Lec 09 二叉搜索树

(Randomly Built) Binary Search Trees, BSTs for short

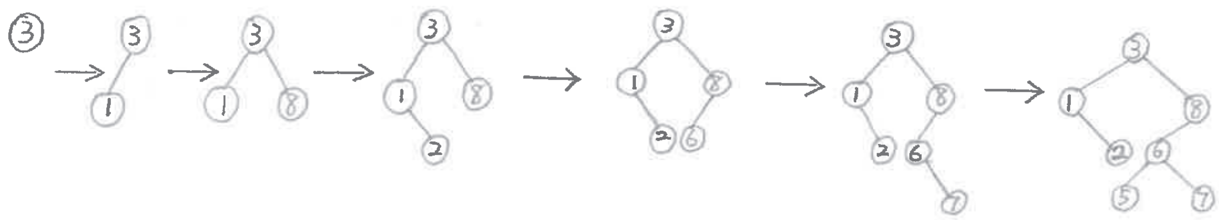
△ BST sort (A):

// build the BST and then traverse it in order

```
T ← ∅
for i ← 1 to n
    do Tree-Insert (T, A[i])
Inorder-Tree-Walk (T.root)
```

Example:

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



Time:

$O(n)$ for walk,

$\Omega(n\lg n)$ for n Tree-Inserts, meanwhile, $O(n^2)$ for n Tree-Inserts

best case is a perfectly balanced tree

worst case is the array is already sorted/reverse-sorted

if already sorted/reverse-sorted, then it's a bad shape!

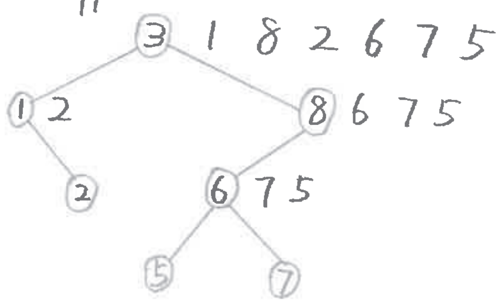
if lucky, it is a balanced tree with $O(\lg n)$ height $\Rightarrow$ $O(n\lg n)$ time

Quicksort?

it turns out the running time of this algorithm is the same as the running time of quicksort.

Relation to Quicksort

BST sort and Quicksort make the same comparisons, but in a different order.



△ Randomized BST sort

① randomly permuted the array A

② BST sort(A)

Time = time(randomized Quicksort), i.e.

$E[\text{time(randomized BST sort)}] = E[\text{time(randomized Quicksort)}] = \Theta(n\lg n)$

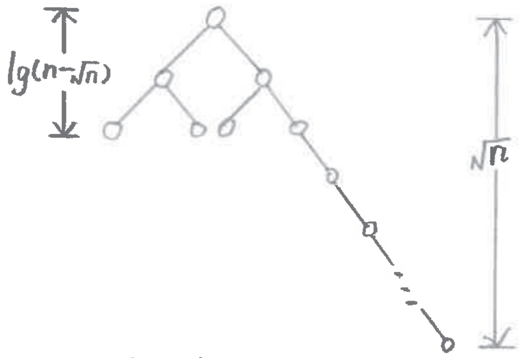
randomly built BST = the tree resulted from randomized BST sort

$\text{time(BST sort)} = \sum_{node} \text{depth(node)}$ ⟵ "所有结点的深度的和"

$\Rightarrow E(\text{BST sort}) = \Theta(n\lg n)$

$E\left[\frac{1}{n}\sum_{node} \text{depth(node)}\right] = \Theta(n\lg n)/n = \Theta(\lg n)$ ⟵ "树结点的平均深度" 并不意味着树的高度就是lgn

Example:



$\text{avg-depth} \le \frac{1}{n}(n\lg n + \sqrt{n}\cdot\sqrt{n}) = O(\lg n)$

说明: 只知道平均深度是 lgn 的话,并不代表树的高度就是 lgn

Theorem: $E(\text{height of randomized built BST}) = O(\lg n)$ "n个结点"

proof outline:

① prove Jensen's inequality $f[E(X)] \le E[f(X)]$ for convex function f "凹函数"

② instead of analyzing X_n = r.v. of height of BST on n nodes, "random variable"

analyze $Y_n = 2^{X_n}$

③ prove that $E(Y_n) = O(n^3)$

④ conclude that

$$\begin{cases} E(2^{X_n}) = E(Y_n) = O(n^3) \\ 2^{E(X_n)} \le E(2^{X_n}) \end{cases}$$

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

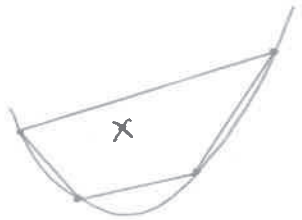
proof: ① $f: R \to R$ is convex if for all x, y and all $\alpha, \beta \geq 0$, $\alpha+\beta=1$,

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

Lemma if $f: R\to R$ is convex,

$x_1, x_2, \cdots, x_n \in R$, $\alpha_1, \alpha_2, \cdots, \alpha_n \geq 0$ with $\alpha_1+\alpha_2+\cdots+\alpha_n=1$,

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



proof: (induction)

base $n=1$, $f(x_1) \leq f(x_2)$

step

$$\begin{aligned}
& f\left(\sum_{k=1}^{n}\alpha_k x_k\right) \\
&= f\left(\alpha_n x_n+\sum_{k=1}^{n-1}\alpha_k x_k\right) \\
&= f\left(\alpha_n x_n+(1-\alpha_n)\sum_{k=1}^{n-1}\frac{\alpha_k}{1-\alpha_n}x_k\right) \\
&\leq \alpha_n f(x_n)+(1-\alpha_n)f\left(\sum_{k=1}^{n-1}\frac{\alpha_k}{1-\alpha_n}x_k\right) && \leftarrow \text{"凹函数的定义"/n=2 case} \\
&\leq \alpha_n f(x_n)+(1-\alpha_n)\sum_{k=1}^{n-1}\frac{\alpha_k}{1-\alpha_n}f(x_k) && \leftarrow \text{induction hypothesis} \\
&= \alpha_n f(x_n)+\sum_{k=1}^{n-1}\alpha_k f(x_k) \\
&= \sum_{k=1}^{n}\alpha_k f(x_k) \quad \text{Q.E.D.}
\end{aligned}$$

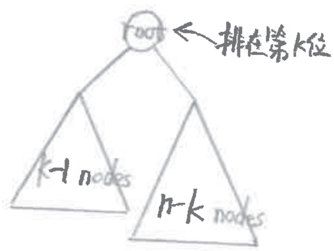
next, to prove Jensen's inequality, suppose x is an integer

$$\begin{aligned}
f[E(X)] &= f\Big(\sum_{x=-\infty}^{+\infty} x\cdot \underbrace{P(X=x)}_{\text{"sum to 1"}}\Big) \\
&\leq \sum_{x=-\infty}^{+\infty} P(X=x)\cdot f(x) && \leftarrow \text{by Lemma} \\
&= E[f(X)]
\end{aligned}$$

② expected BST height analysis

X_n = random variable of height of a randomly built BST on n nodes

$Y_n = 2^{X_n}$ ($y=2^x$ is a convex function)



if root r has rank k,

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

$\quad Y_n = 2\max\{Y_{k-1}, Y_{n-k}\}$

define indicator random variables,

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

$$P(Z_{nk}=1) = E(Z_{nk}) = \frac{1}{n}$$

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

$$\begin{aligned}
E(Y_n) &= E\left[\sum_{k=1}^{n} Z_{nk} \cdot \left(2\max\{Y_{k-1}, Y_{n-k}\}\right)\right] \\
&= \sum_{k=1}^{n} E\left[Z_{nk} \cdot \left(2\max\{Y_{k-1}, Y_{n-k}\}\right)\right] && \longleftarrow \text{期望的线性性质} \\
&= 2\sum_{k=1}^{n} \left[\underbrace{E(Z_{nk})}_{=\frac{1}{n}} \cdot E\left(\max\{Y_{k-1}, Y_{n-k}\}\right)\right] && \longleftarrow \text{事件的独立性} \\
&= \frac{2}{n} \sum_{k=1}^{n} E\left(\max\{Y_{k-1}, Y_{n-k}\}\right) \\
&\leq \frac{2}{n} \sum_{k=1}^{n} E(Y_{k-1} + Y_{n-k}) \\
&= \frac{2}{n} \sum_{k=1}^{n} \left[E(Y_{k-1}) + E(Y_{n-k})\right] && \longleftarrow \text{期望的线性性质} \\
&= 2 \times \frac{2}{n} \sum_{k=1}^{n} E(Y_{k-1}) \\
&= \frac{4}{n} \sum_{k=0}^{n-1} E(Y_k)
\end{aligned}$$

③ (substitution method to solve the recurrence)

claim: $E(Y_k) \leq cn^3$

proof: (substitution method = induction)

base $n = \Theta(1)$ true if c is sufficiently large

step
$$\begin{aligned}
E(Y_n) &\leq \frac{4}{n} \sum_{k=0}^{n-1} E(Y_k) && \leftarrow k<n \\
&\leq \frac{4}{n} \sum_{k=0}^{n-1} c \cdot k^3 && \text{(induction hypothesis)} \\
&\leq \frac{4c}{n} \int_0^n x^3\,dx \\
&= c \cdot n^3
\end{aligned}$$

so $E(Y_k) = O(n^3)$

④ 略