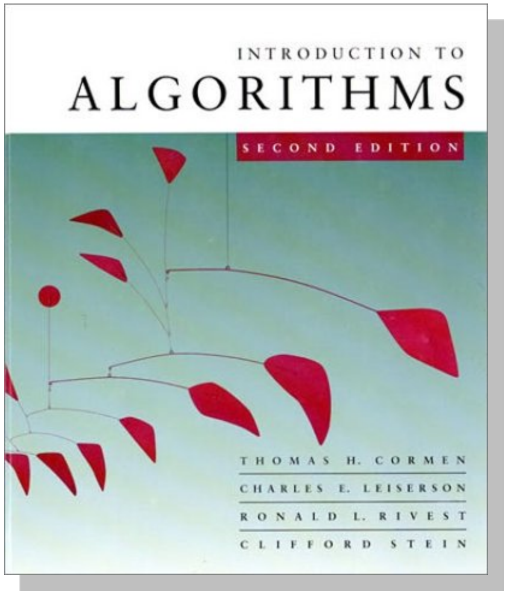
Introduction to Algorithms

6.046J/18.401J

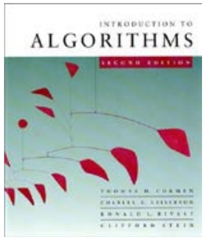


LECTURE 9

Randomly built binary search trees

- Expected node depth
- Analyzing height
  - Convexity lemma
  - Jensen's inequality
  - Exponential height
- Post mortem

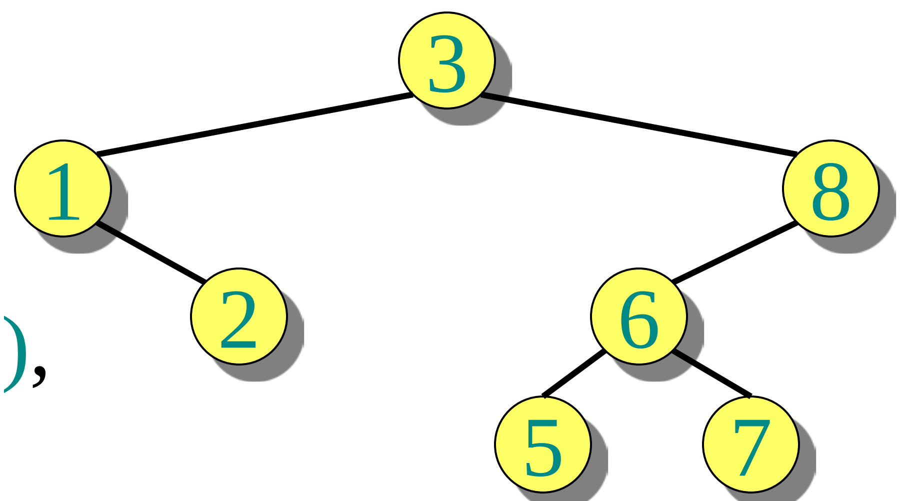
Prof. Erik Demaine



Binary-search-tree sort

$T \leftarrow \emptyset$ ⊳ Create an empty BST
for $i = 1$ to n
  do TREE-INSERT$(T, A[i])$
Perform an inorder tree walk of T.

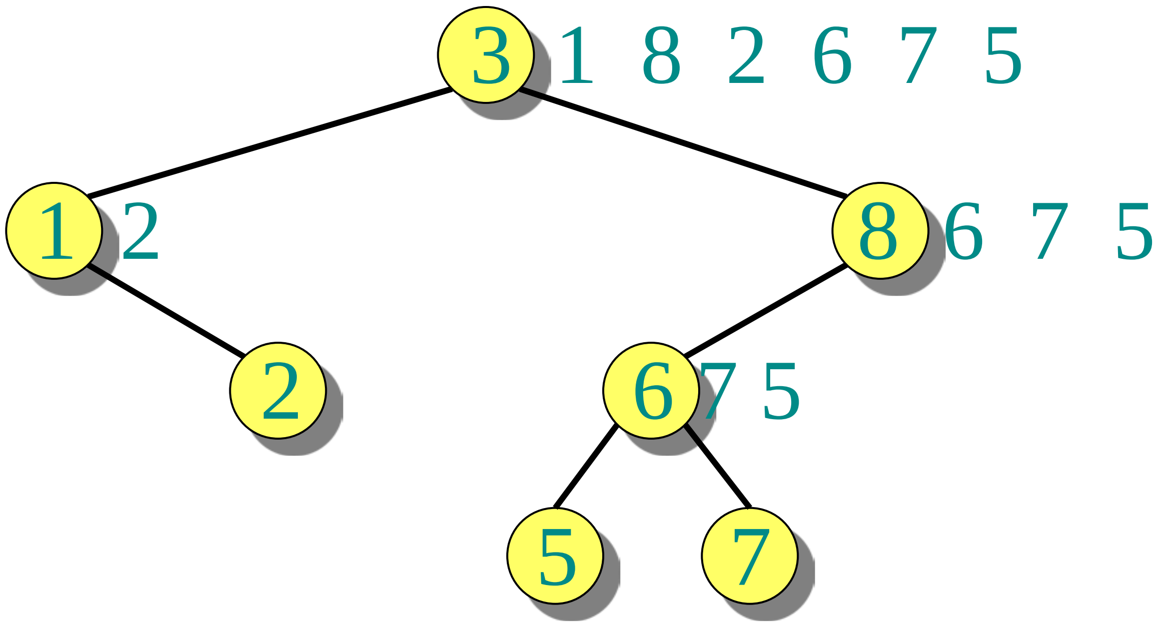
Example:
$A = [3\ 1\ 8\ 2\ 6\ 7\ 5]$



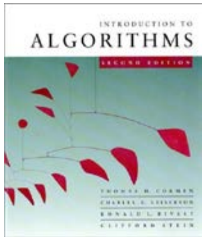
Tree-walk time $= O(n)$, but how long does it take to build the BST?

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

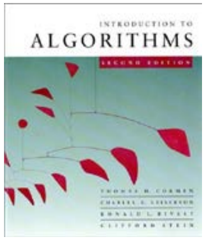
The depth of a node = the number of comparisons made during TREE-INSERT. Assuming all input permutations are equally likely, we have

Average node depth

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

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

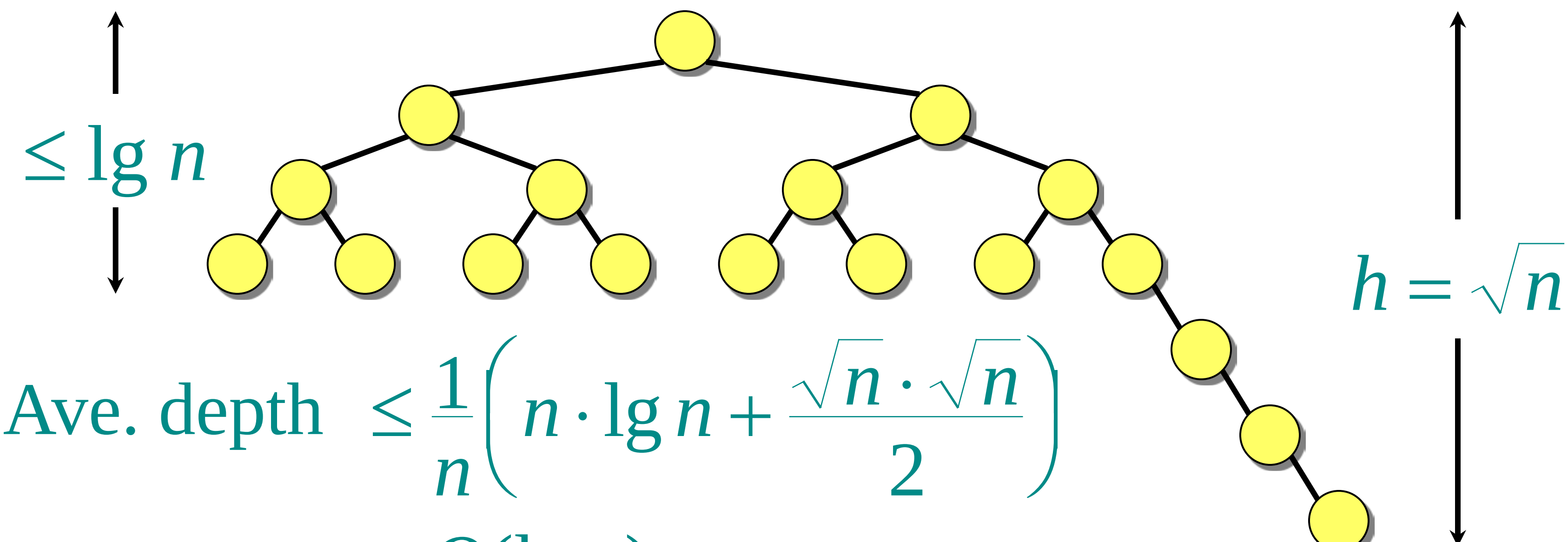
$$= O(\lg n)\ .$$



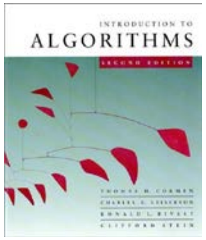
Expected tree height

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.



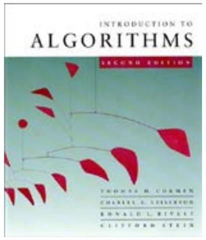
$$\text{Ave. depth} \le \frac{1}{n}\left(n \cdot \lg n + \frac{\sqrt{n} \cdot \sqrt{n}}{2}\right)$$
$$= O(\lg n)$$



Height of a randomly built binary search tree

Outline of the analysis:

- Prove ***Jensen's inequality***, which says that $f(E[X]) \le E[f(X)]$ 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 $2^{E[X_n]} \le E[2^{X_n}] = E[Y_n] = O(n^3)$, and hence that $E[X_n] = O(\lg n)$.

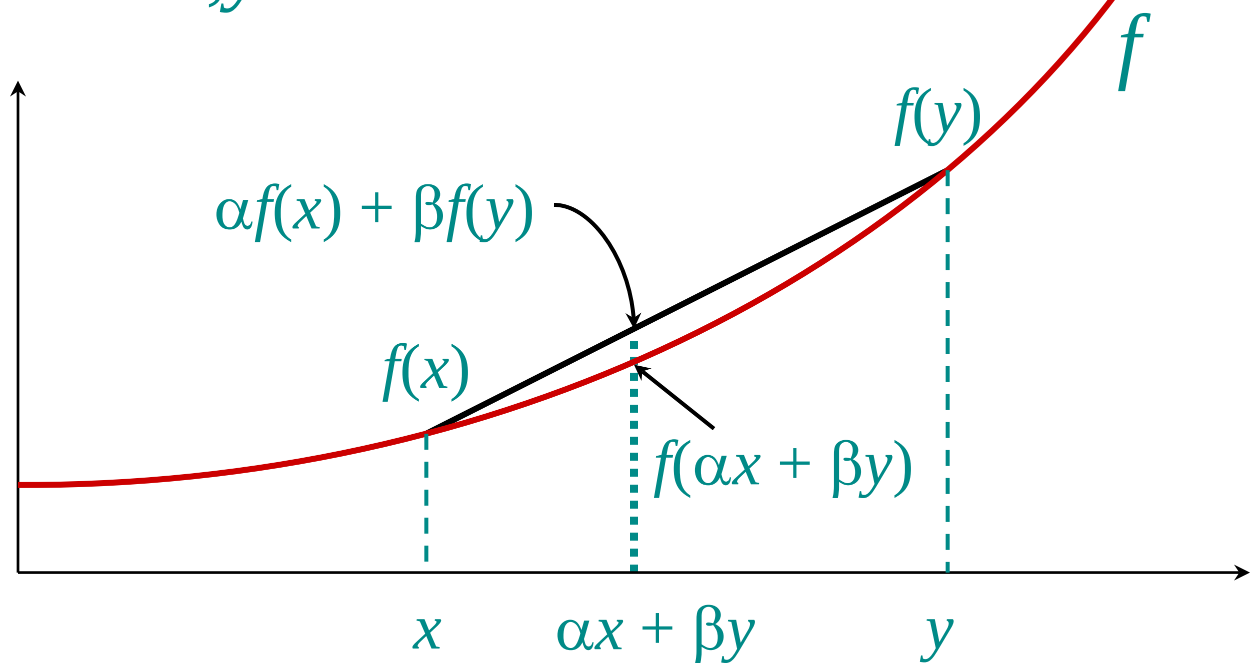


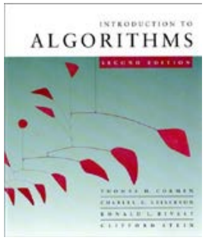
Convex functions

A function $f : \mathsf{R} \rightarrow \mathsf{R}$ 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 $x,y \in \mathsf{R}$.



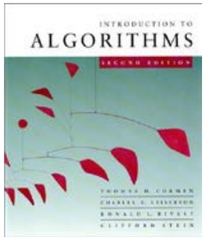


Convexity lemma

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

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

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

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

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

Convexity.

Proof (continued)

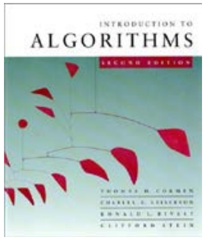
Inductive step:

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

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

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

Induction.

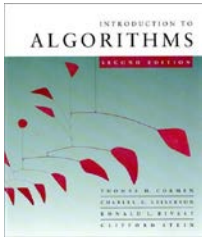


Proof (continued)

Inductive step:

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

Algebra.

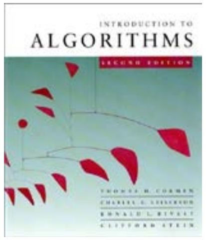


Convexity lemma: infinite case

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

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

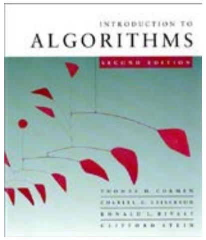
assuming that these summations exist.



Convexity lemma: infinite case

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

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



Convexity lemma: infinite case

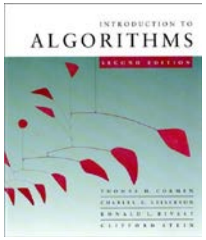
Proof. By the convexity lemma, for any $n \geq 1$,

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

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

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

$\blacksquare$



Jensen's inequality

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

Proof.

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

Definition of expectation.

Jensen's inequality

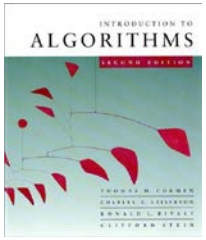
Lemma. Let f be a convex function, and let X be a random variable. Then, $f(E[X]) \le E[f(X)]$.

Proof.

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

$$\le \sum_{k=-\infty}^{\infty} f(k) \cdot \Pr\{X = k\}$$

Convexity lemma (infinite case).



Jensen's inequality

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

Proof.

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

$$\le \sum_{k=-\infty}^{\infty} f(k) \cdot \Pr\{X = k\}$$

$$= E[f(X)]. \quad \square$$

Tricky step, but true—think about it.

Analysis of BST height

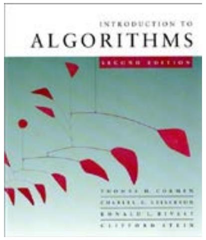
Let X_n be the random variable denoting the height of a randomly built binary search tree on n nodes, and let $Y_n = 2^{X_n}$ be its exponential height.

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

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

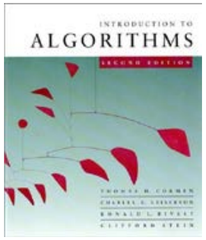
Thus, $\Pr\{Z_{nk} = 1\} = E[Z_{nk}] = 1/n$, and

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

Exponential height recurrence

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

Take expectation of both sides.

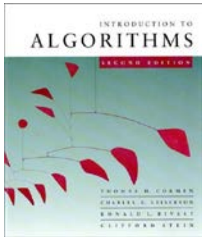


Exponential height recurrence

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

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

Linearity of expectation.



Exponential height recurrence

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

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

Exponential height recurrence

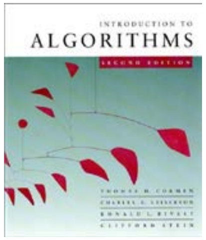
$$E[Y_n] = E\left[\sum_{k=1}^{n} Z_{nk}\left(2 \cdot \max\{Y_{k-1}, Y_{n-k}\}\right)\right]$$

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

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

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

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



Exponential height recurrence

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

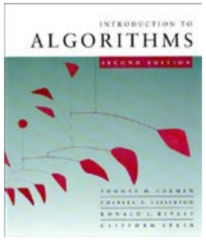
$$= \sum_{k=1}^{n} E[Z_{nk}\left(2 \cdot \max\{Y_{k-1}, Y_{n-k}\}\right)]$$

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

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

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

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.

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

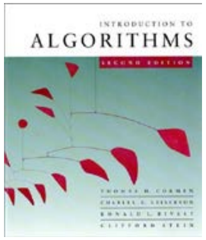
Solving the recurrence

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

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

$$\le \frac{4}{n} \sum_{k=0}^{n-1} ck^3$$

Substitution.



Solving the recurrence

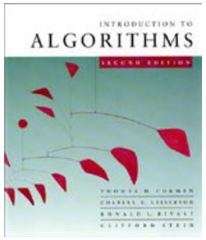
Use substitution to show that $E[Y_n] \le cn^3$ for some positive constant c, which we can pick sufficiently large to handle the initial conditions.

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

$$\le \frac{4}{n}\sum_{k=0}^{n-1} ck^3$$

$$\le \frac{4c}{n}\int_0^n x^3 dx$$

Integral method.



Solving the recurrence

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

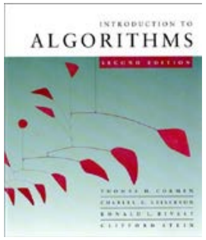
$$E[Y_n] = \frac{4}{n}\sum_{k=0}^{n-1} E[Y_k]$$

$$\le \frac{4}{n}\sum_{k=0}^{n-1} ck^3$$

$$\le \frac{4c}{n}\int_0^n x^3 dx$$

$$= \frac{4c}{n}\left(\frac{n^4}{4}\right)$$

Solve the integral.



Solving the recurrence

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

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

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.

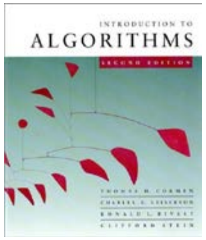
The grand finale

Putting it all together, we have

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

$$= E[Y_n]$$

Definition.

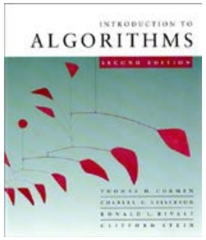


The grand finale

Putting it all together, we have

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

What we just showed.



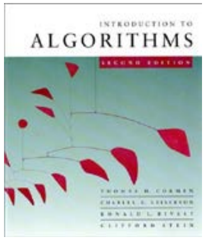
The grand finale

Putting it all together, we have

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

Taking the lg of both sides yields

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



Post mortem

Q. Does the analysis have to be this hard?

Q. Why bother with analyzing exponential height?

Q. Why not just develop the recurrence on

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

directly?

Post mortem (continued)

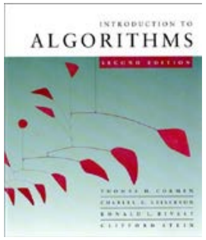
A. The inequality

$$\max\{a, b\} \le a + b \,.$$

provides a poor upper bound, since the RHS approaches the LHS slowly as $|a - b|$ increases. The bound

$$\max\{2^a, 2^b\} \le 2^a + 2^b$$

allows the RHS to approach the LHS far more quickly as $|a - b|$ increases. By using the convexity of $f(x) = 2^x$ via Jensen's inequality, we can manipulate the sum of exponentials, resulting in a tight analysis.



Thought exercises

- See what happens when you try to do the analysis on X_n directly.
- Try to understand better why the proof uses an exponential. Will a quadratic do?
- See if you can find a simpler argument. (This argument is a little simpler than the one in the book—I hope it's correct!)

