

Lecture 7: "Average root → leaf length in a binary tree"

BOARD NOTES

- We have an arbitrary binary tree T with leaf set $L(T)$.
- We are interested in the average root-leaf length, defined as

$$\text{AvgRL}(T) = \frac{1}{|L(T)|} \sum_{l \in L(T)} d(r, l)$$

where $d(r, l)$ is the number of edges between the root r of T and the node (leaf in fact) l .

- OUR GOAL is to get a LOWER BOUND on $\text{AvgRL}(T)$ in terms of $\lg(|L(T)|)$.

"nice" binary trees, called "near complete", are binary trees where every internal node has two child nodes and every leaf is either at depth h or depth $h-1$ (for h the height of T)

Thm 11 (getting the desired result for "near complete" trees)

If T is "near complete" and has at least 4 leaves, then

$$\text{AvgRL}(T) \geq \frac{\lg(|L(T)|)}{2}$$

proof:

We can partition $L(T)$ into two sets $L_h(T)$ and $L_{h-1}(T)$, where the index (h or $h-1$) refers to the depth of the leaves (note this is using the 'leaf' property of near-complete trees)

Also note (using two-child property for internal nodes) that $2^{h-1} < |L(T)| \leq 2^h$. using this in eq.

Then,

$$\begin{aligned} \text{AvgRL}(T) &= \frac{(h|L_h(T)| + (h-1)|L_{h-1}(T)|)}{|L(T)|} \\ &\geq (h-1) \frac{|L(T)|}{|L(T)|} = (h-1) \geq \lg(|L(T)|) - 1 \\ &\geq \lg(|L(T)|) / 2 \quad \text{if } |L(T)| \geq 4 \end{aligned}$$

Now have the desired lower bound on $AvgRL(T)$ for "near complete" trees - want a similar lower bound for all binary trees.

Thm 12

For any binary tree T , we can make a "near complete" tree T' with the same number of leaves such that

$$AvgRL(T) \geq AvgRL(T')$$

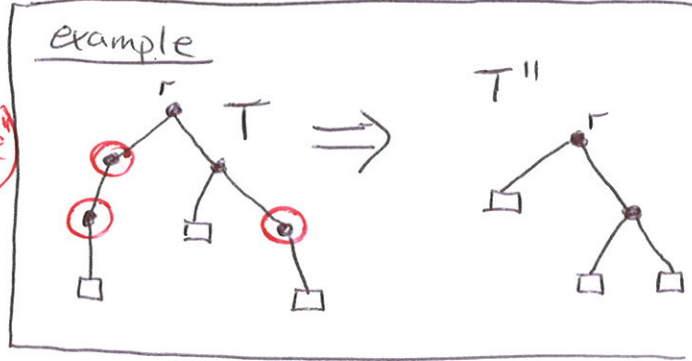
and we already know $AvgRL(T') \geq \frac{\lg(|L(T')|)}{2}$ by Thm 11

proof:

STEP 1 Take T and CONTRACT any internal nodes with only one child, to get an intermediate tree T'' where every internal node in T'' has two child nodes

(this is one of the "near-complete" properties)

$$L(T'') = L(T)$$



Note $d_{T''}(r, l) \leq d_T(r, l)$

for any leaf l , because contractions only decrease distance. So $AvgRL(T'') \leq AvgRL(T)$.

STEP 2

Take T'' and perform a series of pruning-and-reattaching moves on "low-lying" leaf siblings, until all leaves are at depth h or $(h-1)$ (for the appropriate height h)

one move:

Suppose x is a "lowest leaf" and z is a "highest leaf".

If $d(r, x) - d(r, z) \leq 1$, we have a "near complete" tree.

We take the current tree to be T' .

If $d(r, x) - d(r, z) \geq 2$, not a "near complete" tree.

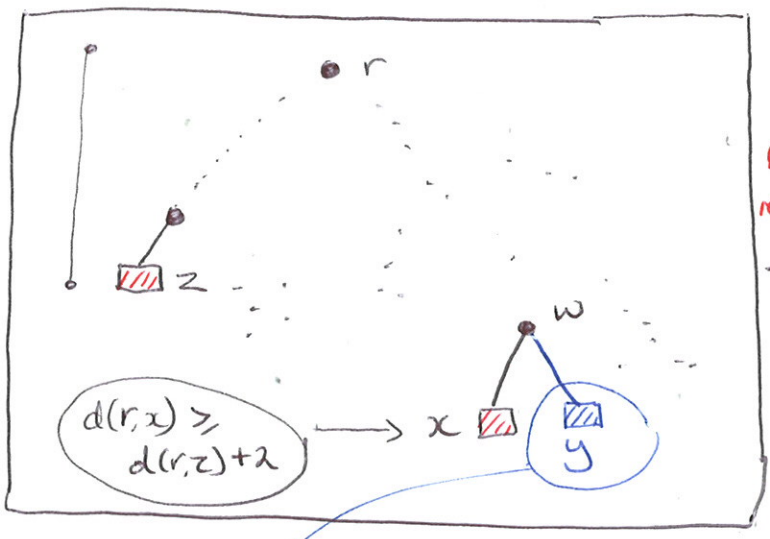
See details on next page of how we will "prune and re-attach" to rectify the height imbalance.

continuing proof of Thm 12
(working on STEP 2)

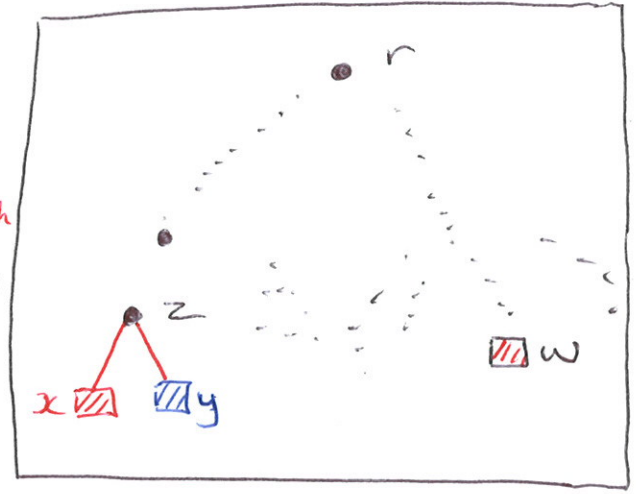
x is a "lowest leaf", depth $d(r,x)$
 z is a "highest leaf", depth $d(r,z)$.

Difference at least 2. (so NOT near complete)

current tree



prune and re-attach



x must have a sibling which is a leaf because x 's parent w must have 2 child nodes (STEP 1) and x is "lowest", so y cannot be an internal node.

our "move"

"chop off" the low sibling pair (x) and (y) (making parent (w) a LEAF) and reattach under the high leaf (z) (which stops being a LEAF)

Result of "move"

(I) x, y still leaves; z is NO LONGER A LEAF \Rightarrow SAME NUMBER of leaves
 w has BECOME a leaf

(II) change in "depths":

For x $d(r,z) + 1 - d(r,x)$ For y $d(r,z) + 1 - d(r,y)$

For z (no longer leaf) $-d(r,z)$ For w (now a leaf) $d(r,w) - 1$

Total change $2d(r,z) + 2 - 2d(r,x) - d(r,z) + d(r,w) - 1$
 $= d(r,z) + 1 - d(r,x) < 0$ (by assumption $d(r,x) \geq d(r,z) + 2$)

So AvgRL has decreased ↓.

(this "move" only decreases AvgRL)

III

Also, after performing the move, we have decreased the overall height imbalance.

It is true that there could be other z, x examples still in the tree (ie $d(r, z)$ and $d(r, x)$ at least 2 away)

However, we have "made progress" in the following way.

Let h_{min} be the value of $d(r, z)$ BEFORE (for z a "highest" leaf) the move.

Consider the measure

$$Imb(\hat{T}) = \sum_{\substack{l \in L(\hat{T}) \\ d(r, l) \geq h_{min}(\hat{T}) + 2}} (d(r, l) - h_{min}(\hat{T}))$$

NOTICE that $Imb(\hat{T})$ would be 0 for a "near complete" tree

Then after the "move" is performed \hat{T} (say tree is \hat{T})

we can delete $2(d(r, x) - h_{min}(\hat{T})) = 2(d(r, x) - d(r, z))$ FROM Imb

(because now x and y are either at $h_{min} + 1$ or maybe even the new $h_{min}(\hat{T})$)

we might have to add (for new leaf w)

$$(d(r, x) - 1 - h_{min}(\hat{T})) \leq (d(r, x) - 1 - h_{min}(\hat{T}))$$

(only will have to add this if x was further than 2 down in the first place)

Overall the change is AT most

$$\begin{aligned} Imb(\hat{T}) &\leq Imb(\hat{T}) - 2(d(r, x) - h_{min}(\hat{T})) \\ &\quad + (d(r, x) - 1 - h_{min}(\hat{T})) \\ &= Imb(\hat{T}) - (d(r, x) - h_{min}(\hat{T})) - 1 \\ &< Imb(\hat{T}). \end{aligned}$$

so we strictly DECREASE Imb value.

So After enough "moves" we will get "Imb" to 0 (satisfying rules of a "near complete" tree) already known by step 1

T' on final line is the final tree after all "moves" done, all leaves are at depth h_{min} or $h_{min} + 1$.
by inductively applying II over many moves.

Also, by II, $AvgRL(T') \leq AvgRL(T'') \leq AvgRL(T)$
So the lower bound of Thm II for "near complete" holds for all T .