Depth of a node in a random search tree

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 subtree of the root k. Consider the randomized QuickSort 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. Using a variation of the analysis with indicator variables, prove that the expected depth of a node (i.e. the random variable representing the distance of the node from the root) is nearly $2\ ln\ n$.
Prove that the probability that the expected depth of a node exceeds $c\ 2\ ln\ n$ is small for any given constant $c > 1$. [Note: the latter point can be solved after we see Chernoff's bounds.[1]]

SOLUTION
To give an estimation on the depth of a given node, i, we would need to consider how many of the other nodes are its ancestors. For this, an indicator variable can be defined as follows:

$$X_{ij} = \begin{cases} 1 & \text{IF } j \text{ IS AN ANCESTOR OF } i, \\ 0 & otherwise \end{cases}$$

With this indicator variable, analysis can be performed taking the indices as those of the sorted set when in order: $z_1, z_2, ..., z_n$. For two arbitrary indices z_i, z_j only three possible scenarios apply:

1. z_i was selected as a key on the tree before z_j, so z_j is a successor of z_i and thus $X_{ij} = 0$.
2. Neither z_i nor z_j were selected as a key on the tree before, so a key in the range $(i, ..., j)$ or $(j, ..., i)$ effectively splits the range and thus $X_{ij} = 0$.
3. z_j was selected as a key on the tree before z_i, so z_i will eventually be found and selected as a key preceded by z_j and thus $X_{ij} = 1$.

With this analysis in mind, the expectation of the indicator variable can be computed as follows:

$$E[\sum_{\substack{j=1 \\ j \neq i}}^{n} X_{ij}] = \sum_{\substack{j=1 \\ j \neq i}}^{n} P(X_{ij} = 1) = ...$$

Given the previous analysis, the probability of X_{ij} in the interval containing z_i and z_j on both extremes is that of the only case in which z_j may be an ancestor of z_i over the total amount of cases. In this case that means the number of elements in the range:

$$... = \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{1}{|j - i| + 1} = ...$$

To explicitly take into account the two orderings between z_i and z_j, we split the computation in two terms. The result resembles two instances of the harmonic series, which are then approximated to $ln(n)$:

$$... = \sum_{j=1}^{i-1} \frac{1}{i - j + 1} + \sum_{j=i+1}^{n} \frac{1}{j - i + 1} < ln(n) + ln(n) = 2ln(n)$$

Finally, the depth of a node in a random search tree is expected to be $2ln(n)$. A variation on the same analysis can be performed to estimate the size of the subtree spanning from a given node. In this case, a similar indicator variable is defined with slightly different semantics:

$$X_{ij} = \begin{cases} 1 & \text{IF } j \text{ IS AN SUCCESSOR OF } i, \\ 0 & otherwise \end{cases}$$

The analysis and computations to be performed afterwards will follow the same structure as before, producing in the same expectancy results for the predicate.

[1] Chernoff's bound

SECOND SOLUTION

Let's start with some key observations: the comparisons are just made with the chosen root k, any two elements are compared at most once, and every time a node is compared with the root k, it will increase its depth in the tree. Let denote with $n_1, \ldots, n_k$ the node of a BST (Binary Search Tree), where an $n_t \leq n_p \forall t \leq p$. Let's fix a generic node n_i then we have:

$$X_j = \begin{cases} 1 & \text{NODE } n_i \text{ IS A DESCENDENT OF } n_j \\ 0 & \text{OTHERWISE} \end{cases}$$

Therefore $X = \sum_{j=i}^{n} X_j$ is the hight (or the depth) of a generic node n_1. Therefore now we need to calculate the $E[X]$ (its expected value). Since the expected value is linear we have that $E[X] = \sum_{j=i}^{n} E[X_j]$, and since we know that $E[X_j] = P[X_j = 1]$ we should approximate the latter probability. Since a couple of element can be compared at most once and every comparison means a comparison with the root (an increasing of the depth), we can can assume that: if n_i is in the left(right) subtree of n_j it means that there are at most $j-i+1$ elements in the left(right) subtree. Since the subtree has $j-i+1$ elements, and because root are chosen randomly and independently, the probability that any given element is the first one chosen as a root is $\frac{1}{j-i+1}$.Therefore we have:

$$\begin{aligned} P(X_j = 1) =& P[n_i \text{ IS A DESCENDENT OF } n_j] \\ \leq& P[n_i \text{ IS IN THE LEFT SUBTREE } n_i \text{ IS IN THE RIGHT SUBTREE}] \\ =& \frac{1}{\text{NUMBER OF NODE IN THE LEFT SUBTREE}} + \frac{1}{\text{NUMBER OF NODE IN THE RIGHT SUBTREE}} \\ \leq& \frac{2}{j-i+1} \end{aligned}$$

Therefore we have $E[X] = \sum_{j=i}^{n} \frac{2}{j-i+1}$, if we change of variables[2] $k = j - i$ and we bound the harmonic series we have:

$$\sum_{k=1}^{n-i} \frac{2}{k+1} < \sum_{k=1}^{n} \frac{2}{k} = 2ln(n)$$

Let's write down the Chernoff Bound:

Theorem 1 *(Chernoff Bounds). Let $X = \sum_{i=1}^{n} X_i$, where $X_i = 1$ with probability p_i and $X_i = 0$ with probability $1 - p_i$, and all X_i are independent. Let $\mu = E(X) = \sum_{i=1}^{n} p_i$[3]. Then*

$$\begin{aligned} &(i)\textbf{\textit{Upper Tail:}} P(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2}{2+\delta}\mu} && \text{for all } \delta > 0 \\ &(ii)\textbf{\textit{Lower Tail:}} P(X \leq (1-\delta)\mu) \leq e^{-\frac{\mu\delta^2}{2}} && \text{for all } 0 < \delta < 1 \end{aligned}$$

Therefore we have:

$$\begin{aligned} P(X \geq (1+c)2ln(n)) \leq& e^{-\frac{c^2}{2+c}2ln(n)} \\ =& \frac{1}{e^{\frac{c^2}{2+c}2ln(n)}} \\ =& \frac{1}{n^{\frac{2c^2}{2+c}}} \\ =& \frac{1}{n^c} \text{WITH } c > 2 \end{aligned}$$

[2] LINK1

[3] Indicator variable