Lecture 9

Today: randomly built binary search trees (BSTs)
- relation to Quicksort
  ⇒ expected node depth
- expected height: big analysis

PS4 out
BST sort:
\[ T \leftarrow \emptyset \] //empty BST
for \( i \leftarrow 1 \) to \( n \)
    do Tree-Insert \( (T, A[i]) \)
Inorder-Tree-Walk \( (\text{root}[T]) \)

Example: \( A = \begin{array}{ccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 3 & 1 & 8 & 2 & 6 & 7 & 5 \end{array} \)

Time: \( O(n) \) for Inorder-Tree-Walk
\( n \) Tree-Inserts?
- worst case: \( \Theta(n^2) \)
  (sorted or reverse sorted)
- lucky case: \( \Theta(n \log n) \)
  \( \Theta(\log n) \) height

\[ \text{sound familiar?} \]
Relation to Quicksort:
BST sort & Quicksort perform the same comparisons, but in a different order.

Example:

3 1 8 2 6 7 5
1 2
2 6 7 5
5
7

[assuming "stable" Partition algorithm]

Randomized BST sort (A):
1. randomly permute A (uniformly)
2. BST sort (A)
   - same running time as Randomized Quicksort as random variable, or in expectation

\[ E[\text{average node depth}] \]

\[ = E \left[ \frac{1}{n} \sum_{i=1}^{n} \left( \text{# comparisons to insert node } i \right) \right] \]

\[ = \frac{1}{n} E[\text{running time}] \]

\[ = \Theta(n \lg n) \]

\[ = \Theta(\lg n). \]
Randomly built BST

- BST resulting from random BST sort
- \( E[\text{average node depth}] = O(\log n) \)
- \( E[\text{height (max. node depth) }] = O(\log n) \)?

- not in general:

\[
\frac{1}{n} \left( n \log n + \sqrt{n} \cdot \sqrt{n} \right) = O(\log n)
\]

Theorem: \( E[\text{height of randomly built BST}] = O(\log n) \)

Proof outline:

1. Prove Jensen's inequality: \( f(E[X]) \leq E[f(X)] \) for any convex function \( f \)

2. Instead of \( X_n = \text{random variable of height} \)
   analyze \( Y_n = 2^{X_n} = \text{exponential height} \)

3. Prove that \( E[Y_n] = O(n^3) \)

4. Conclude \( 2E[X_n] \leq E[2^{X_n}] = E[Y_n] = O(n^3) \)

\( \Rightarrow E[X_n] = O(\log n) \)
Jensen's inequality: 
\[ f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)] \] for any convex function \( f \)

- \( f: \mathbb{R} \rightarrow \mathbb{R} \) is convex if
for all \( x, y \in \mathbb{R} \) and \( \alpha, \beta \geq 0 \) with \( \alpha + \beta = 1 \):
\[ f(\alpha x + \beta y) \leq \alpha f(x) + \beta f(y) \]

Lemma: if \( f: \mathbb{R} \rightarrow \mathbb{R} \) is convex; \( x_1, x_2, \ldots, x_n \in \mathbb{R} \);
and \( \alpha_1, \alpha_2, \ldots, \alpha_n \geq 0 \) and \( \sum_{k=1}^{n} \alpha_k = 1 \),
then \( f(\sum_{k=1}^{n} \alpha_k x_k) \leq \sum_{k=1}^{n} \alpha_k f(x_k) \).

Proof: By induction on \( n \).

Base: \( n=1 \Rightarrow \alpha_1 = 1 \Rightarrow f(\alpha_1 x_1) = f(x_1) = \alpha_1 f(x_1) \)

Step: \( f(\sum_{k=1}^{n} \alpha_k x_k) = f(\alpha_n x_n + (1-\alpha_n) \sum_{k=1}^{n-1} \frac{\alpha_k}{1-\alpha_n} x_k) \)
- algebra
\[ \leq \alpha_n f(x_n) + (1-\alpha_n) f\left(\sum_{k=1}^{n-1} \frac{\alpha_k}{1-\alpha_n} x_k\right) \]
- convexity
\[ \leq \alpha_n f(x_n) + (1-\alpha_n) \sum_{k=1}^{n-1} \frac{\alpha_k}{1-\alpha_n} x_k \]
- induction hypothesis
\[ = \sum_{k=1}^{n} \alpha_k f(x_k) \] - algebra
Jensen's inequality:
\[ f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)] \]
for convex function \( f \).

Here: assume \( X \) is integral
& has bounded range \([0, M]\).

Proof:
\[
\begin{align*}
\mathbb{E}[f(X)] &= f(\sum_{k=0}^{M} \alpha_k \cdot \Pr\{X = k\}) \quad \text{def. of E} \\
&\leq \sum_{k=0}^{M} f(k) \cdot \Pr\{X = k\} \quad \text{convexity lemma} \\
&= \sum_{x \in \text{range}(f)} x \cdot \sum_{k: f(k) = x} \Pr\{X = k\} \\
&= \mathbb{E}[f(X)] \quad \text{def. of E}
\end{align*}
\]

Generalizes to arbitrary integral R.V. \( X \):
\[
\begin{align*}
&\frac{\sum_{k=0}^{M} \alpha_k x_k}{\sum_{i=1}^{M} \alpha_i} \leq \frac{\sum_{k=0}^{M} \frac{\alpha_k}{\sum_{i=1}^{M} \alpha_i} f(x_k)}{\sum_{i=1}^{M} \alpha_i} \\
\Rightarrow \quad &\lim_{M \to \infty} f\left(\frac{\sum_{k=0}^{M} \alpha_k x_k}{\sum_{i=1}^{M} \alpha_i}\right) \leq \lim_{M \to \infty} \frac{\sum_{k=0}^{M} \frac{\alpha_k}{\sum_{i=1}^{M} \alpha_i} f(x_k)}{\sum_{i=1}^{M} \alpha_i} \\
&\quad \to 1 \quad \to \frac{\sum_{k=0}^{M} \alpha_k x_k}{\sum_{i=1}^{M} \alpha_i} \quad \to 1 \quad \to \frac{\sum_{k=0}^{M} \alpha_k f(x_k)}{\sum_{i=1}^{M} \alpha_i}
\end{align*}
\]

Generalizes to continuous random variables too.
**Expected BST height analysis:**

\[ X_n = \text{R.V. of height of randomly built BST on } n \text{ nodes} \]

\[ Y_n = 2^X_n \]

- if root of tree has rank \( k \)
  - then \( X_n = 1 + \max \{ X_{k-1}, X_{n-k} \} \)
  - \( Y_n = 2 \max \{ Y_{k-1}, Y_{n-k} \} \).

- define indicator random variables:

\[ Z_{nk} = \begin{cases} 1 & \text{if root has rank } k \\ 0 & \text{otherwise} \end{cases} \]

\[ E[Z_{nk}] = \Pr \{ Z_{nk} = 1 \} = \frac{1}{2} = Y_n. \]

\[ Y_n = \sum_{k=1}^{n} Z_{nk} \left( 2 \max \{ Y_{k-1}, Y_{n-k} \} \right). \]

\[ \Rightarrow E[Y_n] = E \left[ \sum_{k=1}^{n} Z_{nk} \left( 2 \max \{ Y_{k-1}, Y_{n-k} \} \right) \right] \]

\[ = \sum_{k=1}^{n} E[Z_{nk}] \cdot E \left[ 2 \max \{ Y_{k-1}, Y_{n-k} \} \right] \]

- linearity of expectation

\[ = 2 \sum_{k=1}^{n} E[Z_{nk}] \cdot E[Y_n] \]

- independence of root from subtrees

\[ = \frac{2^n}{n} \sum_{k=1}^{n} E \left[ \max \{ Y_{k-1}, Y_{n-k} \} \right] \]

\[ = \frac{2^n}{n} \sum_{k=1}^{n} E[Y_{k-1} + Y_{n-k}] \quad - \text{max } \leq \text{sum} \]

\[ = \frac{4^n}{n} \sum_{k=0}^{n} E[Y_k] \quad - \text{rearrange terms} \]
Claim: \( E[Y_n] \leq cn^3 \)

Proof: Substitution method.

Base: \( n = O(1) \Rightarrow E[Y_n] = O(1) \leq cn^3 \) for \( c \) sufficiently large.

Step: \( E[Y_n] = \frac{4}{n} \sum_{k=0}^{n-1} E[Y_k] \)

\[ \leq \frac{4}{n} \sum_{k=0}^{n-1} c k^3 \quad \text{--- induction hyp.} \]

\[ \leq \frac{4c}{n} \int_0^n x^3 \, dx \quad \text{--- integral method} \]

\[ = \frac{4c}{n} \cdot \frac{n^4}{4} \]

\[ = cn^3 \quad \Box \]

Conclusion: \( 2E[X_n] \leq E[2^{X_n}] \)

\[ = E[Y_n] \quad \text{--- definition of } Y_n \]

\[ \leq cn^3 \quad \text{--- just shown} \]

\[ \Rightarrow E[X_n] \leq 3 \log n + O(1) \quad \text{--- } \log \text{ of both sides} \]

Question: Why exponential height \( Y_n \), not height \( X_n \)?

Idea: \( \max \{ 2^a, 2^b \} \leq 2^{a+b} \) gets tighter faster than \( \max \{ a, b \} \leq a+b \) as \( |a-b| \) grows.

Exercise: Try it! (or with quadratic).

Right constant: \( \approx 2.9882 \log n \quad [\text{Devroye 1986}] \)