IE531: Algorithms for Data Analytics
Spring, 2020
Homework 1: The Median-of-Medians Algorithm
Due Date: February 7, 2020

Instructions

1. You can modify any of the Python code on Compass to solve these problems, if you want. It might help you with honing your programming skills.
2. You will submit a PDF-version of your answers on Compass on-or-before midnight of the due date.

Instructions

When we discussed the Median-of-Medians Algorithm in class, we stopped the recursion when the array-length is less-than-or-equal-to 5; when this condition was true, we just used a sort-and-pick method to pick the appropriate k-th smallest element. If $T_5(n)$ is the running-time for this version of the Median-of-Medians algorithm for an array-size of n, we get the recursion

$$T_5(n) \leq c_5 n + T_5\left(\frac{n}{5}\right) + T_5\left(\frac{7n}{10}\right) \Rightarrow T_5(n) = 10c_5 n \tag{1}$$

We will consider versions of this algorithm where we stop the recursion when the array-length is less-than-or-equal-to some m (i.e. m is odd, and $m \neq 5$), and we are interested in figuring out what $T_m(n)$ would be for different values of m. In fact, $T_m(n) = \alpha c_m n$, for an appropriate α and c_m. This homework is about the nitty-gritty details of this process.

1. (*50 points*) Show that for any odd $m \geq 5$, the running-time of the Median-of-Medians algorithm will be

$$T_m(n) = \frac{4m}{m-3} \times c_m \times n$$

The asymptotic running time for the sort-and-pick algorithm is $O(nlog(n))$, where n is the array-size. That is, for large n, the running-time is $\widehat{c} \times n \times log(n)$, for some $\widehat{c}$ (that depends on your computer). While recognizing the fact that these asymptotic expressions are valid only when n is very large, if we blindly used this to estimate the running-time for the sort-and-pick algorithm when $n = 5$, we would get $\widehat{c} \times 5 \times log(5)$ as the possible running-time. When the stopping-length of the recursion $m = 5$, Lines 5-6-7 in figure 7 of my notes (where I reviewed the Computation pre-requisites) will be executed for a total of

$$\frac{n}{\cancel{5}} \times \widehat{c} \times \cancel{5} \times log(5) = \underbrace{\widehat{c} \times log(5)}_{=c_5 \text{ of Eq.1}} \times n$$

time. While acknowledging the inappropriateness of using asymptotic results for arrays of smaller size, through a sequence of deductive steps, we must conclude that the (asymptotic) running time of the Median-of-Medians Algorithm when we stop the recursion when the array-size $m = 5$, will be $10 \times \widehat{c} \times log(5) \times n$.

1. (*25 points*) Using a similar reasoning, derive an expression (in terms of $\widehat{c}$ and other constants) show that

$$T_m(n) = \frac{4mlog(m)}{m-3} \times \widehat{c}.$$

2. (*25 points*) If the above mentioned inappropriateness of using asymptotic results can be overlooked, show that there is an optimal value for m (i.e. there is a unique m for which $T_m(n)$ will be smallest, for any n).