University of Pisa

Department computer science

Second hands-on: Depth of a node in a random search tree

Domenico Erriquez

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 $2 \ln n$.

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

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

2. SOLUTION

2.1 Proof for the first point

To compute the expected depth of a node in RBST(random binary search tree), an indicator variable for each pair of nodes (z_i, z_j) is defined as follows:

$$X_{ij} = \begin{cases} 1, & if\ z_j\ is\ ancestor\ of\ z_i \\ 0, & otherwise \end{cases}$$

The depth of a node (e.g. z_i) in RBST can be expressed in terms of the sums of indicator variables as follows:

$$depth(z_i) = \sum_{j=1}^{n} X_{ij} = \#\ ancestors\ of\ z_i$$

We would like to compute the expected value of the depth of a node z_i. This value can be calculated as:

$$E[depth(z_i)] = E\left[\sum_{j=1}^{n} X_{ij}\right] = \sum_{j=1}^{n} E[X_{ij}] = \sum_{j=1}^{n} \Pr[X_{ij} = 1]$$

The next step is to estimate the probability of an indicator variable X_{ij} to be equal to 1 (i.e. z_j is an ancestor of z_i).

$$\Pr[X_{ij} = 1]\; Pr[z_j\ is\ a\ pivot \mid z_j\ and\ z_i\ are\ in\ the\ same\ partition] = \frac{1}{|j-i|+1}$$

We can now define the probability of the indicator variable X_{ij} to be equal to 1 as:

$$\Pr\,[X_{ij} = 1] \begin{cases} \dfrac{1}{j-i+1}, & if\ i < j \\ 0, & if\ i = j \\ \dfrac{1}{i-j+1}, & if\ i > j \end{cases}$$

Finally, the expected depth of node z_i can be computed as follows:

$$E[depth(z_i)] = \sum\nolimits_{j=1}^{n} \Pr[X_{ij} = 1] = \sum_{j=1}^{i-1} \frac{1}{i-j+1} + \sum_{j=i+1}^{n} \frac{1}{j-i+1}$$

$$= \frac{1}{i} + \frac{1}{i-1} + \cdots + 1 + \frac{1}{2} + \cdots + \frac{1}{n-i+1}$$

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

2.2 Proof for the second point

To demonstrate the second point, an approach similar to the one utilized in the first point can be used. Given a node z_i we have to approximate the number of descendant nodes. To begin with, let's define the indicator variables as:

$$Y_{ij} = \begin{cases} 1, & if\ z_j\ is\ descendant\ of\ z_i \\ 0, & otherwise \end{cases}$$

Therefore, the size of the subtree of z_i is:

$$subtree\_size(zi) = \sum_{j=1}^{n} Y_{ij} = \#\ descendant\ of\ zi$$

While the expected size of the subtree is:

$$E[subtree\_size(z_i)] = E\left[\sum_{j=1}^{n} Y_{ij}\right] = \sum_{j=1}^{n} E[Y_{ij}] = \sum_{j=1}^{n} \Pr[Y_{ij} = 1]$$

Now as the precedent point, we need to compute the probability of an indicator variable Y_{ij} to be equal to 1, that is the probability of z_j to be descendant of z_i. Saying that z_j is a descendant of z_i is the same as saying that z_i is an ancestor of z_j. So, the probability of z_j to be descendant of z_i is the same of the probability of z_i to be an ancestor of z_j, and we know this probability from the previous point and it's the following:

$$P\,r[Y_{ij} = 1] = P\,r[X_{ij} = 1] = \frac{1}{|j-i|+1}$$

Finally, we can compute the expected size of the subtree of z_i in the same way we computed the expected depth of the node z_i in the previous point.

$$E[subtree\_size(z_i)] = \sum_{j=1}^{n} \Pr[Y_{ij} = 1] = \sum_{j=1}^{n} \frac{1}{|j-i|+1} \approx 2 * \ln(n)$$

2.3 Proof for the third point

To prove that the probability that the depth of a node exceeds $c * 2 * \ln n$ is small for any given constant $c > 2$ the Chernoff bound can be used.

Let's start by defining what Chernoff bound is:

$$\Pr[X > \mu + \lambda] \le e^{-\frac{\lambda^2}{2\mu+\lambda}}$$

Where X is the sum of indicator random variables, $\mu = E[X]$ is the expected value of the sum and $\lambda > 0$ a constant.

Now we can proceed with proving that the probability for a node z_i to have depth greater than $2 * c * \ln(n)$ is small for a constant $c > 2$.

Given a node z_i we know from the first point that $depth(z_i) = \sum_{j=1}^{n} X_{ij}$ and $E[depth(z_i)] \approx 2\ln(n)$ for a RBST with n nodes. So, we want to prove that $Pr[depth(z_i) > 2c * \ln(n)]$ is small for a constant $c > 2$.

In order to use Chernoff bound we have to write $2c * \ln(n)$ as the sum between μ and λ, knowing that $\mu = E[depth(z_i) \approx 2\ln(n)$, we can find λ as follows:

$$\lambda + \mu = 2 * c * \ln(n)$$
$$\lambda = 2c\ln(n) - \mu = 2c\ln(n) - 2\ln(n) = (2c-2)\ln(n)$$

Finally, Chernoff bound becomes:

$$\begin{aligned}\Pr[depth(z_i) > \lambda + \mu] = \Pr[depth(z_i) > 2c\ \ln(n)] &\le e^{-\frac{((2c-2)\ln(n))^2}{2(2\ln(n)) + (2c-2)\ln(n)}} \\ &= e^{-\frac{(2c-2)^2\ln(n)^2}{2c\ln(n) + 2\ln(n)}} \\ &= e^{-\frac{(2c-2)^2\ln(n)^2}{(2c+2)\ln(n)}}\end{aligned}$$

$$
\begin{aligned}
&= e^{-\frac{(2c-2)^2 \ln n}{(2c+2)}} \\
&= e^{\ln(n)^{-\frac{(2c-2)^2}{(2c+2)}}} \\
&= n^{-\frac{(2c-2)^2}{2c+2}} \\
&= n^{-\frac{2c^2-4c+2}{c+1}} \\
&= \frac{1}{n^{\frac{2c^2-4c+2}{c+1}}}
\end{aligned}
$$

We can conclude that the probability that the depth of a node to exceed $2c * \ln(n)$ is small for a constant $c > 2$. More c is big, lower is this probability.