

Design and Analysis of Algorithms

CSE 5311

Lecture 9 Randomly Built Binary Search Trees.

Song Jiang, Ph.D.

Department of Computer Science and Engineering



Binary-search-tree sort

$T \leftarrow \emptyset$ \triangleright Create an empty BST

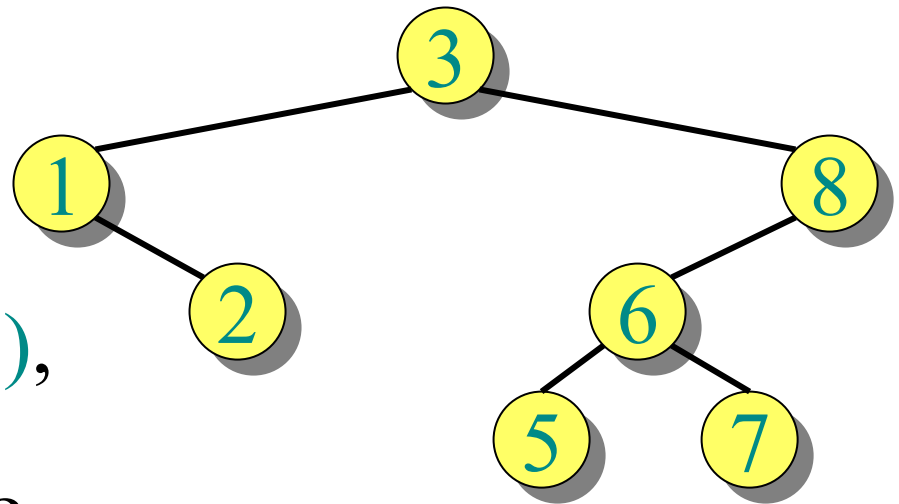
for $i = 1$ to n

do TREE-INSERT($T, A[i]$)

Perform an inorder tree walk of T .

Example:

$A = [3 \ 1 \ 8 \ 2 \ 6 \ 7 \ 5]$

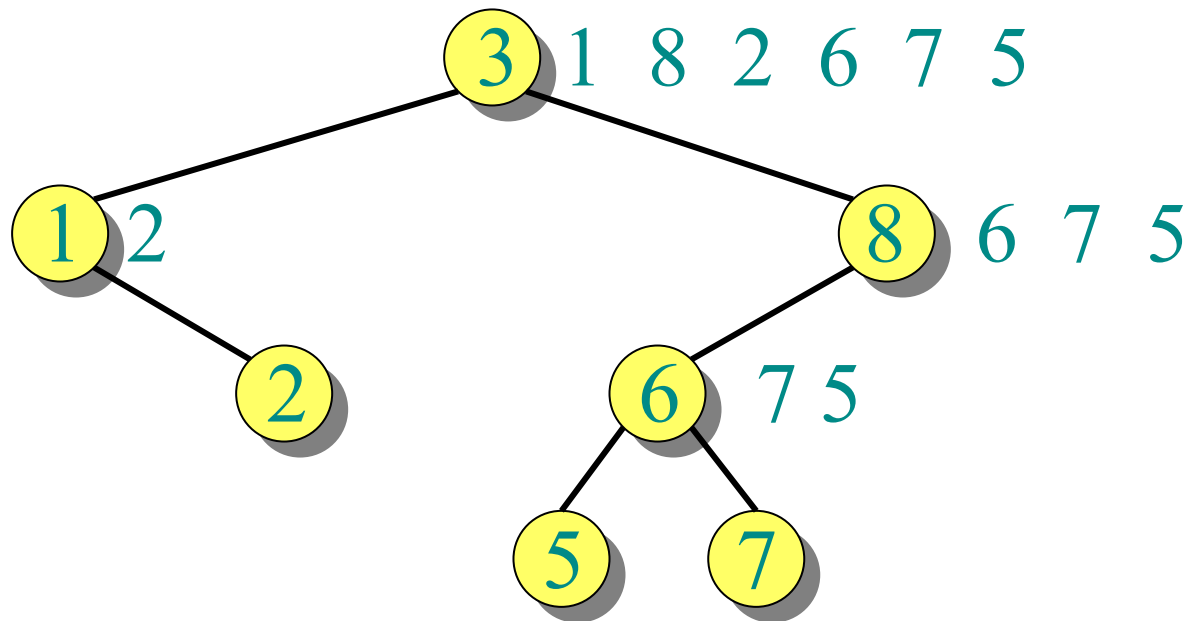


Tree-walk time = $O(n)$,
but how long does it
take to build the BST?



Analysis of BST sort

BST sort performs the same comparisons as quicksort, but in a different order!



The expected time to build the tree is asymptotically the same as the running time of quicksort.



Node depth

The depth of a node = the number of comparisons made during TREE-INSERT. Assuming all input permutations are equally likely, we have

Average node depth

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

$$= \frac{1}{n} O(n \lg n) \quad (\text{quicksort analysis})$$

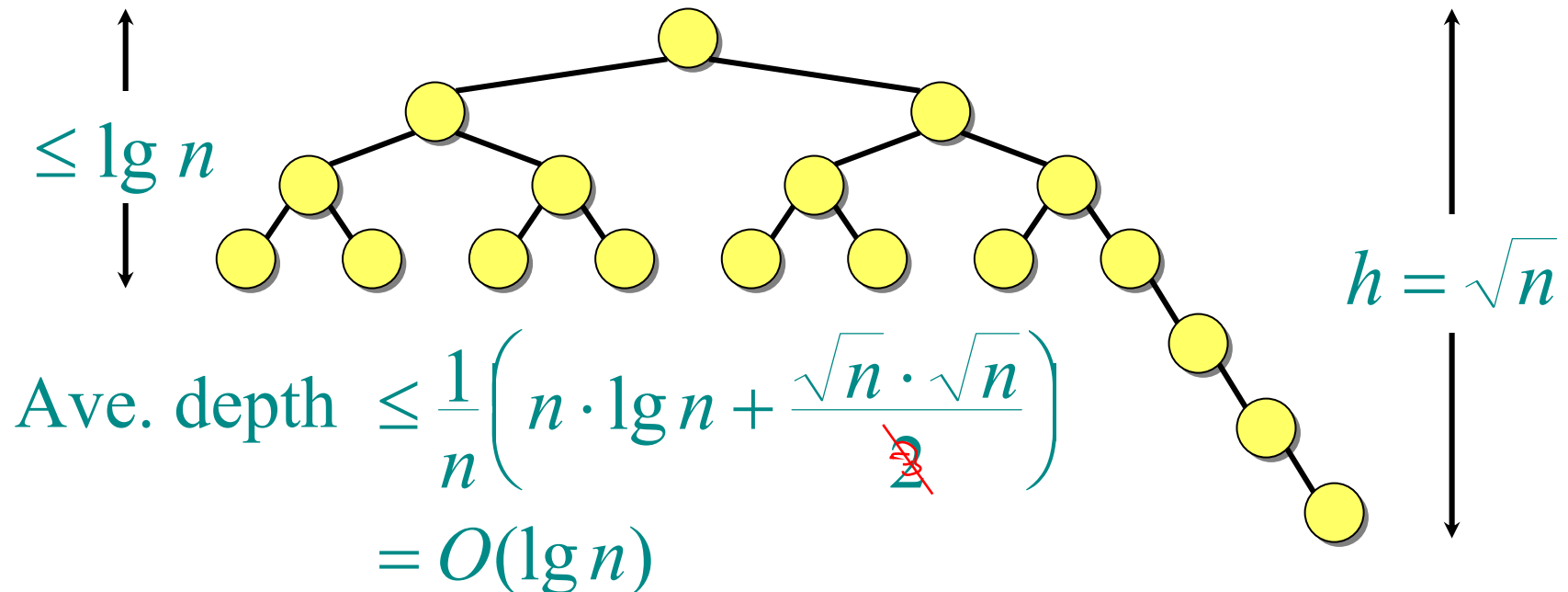
$$= O(\lg n) .$$

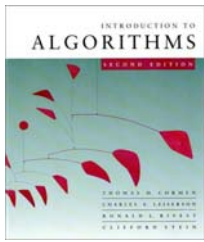


Expected tree height

But, average node depth of a randomly built BST = $O(\lg n)$ does not necessarily mean that its expected height is also $O(\lg n)$ (although it is).

Example.





Height of a randomly built binary search tree

Outline of the analysis:

- Prove *Jensen's inequality*, which says that $f(E[X]) \leq E[f(X)]$ for any convex function f and random variable X .
- Analyze the *exponential height* of a randomly built BST on n nodes, which is the random variable $Y_n = 2^{X_n}$, where X_n is the random variable denoting the height of the BST.
- Prove that $2^{E[X_n]} \leq E[2^{X_n}] = E[Y_n] = O(n^3)$, and hence that $E[X_n] = O(\lg n)$.

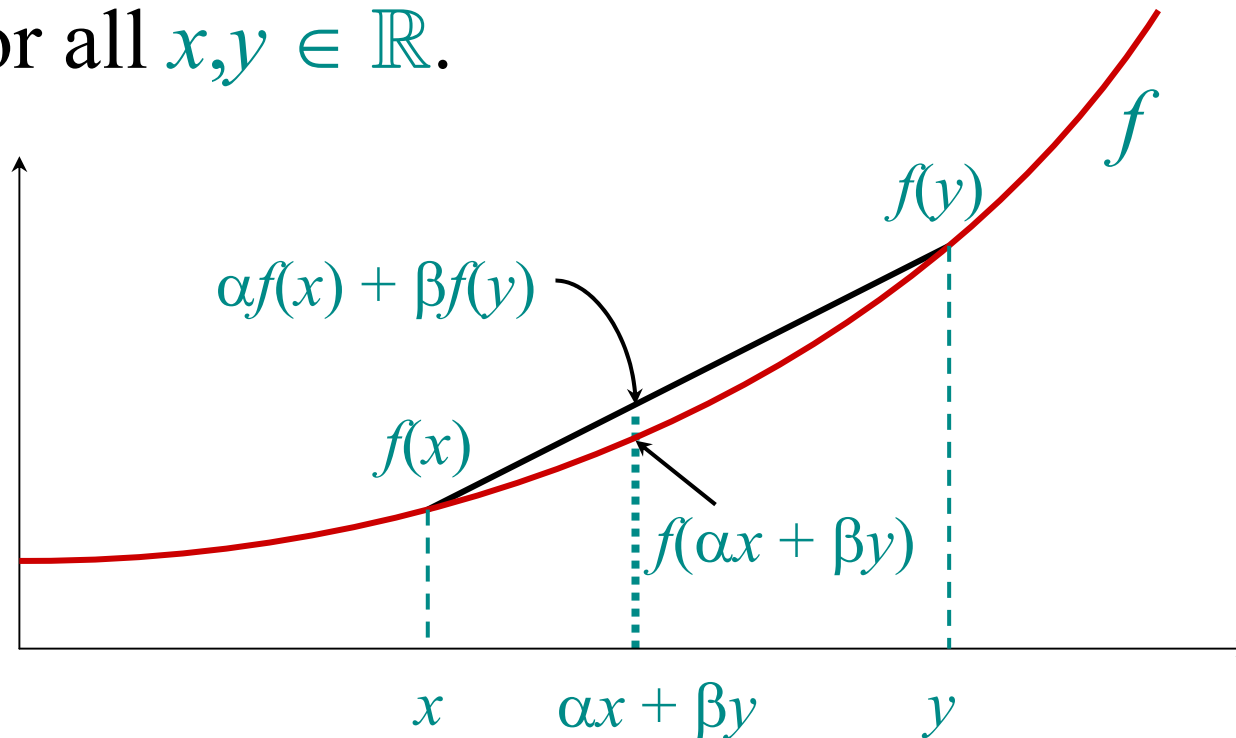


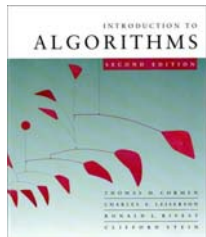
Convex functions

A function $f: \mathbb{R} \rightarrow \mathbb{R}$ is **convex** if for all $\alpha, \beta \geq 0$ such that $\alpha + \beta = 1$, we have

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

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





Convexity lemma

Lemma. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be a convex function, and let $\alpha_1, \alpha_2, \dots, \alpha_n$ be nonnegative real numbers such that $\sum_k \alpha_k = 1$. Then, for any real numbers x_1, x_2, \dots, x_n , we have

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

Proof. By induction on n . For $n = 1$, we have $\alpha_1 = 1$, and hence $f(\alpha_1 x_1) \leq \alpha_1 f(x_1)$ trivially.



Proof (continued)

Inductive step:

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

Algebra.



Proof (continued)

Inductive step:

$$\begin{aligned} f\left(\sum_{k=1}^n \alpha_k x_k\right) &= f\left(\alpha_n x_n + (1 - \alpha_n) \sum_{k=1}^{n-1} \frac{\alpha_k}{1 - \alpha_n} x_k\right) \\ &\leq \alpha_n f(x_n) + (1 - \alpha_n) f\left(\sum_{k=1}^{n-1} \frac{\alpha_k}{1 - \alpha_n} x_k\right) \end{aligned}$$

Convexity.

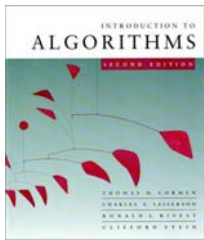


Proof (continued)

Inductive step:

$$\begin{aligned} f\left(\sum_{k=1}^n \alpha_k x_k\right) &= f\left(\alpha_n x_n + (1 - \alpha_n) \sum_{k=1}^{n-1} \frac{\alpha_k}{1 - \alpha_n} x_k\right) \\ &\leq \alpha_n f(x_n) + (1 - \alpha_n) f\left(\sum_{k=1}^{n-1} \frac{\alpha_k}{1 - \alpha_n} x_k\right) \\ &\leq \alpha_n f(x_n) + (1 - \alpha_n) \sum_{k=1}^{n-1} \frac{\alpha_k}{1 - \alpha_n} f(x_k) \end{aligned}$$

Induction.

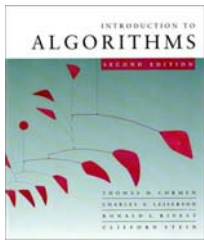


Proof (continued)

Inductive step:

$$\begin{aligned} f\left(\sum_{k=1}^n \alpha_k x_k\right) &= f\left(\alpha_n x_n + (1 - \alpha_n) \sum_{k=1}^{n-1} \frac{\alpha_k}{1 - \alpha_n} x_k\right) \\ &\leq \alpha_n f(x_n) + (1 - \alpha_n) f\left(\sum_{k=1}^{n-1} \frac{\alpha_k}{1 - \alpha_n} x_k\right) \\ &\leq \alpha_n f(x_n) + (1 - \alpha_n) \sum_{k=1}^{n-1} \frac{\alpha_k}{1 - \alpha_n} f(x_k) \\ &= \sum_{k=1}^n \alpha_k f(x_k). \quad \square \end{aligned}$$

Algebra.

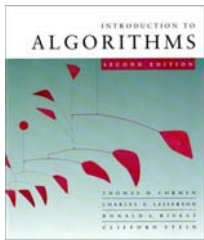


Convexity lemma: infinite case

Lemma. Let $f: \mathbb{R} \rightarrow \mathbb{R}$ be a convex function, and let $\alpha_1, \alpha_2, \dots$, be nonnegative real numbers such that $\sum_k \alpha_k = 1$. Then, for any real numbers x_1, x_2, \dots , we have

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

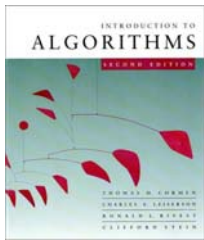
assuming that these summations exist.



Convexity lemma: infinite case

Proof. By the convexity lemma, for any $n \geq 1$,

$$f\left(\frac{\sum_{k=1}^n \alpha_k x_k}{\sum_{i=1}^n \alpha_i}\right) \leq \frac{\sum_{k=1}^n \alpha_k f(x_k)}{\sum_{i=1}^n \alpha_i}.$$



Convexity lemma: infinite case

Proof. By the convexity lemma, for any $n \geq 1$,

$$f\left(\frac{\sum_{k=1}^n \alpha_k x_k}{\sum_{i=1}^n \alpha_i}\right) \leq \frac{\sum_{k=1}^n \alpha_k f(x_k)}{\sum_{i=1}^n \alpha_i}.$$

Taking the limit of both sides
(and because the inequality is not strict):

$$\lim_{n \rightarrow \infty} f\left(\underbrace{\frac{1}{\sum_{i=1}^n \alpha_i}}_{\rightarrow 1} \underbrace{\sum_{k=1}^n \alpha_k x_k}_{\rightarrow \sum_{k=1}^{\infty} \alpha_k x_k}\right) \leq \lim_{n \rightarrow \infty} \underbrace{\frac{1}{\sum_{i=1}^n \alpha_i}}_{\rightarrow 1} \underbrace{\sum_{k=1}^n \alpha_k f(x_k)}_{\rightarrow \sum_{k=1}^{\infty} \alpha_k f(x_k)}$$





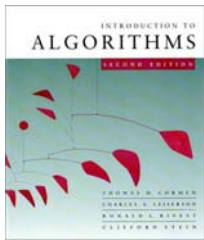
Jensen's inequality

Lemma. Let f be a convex function, and let X be a random variable. Then, $f(E[X]) \leq E[f(X)]$.

Proof.

$$f(E[X]) = f\left(\sum_{k=-\infty}^{\infty} k \cdot \Pr\{X = k\}\right)$$

Definition of expectation.



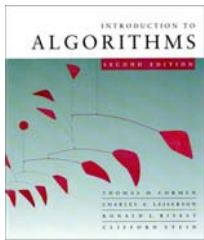
Jensen's inequality

Lemma. Let f be a convex function, and let X be a random variable. Then, $f(E[X]) \leq E[f(X)]$.

Proof.

$$\begin{aligned} f(E[X]) &= f\left(\sum_{k=-\infty}^{\infty} k \cdot \Pr\{X = k\}\right) \\ &\leq \sum_{k=-\infty}^{\infty} f(k) \cdot \Pr\{X = k\} \end{aligned}$$

Convexity lemma (infinite case).



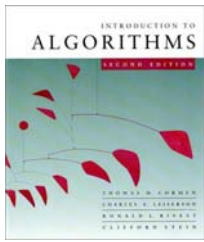
Jensen's inequality

Lemma. Let f be a convex function, and let X be a random variable. Then, $f(E[X]) \leq E[f(X)]$.

Proof.

$$\begin{aligned} f(E[X]) &= f\left(\sum_{k=-\infty}^{\infty} k \cdot \Pr\{X = k\}\right) \\ &\leq \sum_{k=-\infty}^{\infty} f(k) \cdot \Pr\{X = k\} \\ &= E[f(X)]. \quad \square \end{aligned}$$

Tricky step, but true—think about it.



Analysis of BST height

Let X_n be the random variable denoting the height of a randomly built binary search tree on n nodes, and let $Y_n = 2^{X_n}$ be its exponential height.

If the root of the tree has rank k , then

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

since each of the left and right subtrees of the root are randomly built. Hence, we have

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



Analysis (continued)

Define the indicator random variable Z_{nk} as

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

Thus, $\Pr\{Z_{nk} = 1\} = E[Z_{nk}] = 1/n$, and

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



Exponential height recurrence

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

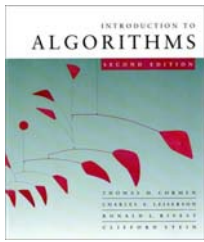
Take expectation of both sides.



Exponential height recurrence

$$\begin{aligned} E[Y_n] &= E\left[\sum_{k=1}^n Z_{nk} (2 \cdot \max\{Y_{k-1}, Y_{n-k}\})\right] \\ &= \sum_{k=1}^n E[Z_{nk} (2 \cdot \max\{Y_{k-1}, Y_{n-k}\})] \end{aligned}$$

Linearity of expectation.



Exponential height recurrence

$$\begin{aligned} E[Y_n] &= E \left[\sum_{k=1}^n Z_{nk} (2 \cdot \max \{Y_{k-1}, Y_{n-k}\}) \right] \\ &= \sum_{k=1}^n E[Z_{nk} (2 \cdot \max \{Y_{k-1}, Y_{n-k}\})] \\ &= 2 \sum_{k=1}^n E[Z_{nk}] \cdot E[\max \{Y_{k-1}, Y_{n-k}\}] \end{aligned}$$

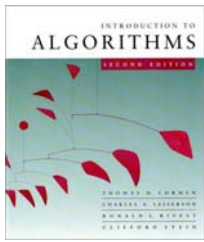
Independence of the rank of the root
from the ranks of subtree roots.



Exponential height recurrence

$$\begin{aligned} E[Y_n] &= E \left[\sum_{k=1}^n Z_{nk} (2 \cdot \max \{Y_{k-1}, Y_{n-k}\}) \right] \\ &= \sum_{k=1}^n E[Z_{nk} (2 \cdot \max \{Y_{k-1}, Y_{n-k}\})] \\ &= 2 \sum_{k=1}^n E[Z_{nk}] \cdot E[\max \{Y_{k-1}, Y_{n-k}\}] \\ &\leq \frac{2}{n} \sum_{k=1}^n E[Y_{k-1} + Y_{n-k}] \end{aligned}$$

The max of two nonnegative numbers is at most their sum, and $E[Z_{nk}] = 1/n$.



Exponential height recurrence

$$\begin{aligned} E[Y_n] &= E\left[\sum_{k=1}^n Z_{nk} (2 \cdot \max\{Y_{k-1}, Y_{n-k}\})\right] \\ &= \sum_{k=1}^n E[Z_{nk} (2 \cdot \max\{Y_{k-1}, Y_{n-k}\})] \\ &= 2 \sum_{k=1}^n E[Z_{nk}] \cdot E[\max\{Y_{k-1}, Y_{n-k}\}] \\ &\leq \frac{2}{n} \sum_{k=1}^n E[Y_{k-1} + Y_{n-k}] \\ &= \frac{4}{n} \sum_{k=0}^{n-1} E[Y_k] \end{aligned}$$

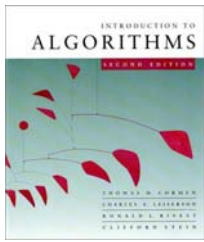
Each term appears twice, and reindex.



Solving the recurrence

Use substitution to show that $E[Y_n] \leq cn^3$ for some positive constant c , which we can pick sufficiently large to handle the initial conditions.

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



Solving the recurrence

Use substitution to show that $E[Y_n] \leq cn^3$ for some positive constant c , which we can pick sufficiently large to handle the initial conditions.

$$\begin{aligned} E[Y_n] &= \frac{4}{n} \sum_{k=0}^{n-1} E[Y_k] \\ &\leq \frac{4}{n} \sum_{k=0}^{n-1} ck^3 \end{aligned}$$

Substitution.

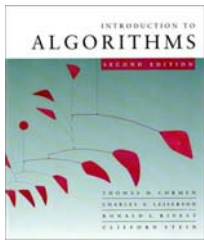


Solving the recurrence

Use substitution to show that $E[Y_n] \leq cn^3$ for some positive constant c , which we can pick sufficiently large to handle the initial conditions.

$$\begin{aligned} E[Y_n] &= \frac{4}{n} \sum_{k=0}^{n-1} E[Y_k] \\ &\leq \frac{4}{n} \sum_{k=0}^{n-1} ck^3 \\ &\leq \frac{4c}{n} \int_0^n x^3 dx \end{aligned}$$

Integral method.

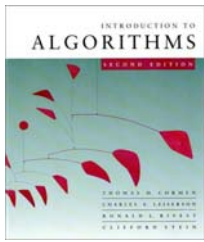


Solving the recurrence

Use substitution to show that $E[Y_n] \leq cn^3$ for some positive constant c , which we can pick sufficiently large to handle the initial conditions.

$$\begin{aligned} E[Y_n] &= \frac{4}{n} \sum_{k=0}^{n-1} E[Y_k] \\ &\leq \frac{4}{n} \sum_{k=0}^{n-1} ck^3 \\ &\leq \frac{4c}{n} \int_0^n x^3 dx \\ &= \frac{4c}{n} \left(\frac{n^4}{4} \right) \end{aligned}$$

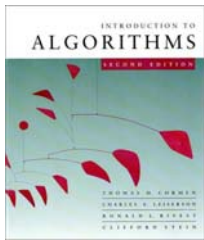
Solve the integral.



Solving the recurrence

Use substitution to show that $E[Y_n] \leq cn^3$ for some positive constant c , which we can pick sufficiently large to handle the initial conditions.

$$\begin{aligned} E[Y_n] &= \frac{4}{n} \sum_{k=0}^{n-1} E[Y_k] \\ &\leq \frac{4}{n} \sum_{k=0}^{n-1} ck^3 \\ &\leq \frac{4c}{n} \int_0^n x^3 dx \\ &= \frac{4c}{n} \left(\frac{n^4}{4} \right) \\ &= cn^3. \quad \text{Algebra.} \end{aligned}$$



The grand finale

Putting it all together, we have

$$2^{E[X_n]} \leq E[2^{X_n}]$$

Jensen's inequality, since $f(x) = 2^x$ is convex.



The grand finale

Putting it all together, we have

$$\begin{aligned} 2^{E[X_n]} &\leq E[2^{X_n}] \\ &= E[Y_n] \end{aligned}$$

Definition.

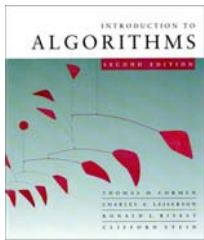


The grand finale

Putting it all together, we have

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

What we just showed.



The grand finale

Putting it all together, we have

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

Taking the \lg of both sides yields

$$E[X_n] \leq 3 \lg n + O(1).$$

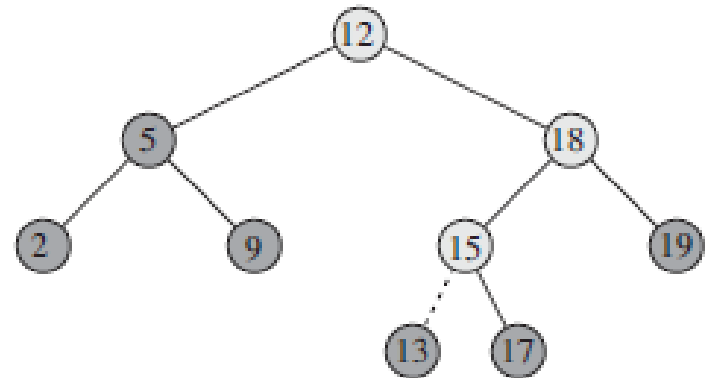
After-class Reading

- Read the first three sections of Chapter 12 (Page 286)

12.1 What is a binary search tree?

INORDER-TREE-WALK(x)

```
1  if  $x \neq \text{NIL}$ 
2      INORDER-TREE-WALK( $x.\text{left}$ )
3      print  $x.\text{key}$ 
4      INORDER-TREE-WALK( $x.\text{right}$ )
```

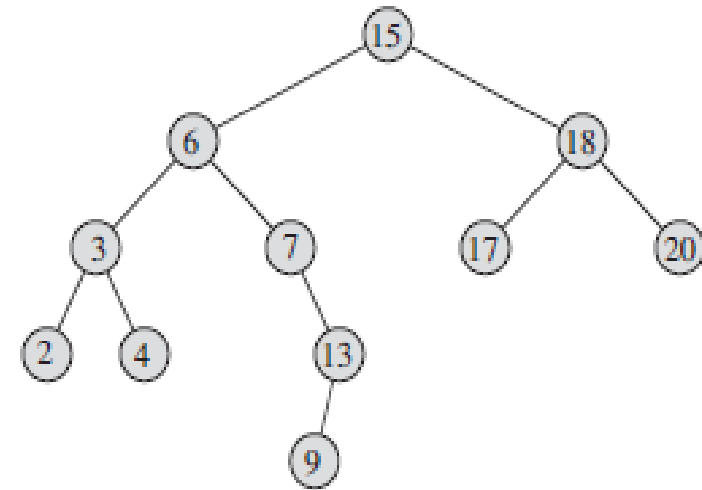


After-class Reading

12.2 Querying a binary search tree

TREE-SUCCESSOR(x)

```
1  if  $x.right \neq \text{NIL}$ 
2      return TREE-MINIMUM( $x.right$ )
3   $y = x.p$ 
4  while  $y \neq \text{NIL}$  and  $x == y.right$ 
5       $x = y$ 
6       $y = y.p$ 
7  return  $y$ 
```



After-class Reading

12.3 Insertion and deletion

TREE-INSERT(T, z)

```
1   $y = \text{NIL}$ 
2   $x = T.\text{root}$ 
3  while  $x \neq \text{NIL}$ 
4       $y = x$ 
5      if  $z.\text{key} < x.\text{key}$ 
6           $x = x.\text{left}$ 
7      else  $x = x.\text{right}$ 
8   $z.p = y$ 
9  if  $y == \text{NIL}$ 
10      $T.\text{root} = z$       // tree  $T$  was empty
11  elseif  $z.\text{key} < y.\text{key}$ 
12      $y.\text{left} = z$ 
13  else  $y.\text{right} = z$ 
```

