Page 1

March 17, 2010
and on Corrections 3/1/2012

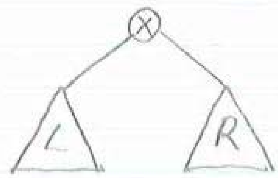
CSC 501/401 Analysis of Algorithms

Lecture Notes

1 Definition of 2-tree.

1. A node with no children is a 2-tree.

2. A tree of the form



where x is a node and L and R are 2-trees, is a 2-tree.

3. Nothing else is a 2-tree.

2 Properties of 2-trees.

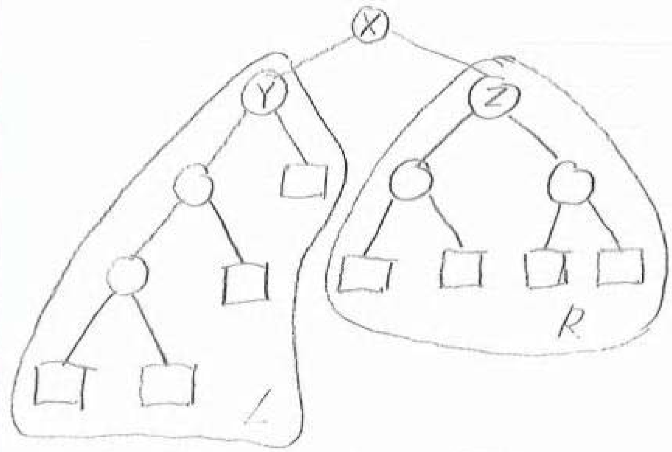
1. Any 2-tree is finite and non-empty.

2. Each node in 2-tree has 0 or 2 children.

3. A 2-tree has n non-leaves and $n+1$ leaves for a total of $2n+1$ nodes, where $n \geq 0$.

Page 2

3 Example of 2-tree T



X is the root of T.
Y and Z are the children of X
Y is the root of L, the left subtree of T
Z is the root of R, the right subtree of T.

external nodes → Squares indicate the leaves of T.
internal nodes → Circles indicate the non-leaves of T.

There are 7 non-leaves in T.
There are 8 leaves in T.
There are 15 nodes in T.

4 Definition of the external path length

The external path length ~~E_T~~ epl(T) in tree T is the sum of lengths of all paths from the root to the leaves of T.

5 Example.

For the tree T visualized in Example 3,

$epl(T)\ \cancel{E_T} = (4+4+3+2)+(3+3+3+3) =$

$= 25$

(Check it!)

For the subtrees L and R of T:

$epl(L)\ \cancel{E_L} = 3+3+2+1 = 9$

$epl(R)\ \cancel{E_R} = 2+2+2+2 = 8$

So, $epl(T) = epl(L) + epl(R) + m$

$\cancel{E_T = E_L + E_R + m}$

where m is the number of leaves in T

(Check it!)

6 Theorem

For any 2-tree T with m leaves,

$epl(T)\ \cancel{E_T} \geq m \lg m$

(In particular, $\cancel{E_T}$ $epl(T)$ $\geq \lceil m \lg m \rceil$.)

Page 4

Proof by induction.

Basis step. T consists of one node x, which is also the only leaf of T.

The only path from the root x of T to its only leaf x has the length 0.
So, $\cancel{E_T}$ $epl(T) = 0$ in this case.

Since the number m of leaves in T is 1,

$$m \lg m = 1 \lg 1 = 1 \cdot 0 = 0.$$

So, $\cancel{E_T}$ $epl(T) \geq m \lg m$ in this case.

This completes the basis step.

Inductive step.

Assume that T has a form indicated in definition 1 item 2, and that the subtrees L and R satisfy the thesis of this theorem, that is,

$$epl(L)\cancel{E_L} \geq m_L \lg m_L \qquad (1)$$

and

$$epl(R)\cancel{E_R} \geq m_R \lg m_R, \qquad (2)$$

where m_L is the number of leaves in L and m_R is the number of leaves in R.

Page 5

Our goal is to prove that

$epl(T)\ \cancel{E_T} \geq m \lg m.$

We have (check it!):

$$m = m_L + m_R \qquad (3)$$

$$epl(T)\ \cancel{E_T} = \overset{epl(L)}{\cancel{E_L}} + \overset{epl(R)}{\cancel{E_R}} + m \geq$$

[by the inductive hypothesis (1) & (2)]

$$\geq m_L \lg m_L + m_R \lg m_R + m \geq$$

[by the convex property of function $f(x) = x \lg x$, which we will prove later, see Lemma 9]

$$\geq 2 \cdot \frac{m_L + m_R}{2} \lg \frac{m_R + m_L}{2} + m =$$

[by (3)]

$$= 2 \frac{m}{2} \lg \frac{m}{2} + m = m(\lg m - \lg 2) + m =$$

$$= m(\lg m - 1) + m = m \lg m - m + m =$$

$$= m \lg m.$$

So, $epl(T)\ \cancel{E_T} \geq m \lg m.$

This completes the inductive step.
This concludes the proof. $\square$

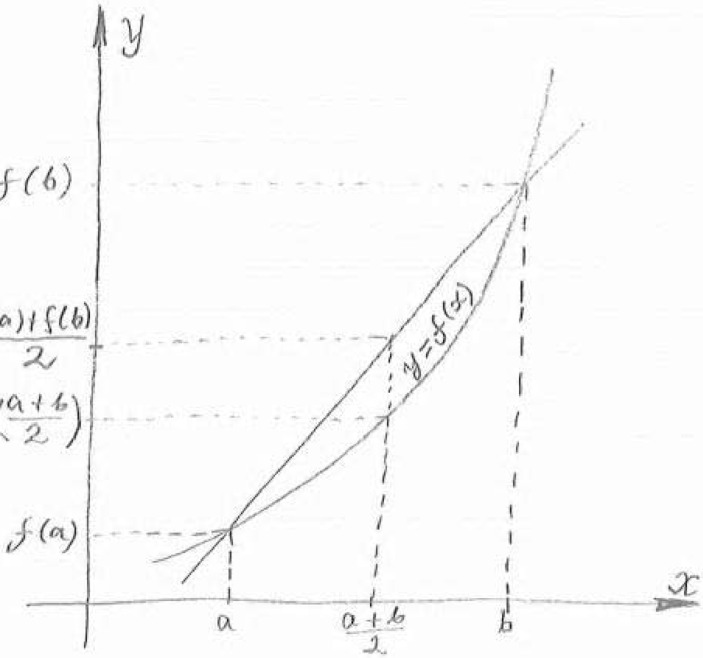
Page 6

7 Lemma (see file ConvexFunctions.pdf)

For any convex function $f(x)$,

$$f(a)+f(b) \geq 2f\left(\frac{a+b}{2}\right). \qquad (4)$$

Proof



We have

$$\frac{f(a)+f(b)}{2} \geq f\left(\frac{a+b}{2}\right).$$

So, multiply both side of the above inequality to get (4)

This concludes the proof. □

Page 7

8 Lemma

$x \lg x$ is a convex function on R^+.

Proof.

Since the domain of $\lg x$ is R^+, $x \lg x$ is a function on R^+.

To prove that $x \lg x$ is convex on R^+, suffices to show that its second derivative is always greater than 0 on R^+.

$$[x \ln x]'' = [[x]' \cdot \ln x + x \cdot [\ln x]']' =$$
$$= [\ln x + x \cdot \tfrac{1}{x}]' = [\ln x + 1]' = \tfrac{1}{x} > 0 \text{ for } x > 0.$$

This completes the proof. □

Do → Exercise Graph completely $x \lg x$.

9 Lemma

For any $a, b \geq 1$

$$a \lg a + b \lg b \geq 2 \frac{a+b}{2} \lg \frac{a+b}{2} \quad (5)$$

Proof. Substitute $x \lg x$ for $f(x)$ in Lemma 7 to conclude (5) from (4).

This completes the proof. □

Page 8

10 Corollary.

For any $a, b \geq 1$,

$$a \lg a + b \lg b \geq (a+b)(\lg(a+b) - 1).$$

Proof

$$2\,\frac{a+b}{2} \lg \frac{a+b}{2} = (a+b)(\lg(a+b) - \lg 2) =$$
$$= (a+b)(\lg(a+b) - 1).$$

Application of (5) completes the proof. □

11 Definition of internal path length

Internal path length ipl(T) $\cancel{I_T}$ in a tree T is the sum of lengths of all paths from the root of T to non-leaves of T.

12 Example

For tree T of example 3,

ipl(T) $\cancel{I_T} = (3+2+1) + (2+1+2) + 0 = 11$

(Check it!)

So, ipl(T) $\cancel{I_T}$ = epl(T) $\cancel{E_T} - 2n$, where n is the number of non-leaves in T. (Check it!)

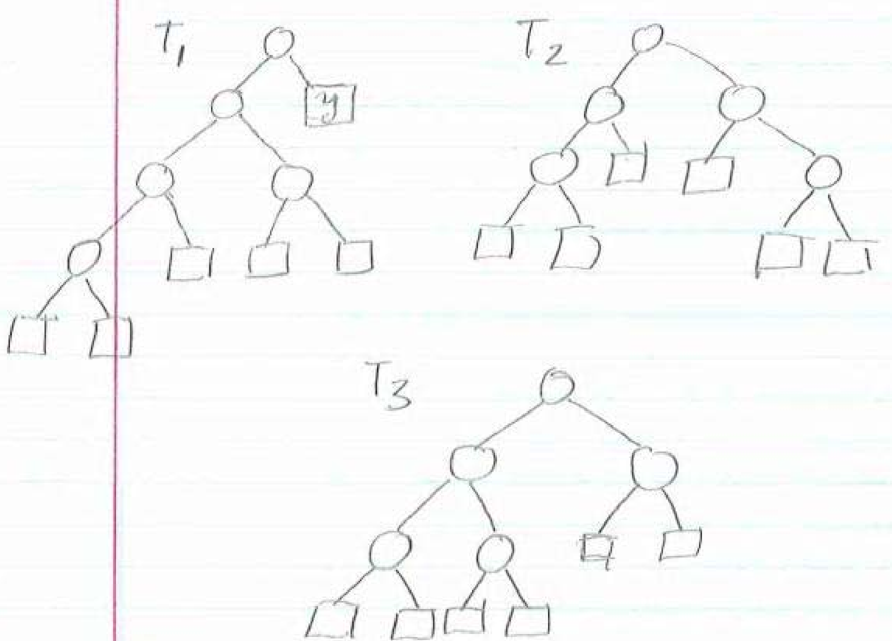
Page 9

13 Theorem. Given the number of nodes n, a 2-tree T that has the smallest external path length ~~E_T~~ epl(T) has leaves on its last level only or on its last two levels only.

Proof in the textbook. and in file 2-trees.PDF. $\square$

14 Example

Consider these 2-trees with 11 nodes.



Their external paths lengths are:

Page 10

$\text{epl}(T_1)\ \cancel{E_{T_1}} = 4+4+3+3+3+1 = 18$

$\text{epl}(T_2)\ \cancel{E_{T_2}} = 3+3+2+2+3+3 = 16$

$\text{epl}(T_3)\ \cancel{E_{T_2}} = 3+3+3+3+2+2 = 16$

Both T_2 and T_3 have leaves only on their last two levels, therefore their extended path lengths are smallest for any 2-tree with 4 nodes.

T_1 has a leaf y on other level than the last two levels, so its external path length is larger.

Also, $\lceil 6 \lg 6 \rceil = \lceil 15.50\ldots \rceil = 16$

so the theorem 6 holds (which should not come as a surprise).

15 Theorem

The shortest external path length in a 2-tree with m leaves is

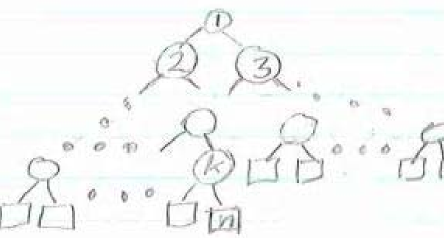
$$E_m^{min} = m \lfloor \lg m \rfloor + 2d \qquad (5)$$

where d is the offset from the largest power of 2 not greater than m. (More precisely, $d = m - 2^n$, where $n = \lfloor \lg m \rfloor$).

Page 11

Proof

We may assume that the 2-tree T with m leaves and smallest external path length has a shape of heap;



Lets enumerate the nodes of T level-by-level, from the left to the right. Node k is the last non-leaf in the sense of this enumeration, and node n is the last leaf.

So, k is the number of non-leaves of T. Therefore, by the property 2 item 3 of 2-trees, $k = m-1$, and $n = 2m-1$.

It follows that the length of path from the root of T to any of its nodes i is equal to $\lfloor \lg i \rfloor$. (Check it!)

So,

$$\mathrm{epl}(T)\ \cancel{E_T} = \sum_{i=k+1}^{n} \lfloor \lg i \rfloor = \sum_{i=m}^{2m-1} \lfloor \lg i \rfloor = \sum_{i=1}^{2m-1} \lfloor \lg i \rfloor - \sum_{i=1}^{m-1} \lfloor \lg i \rfloor =$$

Note: Sipler derivation in file 2-tree. PDF

$$= \left(2m\lfloor \lg 2m\rfloor - 2^{\lfloor \lg 2m\rfloor + 1} + 2\right) -$$
$$\left(m\lfloor \lg m\rfloor - 2^{\lfloor \lg m\rfloor + 1} + 2\right) =$$
$$= 2m\lfloor \lg m + 1\rfloor - 2^{\lfloor \lg m + 1\rfloor + 1} - m\lfloor \lg m\rfloor +$$
$$+ 2^{\lfloor \lg m\rfloor + 1} =$$
$$= 2m(\lfloor \lg m\rfloor + 1) - 2^{\lfloor \lg m\rfloor + 2} - m\lfloor \lg m\rfloor +$$
$$+ 2^{\lfloor \lg m\rfloor + 1} =$$
$$= 2m\lfloor \lg m\rfloor + 2m - 2^2 2^{\lfloor \lg m\rfloor} - m\lfloor \lg m\rfloor +$$
$$+ 2\cdot 2^{\lfloor \lg m\rfloor} =$$
$$= m\lfloor \lg m\rfloor + 2\left(m - 2^{\lfloor \lg m\rfloor}\right) =$$
$$= m\lfloor \lg m\rfloor + 2d.$$

Derived in file 2-tree.PDF

(Here we used the fact:

$$\boxed{\sum_{i=1}^{M} \lfloor \lg i\rfloor = (M+1)\lfloor \lg M\rfloor - 2^{\lfloor \lg M\rfloor + 1} + 2.}$$

Same as: $(M+1)\lfloor \lg(M+1)\rfloor - 2^{\lfloor \lg(M+1)\rfloor + 1} + 2$

proven in lecture notes titled "Balanced tree".)

This concludes the proof. $\square$

Page 13

16 Example

For trees T_2 and T_3 in example 14,

$m = 6$, $\lfloor \lg m \rfloor = 2$, and

$d = m - 2^{\lfloor \lg m \rfloor} = 6 - 2^2 = 6 - 4 = 2$

Also,

$$E_6^{min} = m \lfloor \lg m \rfloor + 2d =$$

$$= 6 \cdot 2 + 2 \cdot 2 = 16.$$

So, $\overset{epl(T_2)}{\cancel{E_{T_2}}} = \overset{epl(T_3)}{\cancel{E_{T_3}}} = E_6^{min}$, and the theorem 15 holds, indeed.

17 Corollary.

For every $m \geq 1$,

$$\underbrace{m \lfloor \lg m \rfloor + 2\left(m - 2^{\lfloor \lg m \rfloor}\right)}_{\text{The tight lower bound}} \geq \overbrace{\lceil m \lg m \rceil}^{\text{Approximation from the textbook}}.$$

Proof by putting together theorems 6 and 15. □

Note. The above is a very close approximation, indeed. Proof would involve Taylor's series.

Page 14

18 Theorem

balanced

1. The average length l_m^{avg} of path from the root to a leaf in a balanced 2-tree T ~~that has a shape of heap (visualised at the beginning of the proof of theorem 15)~~ is

$$l_m^{avg} = \lfloor \lg m \rfloor + \frac{2d}{m} \qquad (6)$$

where m is the number of leaves in tree T and $d \geq 0$ (as before) is the smallest integer that is given by $d = m - 2^n$. (the difference between m and the largest power of 2 not greater than m.

2. If $m = 2^n$ for some n then

$$l_m^{avg} = \lg m = n. \qquad (7)$$

3. For all other 2-trees T with m leaves, the average length l_T of path from the root to a leaf is

$$l_T \geq \lfloor \lg m \rfloor + \frac{2d}{m} \geq \lg m \qquad (8)$$

Proof. To prove part 1 let's note that $l_m^{avg} = \frac{E_m^{min}}{m}$.

Aplication of (5) of theorem 15 yields (6)

Page 15

To prove part 2, let's note that if $m = 2^n$ for some n then $d = 0$ and $\lg m = n = \lfloor n \rfloor = \lfloor \lg m \rfloor$. This yields (7).

Part 3 follows from the fact that heap-shaped 2-trees have smallest external path lengths (theorem 15), and from the lower bound on the external path length in a 2-tree (theorem 6).

These observations complete the proof. □

19 Corollary

The lower bound on the average number of comparisons made by any sorting algorithm that sorts any n-element array by comparisons is

$$LB^{sort}_{avg}(n) = \lfloor \lg n! \rfloor + \frac{2d}{n!} \geq \lg n! \qquad (9)$$

Proof. It follows that $LB^{sort}_{avg}(n)$ is equal to the average length of path from the root to a leaf in a shortest decision tree T for sorting an n-element array by comparisons. Since there are up to $n!$ different arrangements of the array to be sorted, T must have $n!$ nodes.

Plugging-in $m = n!$ into (6) and (8) in theorem 18 yields (9).

This completes the proof. $\square$

It follows that the only cases when $n!$ is a power of 2 is when $n = 1$ or $n = 2$. Therefore, the equality (7) does not hold for any $n > 2$.

So the $\geq$ symbol in (9) may be replaced by $>$ symbol for $n > 2$. Obviously, for $n = 1$ and $n = 2$, the equality holds. This observation allows us to refine collorary 19 to:

$$\text{For } n = 1, 2, \quad LB^{sort}_{avg}(n) = \lg n! \qquad (10)$$

$$\text{For } n > 2, \quad LB^{sort}_{avg}(n) = \lfloor \lg n! \rfloor + \frac{2d}{n!} > \lg n!$$

Approximating $n!$ with $\left(\frac{n}{e}\right)^n \cdot \sqrt{2\pi n}$ (Sterling's formula) yields (check it, using calculator!):

20 Corollary.

For $n > 2$,

$$LB^{sort}_{avg}(n) \approx \lfloor (n + \tfrac{1}{2}) \lg n - 1.45 n + 0.91 \rfloor + \frac{2d}{\sqrt{2\pi n}} \cdot \left(\frac{e}{n}\right)^n$$

$$> (n + \tfrac{1}{2}) \lg n - 1.45 n + 0.91. \qquad \square$$

Page 17

21 Note

The value of $\frac{2d}{n!} \approx \frac{2d}{\sqrt{2\pi n}}\left(\frac{e}{n}\right)^n$ does not converge to 0 as n diverges to ∞, because d is not a constant and it varies within these limits:

$$0 < d < \frac{n!}{2}. \quad \text{(Check it!)}$$

Therefore, $\frac{2d}{n!}$ varies between 0 and 1, which is not surprising if one takes into account that the difference between $\lfloor \lg n! \rfloor$ and $\lg n!$ varies between 0 and 1 as well.

A note regarding my penmanship

Please, keep ~~them~~ in mind that I do not use, intentionally, different "fonts" in my handwriting.

In particular, n, n, n, and n all denote the same symbol. Here is some more of the same:

1 1 1 1 (one)
s S
m, m, m
p P
b b b
etc.