Randomized Selection

Algorithm $RSelect(S, k)$:

// $S$ is a sequence of numbers, $1 \leq k \leq |S|$,
// returns the $k$-th smallest element in $S$

if $|S| = 1$ : return the only element in $S$
if $|S| > 2$: $p =$ uniformly random element in $S$;
by scanning $S$, divide it into

\[
\begin{array}{c|c|c}
< p & = p & > p \\
\hline
S_< & S = & S_>
\end{array}
\]

if $k \leq |S_<|$ : $RSelect(S_<, k)$
if $|S_<| + 1 \leq k \leq |S_<| + |S = | + 1$ : return $p$
if $k > |S_<| + |S = | + 1$:

$RSelect(S_>, k - |S_<| - |S = |)$
Let $n$ be a large integer, let $S$ be a sequence of length $n$. Define random variable $T =$ time when running $RSelect(S,k)$.

We are going to show: $E(T) = O(n)$.

When we run $RSelect(S,k)$, recursive calls are generated on smaller and smaller sequences.

For $i = 0,1,2,...$: a call is in phase $i$ if the length of the sequence in this call is

$$i \leq \left(\frac{3}{4}\right)^i n \quad \text{and} \quad \left(\frac{3}{4}\right)^{i+1} n.$$ 

Initially: phase 0, because $|S| = n$

During recursive calls, we either stay in the same phase or move to a phase with a larger index. Which case happens depends on the pivot $p$ (which is randomly chosen).
Consider a call in phase $i$.

$m = \text{length of the sequence in this call.}$

$$ (\frac{3}{4})^i \cdot n < m \leq (\frac{3}{4})^{i+1} \cdot n $$

For purpose of analysis, consider these $m$ numbers in sorted order:

```
< m/4 |< m/2 |< m/4 |
   |   |   |
   |   | Good pivots |
   |   |
```

pivot $p$ is good if $p$ is in the middle half

* $Pr(\text{pivot } p \text{ is good}) = \frac{1}{2}$

* Assume pivot $p$ is good. If there is a next call, then it is on a sequence of length

$$ \leq m - m/4 = \frac{3}{4} \cdot m \leq \left(\frac{3}{4}\right)^{i+1} \cdot n $$

\[ \therefore \text{ this next call is in a phase } > i+1 \]
From COMP 2804:

- experiment \( \xrightarrow{\text{success}} \) with probability \( \alpha \)
- failure with probability \( 1 - \alpha \)

Repeat experiment until success \( \# \) for the first time.

Then: \( E(\# \text{ times}) = \frac{1}{\alpha} \)

Define random variable

\[ X_i = \# \text{ calls in phase } i \]

Then: \( E(X_i) \leq 2 \)  
[Exercise: Why \( \leq 2 \)?  
Why not \( = 2 \)?]

\* time for one call in phase \( i \) (excluding the recursive call)

\[ \leq cm \leq c^{\frac{3}{4}} \cdot n \]

for some constant \( c \).
Remember: \( T = \text{time when running } RSelect(S, k) \), where \(|S| = n\).

\[
T \leq \sum_{i=0}^{\infty} c(\frac{3}{4}) \cdot n \cdot X_i
\]

Using linearity of expectation:

\[
E(T) \leq \sum_{i=0}^{\infty} c(\frac{3}{4}) \cdot n \cdot E(X_i)
\]

\[
\leq 2cn \sum_{i=0}^{\infty} \left( \frac{3}{4} \right)^i
\]

\[
= 2cn \cdot \frac{1}{1 - \frac{3}{4}}
\]

\[
= 8cn
\]

\[
= O(n)
\]