BLACKBOARD PROOF

CSE202 – WEEK 10

Unsuccessful search in a BST

The analysis of the complexity of a successful or a unsuccessful search in a binary search tree relies on basic properties of those trees.

We first define formally binary trees in an inductive way.

Definition 1. *A binary tree is either reduced to a leaf, or consists of an internal node (its* root*), connected to two binary trees. The number of internal nodes is called the* size *of the tree.*

Numbers of nodes and leaves are closely related.

Lemma 1. *A binary tree of size n has $n+1$ leaves.*

Proof. By induction on the size. The property clearly holds when the tree is reduced to a leaf. Next, assume that the property holds for all binary trees of size n and consider a binary tree of size $n+1$. By induction, at least one of its internal nodes is connected to two leaves. Consider one such node and the binary tree obtained by replacing it with a leaf, i.e., removing one internal node and two leaves and adding a leaf. This new binary tree is smaller by one internal node and one leaf. By the induction hypothesis, it has $n+1$ leaves, which concludes the proof. □

While the average-case analysis of successful searches in a BST is governed by its expected internal path length P_n (the sum of the depths of its internal nodes), the case of unsuccessful searches is obtained from its expected *external path length* E_n, i.e., the sum of the depths of its leaves. These quantities are related as follows.

Lemma 2. *In a binary tree $\mathcal{T}_n$ of size n, internal and external path lengths obey*

$$E(\mathcal{T}_n) - P(\mathcal{T}_n) = 2n.$$

In particular, this difference depends only on the size.

Proof. The proof is again by induction. The property is clear for trees of size 0. Assume it holds for binary trees of sizes up to n and consider a binary tree $\mathcal{T}_{n+1}$ with $n+1$ internal nodes. Let $\mathcal{L}_k$ be its left subtree and k its size, and $\mathcal{R}_{n-k}$ its right subtree. From the path lengths of the subtrees, those of the tree itself are obtained by adding 1 to the depths of each of the nodes. By the previous lemma, these numbers of nodes differ by 1 between internal and external nodes, which gives

$$\begin{aligned} P(\mathcal{T}_{n+1}) &= P(\mathcal{L}_k) + k + P(\mathcal{R}_{n-k}) + n - k, \\ E(\mathcal{T}_{n+1}) &= E(\mathcal{L}_k) + k + 1 + E(\mathcal{R}_{n-k}) + n - k + 1. \end{aligned}$$

Subtracting these equations and using the induction hypothesis yields

$$\begin{aligned} E(\mathcal{T}_{n+1}) - P(\mathcal{T}_{n+1}) &= (E(\mathcal{L}_k) - P(\mathcal{L}_k)) + (E(\mathcal{R}_{n-k}) - P(\mathcal{R}_{n-k})) + 2 \\ &= 2k + 2(n-k) + 2 = 2n + 2, \end{aligned}$$

which concludes the proof. □