IE531: Algorithms for Data Analytics
Spring, 2019
Programming Assignment 1: Median-of-Median Algorithms with different Stopping Lengths
Due Date: February 8, 2019

1 Introduction

In Homework 1 you looked at the Median-of-Median Algorithm where the recursion was stopped as soon as the array-size was less-than-or-equal-to $m \in \{5, 7, 9, 11, 13\}$. We violated a bunch of "rules" by using the formulae for asymptotic running-time on arrays of small size (i.e. size m). If the mathematics of Homework 1 is to be believed, we expect to the running-time to be the smallest for an optimal value of m in the range $\{5, 7, 9, 11, 13\}$. In this programming exercise you are going to experimentally verify/refute this claim.

Programming Assignment

You are going to modify the Python Code for the *Deterministic-Selection* (or *Median-of-Medians Algorithm*) written by me to create a version where the recursion stops whenever the array-size is less than m (cf. figure 1). Following this, you are going to implement the pseudo-code shown in figure 2 that verifies/refutes the theory developed in Homework 1.

Figure 3 shows the plots of mean-running-time versus m for different array-sizes (i.e. different values of n). It is not exactly clear that there is a unique value of m for which the mean-running-time is the smallest for all lengths. This is an artifact that comes out of using the asymptotic formulae (i.e. formulae that are correct only as n is very large) for cases where the array size (i.e. n) is small. It is hard to tell if repeating this experiment for very large values of n will confirm the theory discovered in Homework 1.

What I need from you

You can submit your Python code on Compass – make sure it runs without issues. In addition, I would like you to upload a PDF file that shows the plots (cf. figure 3) and a short explanation of why you were able to confirm/refute the theory developed in Homework 1.

Deterministic Select (vector $\mathbf{A}$ of size n, int $k \leq n$, int $m \leq n$)

if $n \leq m$ **then**
    Sort $\mathbf{A}$ and return the element in k-th position.
end if
Partition $\mathbf{A}$ into vectors $\{\mathbf{B}_i\}_{i=0}^{(n/5)-1}$, where each vector $\mathbf{B}_i$ has 5 elements.
for $0 \leq i < n/5$ **do**
    $\mathbf{C}$[i] = **Deterministic Select**($\mathbf{B}_i$, 3)
end for
{ /* $\mathbf{C}$ is a $(n/5)$-long vector, where the i-th entry is the median of $\mathbf{B}_i$ */}.
(*median-of-medians*) p= **Deterministic Select**($\mathbf{C}$, $(n/10)$)
Partition $\mathbf{A}$ into three sub-vectors $\mathbf{A}_{<p}, \mathbf{A}_{=p}$, and $\mathbf{A}_{>p}$.
{/* $\mathbf{A}_{<p}$ has all elements of $\mathbf{A}$ that are less than p. */}
{/* $\mathbf{A}_{=p}$ has all elements of $\mathbf{A}$ that are equal to p. */}
{/* $\mathbf{A}_{>p}$ has all elements of $\mathbf{A}$ that are greater than p. */}
if $k \leq length(\mathbf{A}_{<p})$ **then**
    return **Deterministic Select**($\mathbf{A}_{<p}$, k)
else
    if $k > length(\mathbf{A}_{<p}) + length(\mathbf{A}_{=p})$ **then**
        return **Deterministic Select**($\mathbf{A}_{>p}$, $k - length(\mathbf{A}_{<p}) - length(\mathbf{A}_{=p})$)
    end if
else
    return p
end if

Figure 1: Pseudo-code for the *median-of-medians, deterministic-selection,* where m can be varied.

for $1000 \leq n \leq 10,000$ in steps of 1000 **do**
  for $m \in \{5, 7, 9, 11, 13\}$ **do**
    for $1 \leq i \leq \#trials$ **do**
      Fill an array of size n with random values.
      Let $k = \lceil \frac{n}{2} \rceil$ (i..e k is almost the median)
      Find the amount of time it took to find the k-th smallest element in the array (for the present value of m, array-size n, and trial i)
    end for
    Compute the mean-running-time of the $\#trials$-many experiments (for the present value of m, array-size n)
  end for
  Plot the mean-running-time as a function of m (for the present array-size n).
  Check if there is a value of m for which the mean-running-time is the smallest (for the present array-size n).
end for

Figure 2: Experimentally verifying/refuting the existence of an optimal value for m.

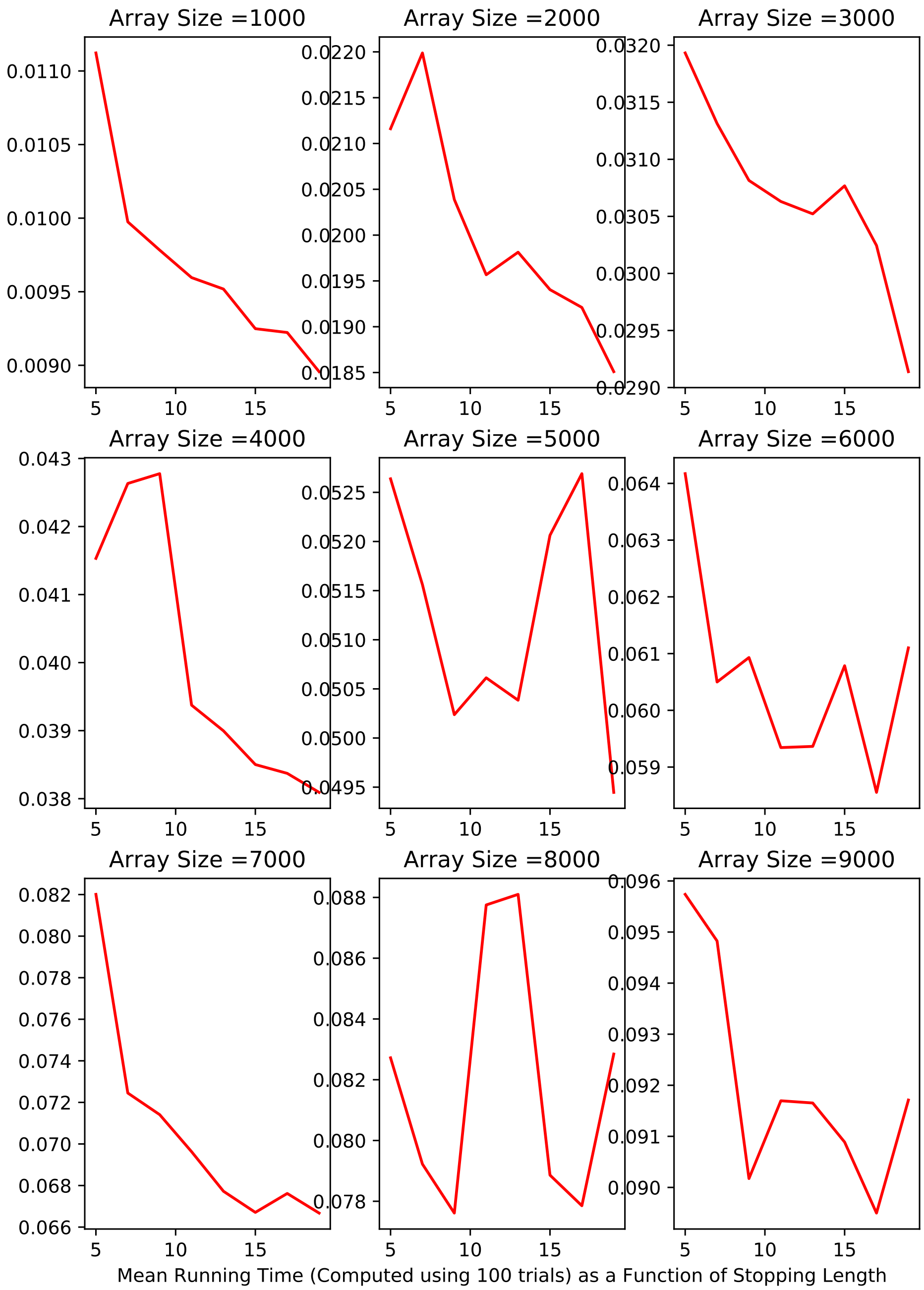

Figure 3: Mean-running-time (from 100 trials) vs. m plots for different values of n