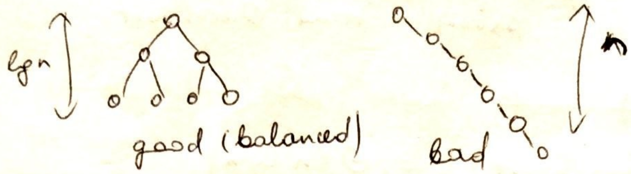
6.046 Lecture 9

Relation of BSTs to Quicksort
Analysis of Random BST

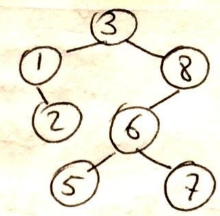


Randomly Built BSTs

BST sort (A)

$T \leftarrow \emptyset$
for $i \leftarrow 1$ to n
  do Tree-Insert (T, $A[i]$)
Inorder-Tree-Walk (root[T])

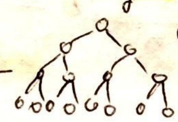
Ex 1 $A =$ | 3 | 1 | 8 | 2 | 6 | 7 | 5 |



inorder traversal 1 2 3 5 6 7 8

Time: $O(n)$ for walk
n Tree_Inserts
$\to \Omega(n \lg n)$ all the time
$\to O(n^2)$

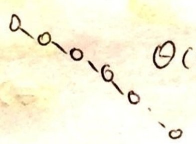
lucky



$\Theta(n \lg n)$

like Quicksort

if already sorted



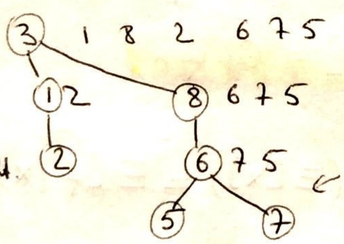
$\Theta(n^2)$

$\text{Time} = \sum_{x=1}^{n} \text{depth}(x)$

Relation to Quicksort

Comparisons that BSTsort makes are exactly the same as comparisons that quicksort makes, though in a different order.

Ex 2 (stable Partition) pivot is first elt.



same tree as in Ex 1

same comparisons as in Ex 1, but in a different order

Randomized BST Sort

① Randomly permute A ← equivalent to picking random elts as pivots in rand. Quicksort

② BSTsort (A)

Time = time (rand. Quicksort)

$E[\text{Time}] = E[\text{time (rand. Quicksort)}] = \Theta(n \lg n)$

Randomly built BST

= tree resulting from randomized BST sort, without the in-order traversal

$\text{Time (BSTsort)} = \sum_{x \in T} \text{depth}(x)$ random variables

$E[\text{Time (BSTsort)}] = \Theta(n \lg n)$

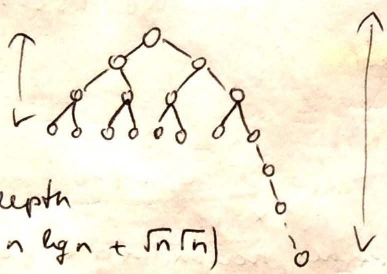
$E\left[\frac{1}{n} \sum_{x \in T} \text{depth}(x)\right] = \frac{\Theta(n \lg n)}{n} = \Theta(\lg n)$

const.

average depth in the tree

Ex 3

$\approx \lg(n - \sqrt{n})$



$\sqrt{n}$

knowing that the average depth is $\Theta(\lg n)$

$\not\Rightarrow$ height is $O(\lg n)$

average depth

$\leq \frac{1}{n}(n \lg n + \sqrt{n}\sqrt{n})$

av. depth $= O(\lg n)$, height $= \sqrt{n}$

Theorem: $E[\text{height of rand. built BST}] = O(\lg n)$

Proof outline:

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

② Instead of analyzing X_n = r.v. of height of BST on n nodes, analyze $Y_n = 2^{X_n}$

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

(4) Conclude that

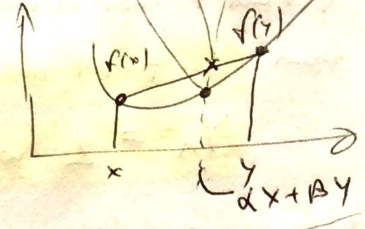
$$2^{E[X_n]} \le E[2^{X_n}] = E[Y_n] = O(n^3)$$

$$\Rightarrow E[X_n] \le \lg O(n^3) = 3\lg n + O(1)$$

linear interpolation

$\alpha f(x) + \beta f(y)$

$f(\alpha x + \beta y)$

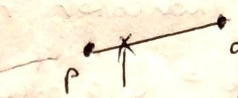


(1) $f: \mathbb{R} \to \mathbb{R}$ is <u>convex</u> if

for all $x, y \in \mathbb{R}$

and all $\alpha, \beta \ge 0$, $\alpha + \beta = 1$

$$f(\alpha x + \beta y) \le \alpha f(x) + \beta f(y)$$



lin. comb $\alpha p + \beta q$, a) $\alpha + \beta = 1$

b) $\alpha, \beta \ge 0$

only a) $\Rightarrow$ entire line

a) and b) only the p-q segment

<u>Lemma</u>: if $f: \mathbb{R} \to \mathbb{R}$ is convex,

and $x_1 \ldots\ldots x_n \in \mathbb{R}$

and $\alpha_1 \ldots\ldots \alpha_n \ge 0$

and $\sum_{k=1}^{n} \alpha_k = 1$

then $f\left(\sum_{k=1}^{n} \alpha_k x_k\right) \le \sum_{k=1}^{n} \alpha_k f(x_k)$

generalization for $n > 2$

<u>Proof</u>: Induction on n

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

$n = 2$ by def. of convexity

<u>Ind. Step</u>

$$f\left(\sum_{k=1}^{n} \alpha_n x_k\right) = f\Big(\underbrace{\alpha_n x_n + (1-\alpha_n)}_{\text{sum to } 1} \underbrace{\sum_{k=1}^{n-1} \frac{\alpha_k}{(1-\alpha_n)}}_{\text{sum up to } 1} x_k\Big)$$

$$\overset{\text{by convexity}}{\le} \alpha_n f(x_n) + (1-\alpha_n) f\left(\sum_{k=1}^{n-1} \frac{\alpha_k}{1-\alpha_k} x_k\right)$$

$$\overset{\text{by IH}}{\le} \alpha_n f(x_n) + (1-\alpha_n) \sum_{k=1}^{n-1} \frac{\alpha_k}{1-\alpha_k} f(x_k)$$

$$= \sum_{k=1}^{n} \alpha_k f(x_k) \checkmark$$

Jensen's inequality

$f(E[X]) \le E[f(X)]$, if f convex, X is integer r.v.

Proof: $f(E[X]) = f\Big(\sum_{x=-\infty}^{\infty} x \underbrace{\Pr\{X = x\}}_{\text{sum to } 1}\Big) \underset{\text{lemma}}{\le} \sum_{x=-\infty}^{\infty} \Pr\{X = x\} f(x)$

reclustering of sum

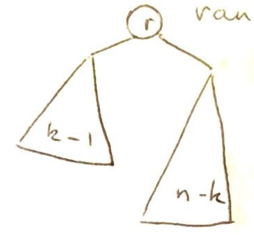
$$= \sum_{y \in \text{range}(f)} y \underbrace{\sum_{x: f(x) = y} \Pr\{X = x\}}_{\Pr\{f(X) = y\}} = E[f(X)] \checkmark$$

Expected BST height analysis

X_n = r.v. of height of randomly built BST on n nodes.

$Y_n = 2^{X_n}$ $\qquad$ 2^x is convex

if root r has rank k



rank k ← index in the sorted order

then $X_n = 1 + \max\{X_{k-1}, X_{n-k}\}$

$Y_n = 2\max\{Y_{k-1}, Y_{n-k}\}$ ← better for recurrence analysis: 2 subproblem

define indicator r.v.s

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

$$\Pr\{Z_{nk}=1\} = E[Z_{nk}] = \frac{1}{n}$$

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

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

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

$$= 2\sum_{k=1}^{n} \underbrace{E[Z_{nk}]}_{1/n} E\left[\max\{Y_{k-1}, Y_{n-k}\}\right] \quad \text{independence}$$

← a bit loose

$$\le \frac{2}{n}\sum_{k=1}^{n} E[Y_{k-1} + Y_{n-k}] = \frac{2}{n}\sum_{k=1}^{n} E[Y_{k-1}] + E[Y_{n-k}]$$

↑ linearity

$$= \frac{4}{n}\sum_{k=0}^{n-1} E[Y_k]$$

Claim: $E[Y_n] \le cn^3$

Proof Substitution, Base $n = \Theta(1)$, true if c is sufficiently large

Inductive Step: $E[Y_n] \le \frac{4}{n}\sum_{k=0}^{n-1} E[Y_k] \le \frac{4}{n}\sum_{k=0}^{n-1} ck^3$ by IH

$$\le \frac{4c}{n}\int_0^n x^3\,dx = \frac{4c}{n}\frac{n^4}{4} = cn^3 \checkmark$$

↑ approximate by integral

$E[X_n] \le \lg[cn^3] = 3\lg n + O(1)$ ← very tight bound

$E[X_n] \approx 2.9882 \cdot \lg n$ [Devroye 1986] ←

$\boxed{\begin{array}{l}\max(a,b) \le a+b \text{ works} \\ \max(2^a,2^b) \le 2^a+2^b \\ \text{↖ better!!}\end{array}}$