Notes, 8(a)

ECE 606

Randomization, Probabilistic Analysis, Approximation Algorithms

We consider several examples of the above themes. We start with so-called randombly built binary search trees.

Recall: binary search tree is a binary tree which satisfies the following binary search property. Suppose $key(u), left(u)$ and $right(u)$, for a node u, are the key, left-child and right-child respectively. Then, the binary search property is: $key(left(u)) \leq key(u) \leq key(right(u))$, if indeed $left(u)$ and $right(u)$ exist.

Also recall the simple algorithm to build a binary search tree: start with an empty tree. Then, insert each new key as a leaf. Depending on the arrival sequence of the new keys, we may end up with either a balanced, or an unbalanced tree. For example, if the keys arrive sorted non-decreasing, then we end up with a right going chain.

How balanced the tree is, in turn, determines how efficiently we can perform lookups of keys. If the height of the tree is h, then the worst-case number of comparisons is $h + 1$. The best possible lower-bound for this, in the worst-case, is $\Theta(\lg n)$. The worst-case is $\Theta(n)$.

So: we know that the simple algorithm above is quite bad in the worst-case. But should that alone cause us to say that the algorithm is bad? How bad is it in the expected case?

To study the expected case, we have to ask what the random event is, or if there is one at all. Suppose we assume that every arrival sequence of the n distinct keys which comprise the keys in the tree is equally likely. This is equivalent to saying: if we seek to construct a tree of n distinct keys, suppose we randomly permute them before we start our algorithm that inserts them in sequence.

Then, we ask: what is the expected height of the resultant binary search tree?

Random variables and other mnemonics:

- $\{1, 2, \ldots, n\}$: the keys we seek to insert.
- X_j: height of tree of j keys.
- Y_j: 2^{X_j}; "exponential height." Define $Y_0 = 0$.
- R_j: key of the root of a tree of j keys.
- $Z_{n,i} = I\{R_n = i\}$.

$$
\begin{aligned}
Y_n &= \sum_{i=1}^{n} Z_{n,i} \cdot (2 \cdot \max\{Y_{i-1}, Y_{n-i}\}) \\
E[Y_n] &= \sum_{i=1}^{n} E[Z_{n,i} \cdot (2 \cdot \max\{Y_{i-1}, Y_{n-i}\})] \\
&= \sum_{i=1}^{n} E[Z_{n,i}] \cdot E[2 \cdot \max\{Y_{i-1}, Y_{n-i}\}] \\
&= \frac{2}{n} \sum_{i=1}^{n} E[\max\{Y_{i-1}, Y_{n-i}\}] \\
&\leq \frac{2}{n} \sum_{i=1}^{n} (E[Y_{i-1}] + E[Y_{n-i}]) = \frac{4}{n} \sum_{i=0}^{n-1} E[Y_i] \\
&\leq \frac{1}{4} \binom{n+3}{3} = O(n^3)
\end{aligned}
$$

$$
2^{E[X_n]} \leq E\left[2^{X_n}\right] = E[Y_n] = O(n^3)
$$

$$
\implies E[X_n] = O(\lg n)
$$