
- QuickSort – Choosing a good pivot
- Probability Review, Part I

The Problem

for simplicity

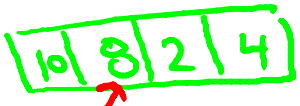
Input: array A with n distinct numbers and a number $i \in \{1, 2, \ldots, n\}$.

Output: i^{th} order statistic (i.e., i^{th} smallest element of input array)

Example: median.

($i = \frac{n+1}{2}$ for n odd,

$i = n/2$ for n even)



3rd order statistic

Reduction to Sorting

O(n log n) algorithm

① apply MergeSort ② return i^{th} element of sorted array

Fact: can't sort any faster [see optional video].

Next: O(n) time (randomized) by modifying QuickSort.

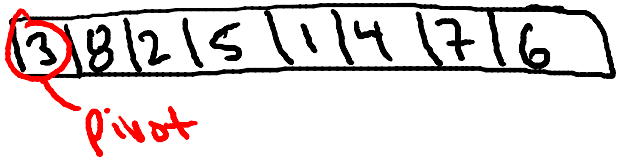
Optional video: O(n) time deterministic algorithm.

- pivot = "median of medians" (warning: not as practical)

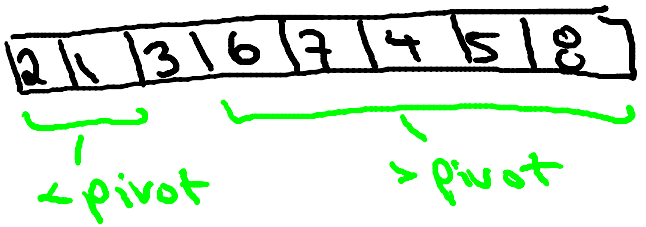
Partitioning Around a Pivot

Key idea: partition array around a pivot element.

- pick element of array



- rearrange array so that:
  - left of pivot $\Rightarrow$ less than pivot
  - right of pivot $\Rightarrow$ greater than pivot



Note: puts pivot in its "rightful position".

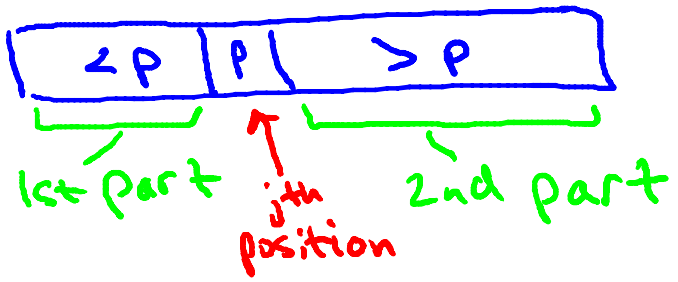
Suppose we are looking for the 5th order statistic in an input array of length 10. We partition the array, and the pivot winds up in the third position of the partitioned array. On which side of the pivot do we recurse, and what order statistic should we look for?

- ○ The 3rd order statistic on the left side of the pivot.
- ○ The 2nd order statistic on the right side of the pivot.
- ○ The 5th order statistic on the right side of the pivot.
- ○ Not enough information to answer question – we might need to recurse on the left or the right side of the pivot.

Randomized Selection

RSelect (array A, length n, order statistic i)

0. if n=1 return A[1]
1. choose pivot p from A uniformly at random
2. partition A around p
   let j = new index of p
3. if j=i return p
4. if j>i return RSelect (1st part of A, j-1, i)
5. [if j<i] return RSelect (2nd part of A, n-j, i-j)



Properties of RSelect

Claim: RSelect is correct (guaranteed to output i^{th} order statistic).

Proof: by induction. [like in optional QuickSort video]

Running Time?: depends on "quality" of the chosen pivots.

What is the running time of the RSelect algorithm if pivots are always chosen in the worst possible way?

- ○ $\theta(n)$
- ○ $\theta(n \log n)$
- ○ $\theta(n^2)$
- ○ $\theta(2^n)$

Example:

- suppose $i = \frac{n}{2}$
- suppose choose pivot = minimum every time

$\Rightarrow \Omega(n)$ time in each of $\Omega(n)$ recursive calls

Running Time of RSelect?

Running Time?: depends on which pivots get chosen.
(could be as bad as $O(n^2)$)

Key: find pivot giving "balanced" split.

Best pivot: the median! (but this is circular)

=> would get recurrence $T(n) \leq T(\frac{n}{2}) + O(n)$

=> $T(n) = O(n)$ [case 2 of Master Method]

Hope: random pivot is "pretty good" "often enough".

Running Time of RSelect

RSelect Theorem: for every input array of length n, the average running time of RSelect is $O(n)$.

- holds for *every* input [no assumptions on data]
- "average" is over random pivot choices made by the algorithm