Expected Height of a Randomly-Built Binary Search Tree

Based on the Cormen et al., Section 12.4, pp 299-303

To randomly build a binary search tree, all permutations of inputs (keys) are equally likely. In other words, given n keys, each of the $n!$ inputs are equally likely. Note that this is different from stating that all binary search trees on n keys are equally likely. Why?

We will first introduce two random variables.

1. X_n is the height of a randomly-built binary search tree on n keys. Note that $X_1 = 0$.
2. $Y_n = 2^{X_n}$ is the exponential height of a randomly-built binary search tree on n keys where:
   (a) $Y_0 = 0$ (by definition),
   (b) $Y_1 = 1$, and
   (c) $Y_n = 2 \cdot \max(Y_{i-1}, Y_{n-i})$.

The probability that a randomly-built binary search has a left subtree with $i-1$ keys and a right subtree with $n-i$ keys, $1 \leq i \leq n$, is $1/n$ since each of the n keys is equally likely to be the root. Hence, the expected value of $E[Y_n]$ is expressed as:

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

We will now prove by induction that $E[Y_n] \leq cn^3$ where $c \geq 1$.

Basis of Induction

$$E[Y_0] = Y_0 = 0 \leq c \cdot 0^3 \text{ where } c \geq 1$$
$$E[Y_1] = Y_1 = 1 \leq c \cdot 1^3 \text{ where } c \geq 1$$

Induction Hypothesis Assume that $E[Y_i] \leq c \cdot i^3,\ 0 \leq i < n$.

Inductive Step

$$\begin{aligned} E[Y_n] &\leq \frac{4}{n} \sum_{i=0}^{n-1} E[Y_i] \\ &\leq \frac{4}{n} \sum_{i=0}^{n-1} c \cdot i^3 \\ &\leq \frac{4}{n} \sum_{i=1}^{n-1} c \cdot i^3 \\ &\leq \frac{4c}{n} \left(\frac{(n-1)^2 \cdot n^2}{4} \right) \\ &\leq cn(n-1)^2 \\ &\leq cn^3 \quad \square \end{aligned}$$

To determine the expected value of X_n, we can use Jensen's inequality which states that:

$$f(E[X]) \leq E[f(X)]$$

provided f is a convex function. In our case, $f(X_n) = Y_n = 2^{X_n}$ is a convex function. Hence,

$$\begin{aligned} f(E[X_n]) &= 2^{E[X_n]} \\ &\leq E[2^{X_n}] \\ &\leq E[Y_n] \\ &\leq cn^3 \end{aligned}$$

Therefore, taking the $\log_2$ on both sides yields our result.

$$E[X_n] \leq \log_2 c + 3\log_2 n \in O(\log n)$$