# Algorithm Design 21/22

## Hands On 2 - Depth of a node in a random search tree

Federico Ramacciotti

## 1 Problem

A random search tree for a set $S$ can be defined as follows: if $S$ is empty, then the null tree is a random search tree; otherwise, choose uniformly at random a key $k \in S$: the random search tree is obtained by picking $k$ as root, and the random search trees on $L = \{x \in S : x < k\}$ and $R = \{x \in S : x > k\}$ become, respectively, the left and right subtrees of the root $k$. Consider the Randomized Quick Sort discussed in class and analyzed with indicator variables [CLRS 7.3], and observe that the random selection of the pivots follows the above process, thus producing a random search tree of $n$ nodes.

1. Using a variation of the analysis with indicator variables $X_{ij}$, prove that the expected depth of a node (i.e. the random variable representing the distance of the node from the root) is nearly $2 \log n$.

2. Prove that the expected size of its subtree is nearly $2 \log n$ too, observing that it is a simple variation of the previous analysis.

3. Prove that the probability that the depth of a node exceeds $c * 2 * \log n$ is small for any given constant $c > 2$. [Note: it can be solved with Chernoff's bounds as we know the expected value.]

Note: $\sum_{k=1}^{n} \frac{1}{k} \leq \log n + 1$

## 2 Solution

Define an indicator variable

$$X_{ij} = \begin{cases} 1 & \text{if } x_i \text{ has been compared to } x_j \\ 0 & \text{otherwise} \end{cases}$$

Building the tree from a non-empty set $S$, every node is compared to all its ancestors.

### 2.1 Height

The height of a node is the sum of all the indicator variables of its ancestors; since we cannot know that precisely, we can only observe that the height of a node $i$ is smaller than or equal to the sum of all the elements compared to $i$ (i.e. the sum of all the indicator variables $X_{ij}$, fixing $i$). So, given that the probability that $X_{ij} = 1$ is $\frac{2}{j-i+1}$, the expected depth of a node is:

$$\begin{aligned} E\left[d\left(i\right)\right] &= E\left[\sum_{j \,\in\, ancestors(i)} X_{ij}\right] \leq E\left[\sum_{j=1}^{n} X_{ij}\right] \\ &= \sum_{j=1}^{n} E\left[X_{ij}\right] = \sum_{j=1}^{n} Pr\left[X_{ij}\right] \\ &= \sum_{j=1}^{n} \frac{2}{j-i+1} = 2\sum_{j=1}^{n} \frac{1}{j-i+1} = O(2 \log n) \end{aligned}$$

## 2.2 Size

The expected size of a subtree rooted in a node $i$ is the number of all the descendants on $i$, since every one of them is compared to $i$ once. As observed in the previous analysis, it is again not possible to count only the descendants of a node. The solution is therefore an approximation, saying that the expected size of the subtree in $i$ is smaller than or equal to the sum of all the indicator variables for a fixed $i$ (i.e. the number of elements compared to $z_i$).

$$E\left[size\left(i\right)\right] = E\left[\sum_{j\ \in\ descendants(i)} X_{ij}\right] \leq E\left[\sum_{j=1}^{n} X_{ij}\right] = O(2\log n)$$

## 2.3 Depth

Given the $X_{ij}$ indicator variables already defined, we define also $Y_i = \sum_{j \in ancestors(i)} X_{ij}$ as the depth of a node $i$. Using Chernoff's bounds with $\mu = E[Y_i] = 2\log n$ and $\lambda = 2c\log n - 2\log n = (2c-2)\log n$, we have:

$$\begin{aligned}
Pr\left[Y_i > 2c\log n\right] &\leq e^{-\frac{((2c-2)\log n)^2}{2(2\log n)+(2c-2)\log n}} \\
&= e^{-\frac{(2c-2)^2(\log n)^2}{2c\log n+2\log n}} \\
&= e^{-\frac{(2c-2)^2(\log n)^2}{(2c+2)\log n}} \\
&= e^{-\frac{(2c-2)^2\log n}{2c+2}} \\
&= n^{-\frac{2c^2-4c+2}{c+1}} \\
&= \frac{1}{n^{\frac{2c^2-4c+2}{c+1}}}
\end{aligned}$$

This means that the probability that the depth of a node exceeds $2c\log n$ is small w.h.p. for any $c > 2$. In fact, the exponent of $n$ is proportional to $c > 2$ (i.e. the bigger $c$ the smaller the final result) and so the probability is small.