1 Selection Sort

Suppose we want to select the kth smallest elements from a group of n numbers. What if we want to find the median element?

1.1 Pseudocode

Algorithm 1 Selection Sort (*Recursive*)

```
function SELECT(A, k, p, r)
    s ← APPROXIMATE_MEDIAN(A, p, r)
    q ← PARTITION(A, p, r, s)
    if k < q − p + 1 then
        SELECT(A, p, q − 1, l)
    else if k > q − p + 1 then
        SELECT(A, q + 1, r, k − (q − p + 1))
    else
        return q
    end if
end function
```

Algorithm 11 Selection Sort (*Non-Recursive*)

```
function SELECT(A, k)
    p ← 1
    r ← n
    repeat
        s ←APPROXIMATE_MEDIAN(A, p, r)
        q ←PARTITION(A, p, r, s)
        if k < q then r ← q − 1
        else if k > q then p ← q + 1
        end if
    until k = q return q
end function
```

Take an approximate median of the list, partition with this approximate median (q), then look to the left or right depending on how k compares to the partition.

1.2 Analysis

Worst Case

Just like all the other bad quadratic sorting algorithms, selection sort in the worst case results in $n(n-1)$ comparisons

Best Case

If we assume that our median is the true median or that our approximate median is close enough we have the reccurence

$$\begin{aligned} T(n) &= T\left(\frac{n}{2}\right) + n - 1 \quad T(1) = 0 \\ &= (n-1) + \left(\frac{n}{2} - 1\right) + \left(\frac{n}{4} - 1\right) + \cdots \\ &\approx 2n \end{aligned}$$

So best case we have $\approx 2n$ comparisons with the assumption that the pivot is exactly in the middle.

Average Case

Assume the average case occurs at either the $\frac{1}{4}$ mark or the $\frac{3}{4}$ mark (similar to quicksort). Under this assumption we can write down the recurrence (we assume a pessimistic approach in that we always choose the larger side)

$$\begin{aligned} T(n) &= T\left(\frac{3n}{4}\right) + n - 1 \quad T(1) = 0 \\ &= (n-1) + \left(\frac{3}{4}n - 1\right) + \left(\left(\frac{3}{4}\right)^2 n - 1\right) + \cdots \\ &= n\left[1 + \left(\frac{3}{4}\right) + \left(\frac{3}{4}\right)^2 + \cdots\right] \\ &= n\frac{1}{1-\frac{3}{4}} = n\frac{1}{\frac{1}{4}} \\ &= 4n \end{aligned}$$

To find a pessimistic view (upper bound, choose the bigger side of the partition) for the average case we then sum over all possible pivots to get the probabalistic analysis

$$\begin{aligned} T(n) &= \sum_{q=1}^{n} \frac{1}{n} T\left(\max(q-1, n-q)\right) + n - 1 \quad T(1) = 0 \\ &= \frac{1}{n}\sum_{q=1}^{n} T\left(\max(q-1, n-q)\right) + n - 1 \\ &= \frac{2}{n}\sum_{q=\frac{n}{2}}^{n-1} T(q) + n - 1 \qquad \text{by change of variable} \end{aligned}$$

Then from here we guess that the algorithm is linear, more specifically we guess $T(n) \leq an$.

$$\begin{aligned}
T(n) &= \frac{2}{n} \sum_{q=\frac{n}{2}}^{n-1} aq + n - 1 \\
&= \frac{2a}{n} \sum_{q=\frac{n}{2}}^{n-1} q + n - 1 \\
&= \frac{2a}{n} \left[\sum_{q=1}^{n-1} q - \sum_{q=1}^{\frac{n}{2}-1} \right] + n - 1 \\
&= \frac{2a}{n} \left[\frac{n(n-1)}{2} - \frac{\frac{n}{2}\left(\frac{n}{2}-1\right)}{2} \right] + n - 1 \\
&= \frac{2a}{n} \left[\frac{n^2}{2} - \frac{n}{2} - \frac{n^2}{8} + \frac{n}{4} \right] + n - 1 \\
&= an - a - \frac{an}{4} + \frac{a}{2} + n - 1 \\
&= \frac{3}{4}an - \frac{a}{2} + n - 1 \\
&= \left(\frac{3}{4}a + 1 \right) n - \frac{a}{2} - 1 \qquad \text{for induction to work we need } \frac{3}{4}a + 1 \leq an \\
T(n) &\approx 4n
\end{aligned}$$

1.3 Finding Median

For this method we will try finding the explicit median to try to get to best case scenario from above.

1. Put the elements into a $5 \times \frac{n}{5}$ grid.
2. Find the median of each column. $\frac{10n}{5} = 2n$ comparisons
3. Within each column move the small elements in the top, large elements in the bottom and median to the middle.
4. Find the median of medians. $T\left(\frac{n}{5}\right)$ comparisons
5. Move the columns with small medians to the left, large medians to the right, and the median of medians to the middle.
6. Partition using median of medians as pivot. $n - 1$ comparisons
7. Recursively call algorithm on proper side. $T\left(\frac{7n}{10}\right)$ comparisons

Then, all together we have the recurrence

$$\begin{aligned} T(n) &\le 2n + T\left(\frac{n}{5}\right) + n - 1 + T\left(\frac{7n}{10}\right) \qquad T(1) = 0 \\ &= T\left(\frac{n}{5}\right) + T\left(\frac{7n}{10}\right) + 3n - 1 \end{aligned}$$

Then from here we guess that the algorithm is linear, more specifically we guess $T(n) \le an$.

$$\begin{aligned} T(n) &= a\frac{n}{5} + a\frac{7n}{10} + 3n - 1 \\ &= a\frac{9}{10}n + 3n - 1 \\ &= \left(\frac{9}{10}a + 3\right)n - 1 \quad \text{for induction to work we need } \frac{9}{10}a + 3 \le an \\ T(n) &\approx 30n \end{aligned}$$

So we see that for $n \ge 2^{30}$ finding the median of medians will be beneficial, but this approach can be extended to a grid of $7 \times \frac{n}{7}$ or even $9 \times \frac{n}{9}$ which results in increasingly more efficient times.