CSE515: Advanced Algorithms
Notes on Lecture 22: Randomized Selection

Antoine Vigneron

May 13, 2021

We give an alternate analysis of the randomized selection algorithm. We make a proof by induction, which is also called the *substitution method* in the MIT textbook and in CSE331.

Remember that the running time of SELECT, ignoring recursive calls, is at most cn for some constant c. We will now show that the expected running time $T(n)$ of the whole algorithm, including recursive calls, is at most an for some larger constant a.

Let $\ell = |S^-|$ be the size of the subset containing elements less than the pivot. Then the algorithm may recurse either on the subset S^- of size ℓ, or on the subset S^+ of size $n - \ell - 1$. So the running time satisfies

$$T(n) \leqslant \max\left(T(\ell), T(n-\ell-1)\right).$$

We make the Induction Hypothesis (IH) that $T(m) \leqslant am$ for all $0 < m < n$, and we want to prove that $T(n) \leqslant an$. Each value of $\ell = 0, \dots, n-1$ occurs with probability $1/n$. Therefore, the expected running time satisfies

$$T(n) \leqslant cn + \frac{1}{n}\sum_{\ell=0}^{n-1} \max\left(T(\ell), T(n-\ell-1)\right).$$

By IH, it implies that

$$\begin{aligned}
T(n) &\leqslant cn + \frac{1}{n}\sum_{\ell=0}^{n-1} \max\left(a\ell, a(n-\ell-1)\right) \\
&= cn + \frac{a}{n}\sum_{\ell=0}^{n-1} \max\left(\ell, (n-\ell-1)\right) \\
&\leqslant cn + \frac{a}{n}\left((n-1)+(n-2)+\cdots+\left\lfloor\frac{n-1}{2}\right\rfloor + \left\lfloor\frac{n-1}{2}\right\rfloor + \cdots + (n-2)+(n-1)\right) \\
&= cn + \frac{2a}{n}\sum_{i=\lfloor (n-1)/2\rfloor}^{n-1} i
\end{aligned}$$

Using the formula $\sum_{i=1}^{N} i = \frac{N(N+1)}{2}$ we obtain:

$$\begin{aligned}
T(n) &\leqslant cn + \frac{2a}{n}\left(\frac{n(n-1)}{2} - \frac{\lfloor (n-1)/2 \rfloor \lfloor (n-1)/2 - 1 \rfloor}{2}\right) \\
&= cn + \frac{2a}{n}\left(\frac{n(n-1)}{2} - \frac{(n-3)(n-5)}{8}\right) \\
&= cn + \frac{a}{4n}\left(4n(n-1) - (n-3)(n-5)\right) \\
&= cn + \frac{a}{4n}(3n^2 + 4n - 15) \\
&\leqslant cn + \frac{a}{4n}(3n^2 + 4n) \\
&= cn + \frac{3}{4}an + a \\
&= cn + an\left(\frac{3}{4} + \frac{1}{n}\right).
\end{aligned}$$

Thus, if $n \geqslant 8$, we have $T(n) \leqslant cn + \frac{7}{8}an$. So if we choose $a \geqslant 8c$, we get $T(n) \leqslant an$. Therefore we need to handle the base cases $m = 1, \dots, 8$ by choosing a larger than $T(m)/m$ for $m = 1, \dots, 8$.

In summary, if we choose $a = \max(8c, T(1)/1, T(2)/2, \dots, T(8)/8)$, we have proved that, if for all $m < n$ we have $T(m) \leqslant am$, then $T(n) \leqslant an$. It proves by induction that $T(n) \leqslant an$ for all n, and thus $T(n) = O(n)$.