Optimal Binary Search Trees
Sections 12, 15.5, 16.2

Searching under the comparison model

Binary search: lg n upper and lower bounds also in "expected case" (probability of search same for each element)

With some balanced binary scheme, updates also in O(lg n)

But what if some elements are requested more than others?

Start with stochastic model: Given a set of n keys, K =<k_1.. k_n> with independent probabilities of access, p_i.

How can we organize a search structure to minimize the expected number of comparisons for a search?

Clearly this is a binary search tree, though in general far from balanced.

How do we find the optimal tree?

Optimal Binary Search Trees

Try a few top down heuristics, and we easily get non-optimal trees.

We need some definitions:

k_i : i^{th} largest of key value i=1..n

d_i : dummy leaf after k_i, before k_{i+1} i=0..n

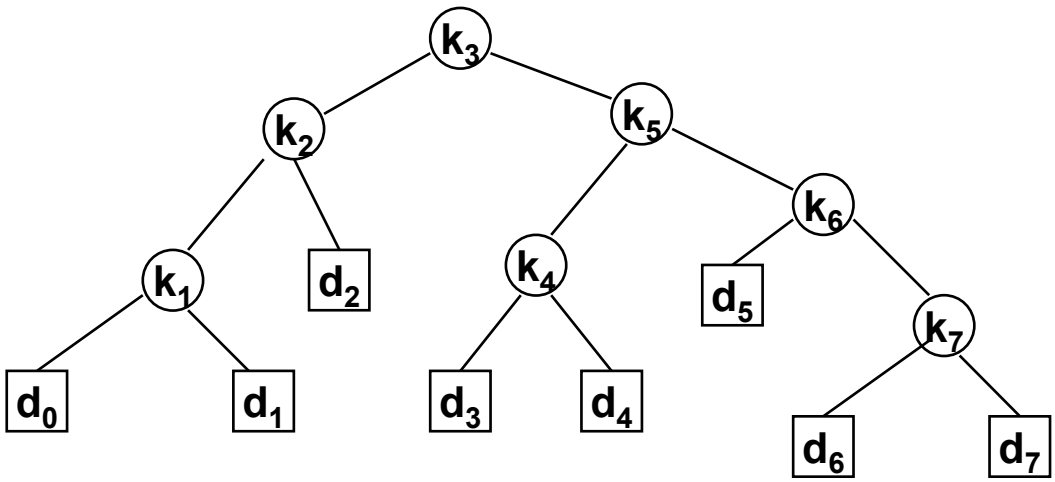
p_i : probability of a request for k_i

q_i: probability of a request for an element after p_i and before p_{i+1}.

w(i,j): probability of element in OPEN INTERVAL (k_{i-1},k_{j+1}), so

$w(i,j)=\sum_{k=i,.j} p_k + \sum_{k=i-1..j} q_k$; $w(1,n)=1$

{find all w(i,j) in $O(n^2)$ time



The Dynamic Program

Expected cost of a search:

E[search in T] =

$\sum(\text{depth}_T(k_i) + 1)p_i + \sum(\text{depth}_T(d_i) + 1)q_i$

So ..

Compute

e[i,j] = cost of optimal i,j tree

$= q_{i-1}$ **if j=i-1 and**

$= \min_{i \le r \le j}\{e[i,r-1]+e[r+1,j]+w(i,j)\}$ **if** $i \le j$

Also keep track of the root as r[i,j]

$r[i+1,i] = d_i$; $r[i,i] = k_i$;

Otherwise r[i,j] determined by e[i,j] calculation

This gives a straightforward dynamic program, which we can do with a loop r=i,..j and recursive calls, of three loops for $\Theta(n^3)$ time, and $\Theta(n^2)$ space.

Improvement

Note: the loop to compute {r,e}[i,j] goes all the way from i to j.

Is all this necessary?

Lemma: r[i,j] cannot precede r[i,j-1] or follow r[i+1,j]. {Omit proof}

So modify the inner loop

Change "r = i..j" to "r = r[i,j-1]..r[i+1,j]"

Look at runtime; series telescopes

Theorem: The optimal binary search tree can be determined in $\Theta(n^2)$ time and $\Theta(n^2)$ space.

This is the best known algorithm, indeed there is no known polynomial time, $o(n^2)$ space method.

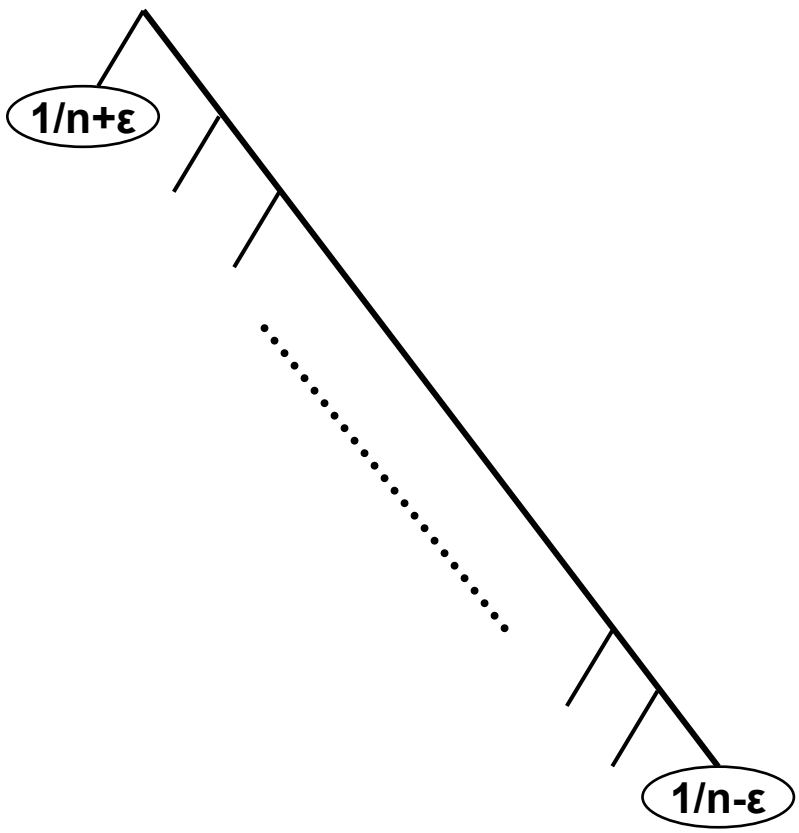
Good Trees in Less Space

Suppose we don't have $\Theta(n^2)$ space.

How can we get a good tree?

Greed !!!

First Attempt: Choose root as key with greatest probability.

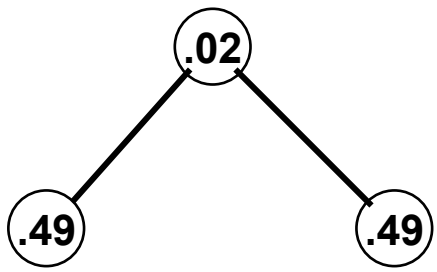


Better Greed

Choose root to balance the weights on either side as well as possible. There are a few (picky) options:

MinMax: Minimizes the weight of largest subtree

"Balance": Choose root to make subtree wts as close as possible



Not optimal; but perhaps, not bad.

Naïve algorithm is $\Theta(n^2)$ (worst case), but O(n) space.

How can we make it faster?

A Faster Greed Algorithm

Given keys in order, and probs: p_i, d_i

Let L_i = probability of being left of k_i

For root of tree in range [st,fin], find, by binary search, key with L_i below and L_{i+1} above $(L_{st}+L_{fin+1})/2$

Hence an O(n lg n) method.

Can we improve this?

Yes … note the method is linear if you always "get lucky" with "split" near the middle.

So … Start at with one comparison in the middle. Then move to the "side still in" and double your way toward the middle, till desired element "bracketed", finish with binary search.

Runtime

Cost of discovering split point:

O(lg v), v is #keys from near end.

Thm: The algorithm runs in O(n) time. i.e. the splits have an amortized cost of O(1).

Proof sketch:

Try induction: $T(n) \geq \alpha n - \beta \lg n$

Tune constants for the base cases

Basic recurrence:

$T(n) \geq T(a) + T(n-a-1) + 2 \lg a \quad \{1<a\leq n/2\}$

Substitute: $T(n) =$

$\alpha a - \beta \lg(a) + \alpha n - \alpha a - \alpha - \beta \lg(n-a-1) + 2 \lg a$

We require

$\beta \lg(a) + \beta \lg(n-a-1) + \alpha > \lg n \quad \{\text{when } a<n/2\}$

This is fine when a is large, α and β have to be tuned to handle the small values of a.

Quality of Solutions

The approximation method is rather good. Define:

- $P = \sum p_i$
- H, the entropy of a distribution, as $H(p_1..p_n,q_0..q_n) = -\sum p_i \lg p_i + \sum q_i \lg q_i$ {Note H is maximized when all probabilities are the same}
- C_{opt} and C_{approx} as tree costs

Then

Thm: $H - P \lg(eH/2P) \leq C_{opt} \leq C_{approx} \leq H+2 - P$

i.e. optimal and approximate tree have costs with lg H of optimal.

Proof: Omitted

But what it probabilities change… or we don't know them?

Could count accesses and update optimal tree based on changing probabilities.

{this has been done for Huffman codes}

Or

Recall linear search and the "move to front" heuristic. Assume list starts empty and element put at the end the first time it is requested

Thm (from CS 240): The cost of a sequence of searches under the move to front heuristic is within a factor of 2 of that of the optimal (static) order.

Indeed

Thm: The amortized cost of a search under move to front is at most twice the optimal we could get if we knew the sequence and updated the list. {Off line updates work by swapping adjacent items}

Splay Trees

See Goodrich and Tamassia Section 3.4

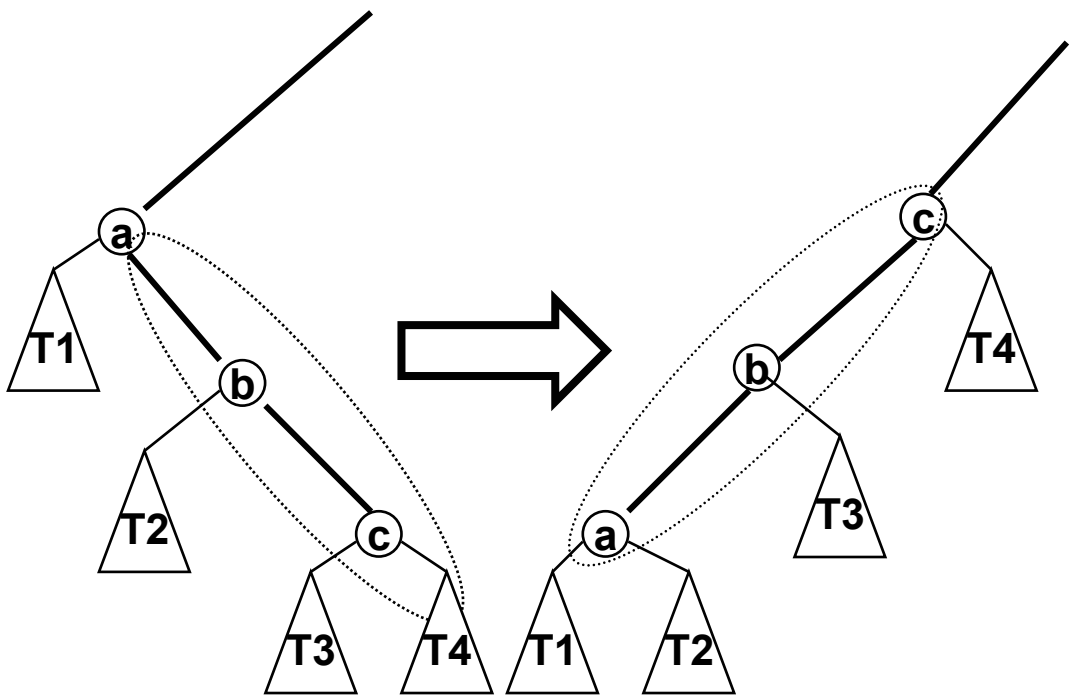
Analogy for search trees:

… when access is made perform rotations to move requested element to root.

Must be careful, a simple rotation does not give the ideal result.

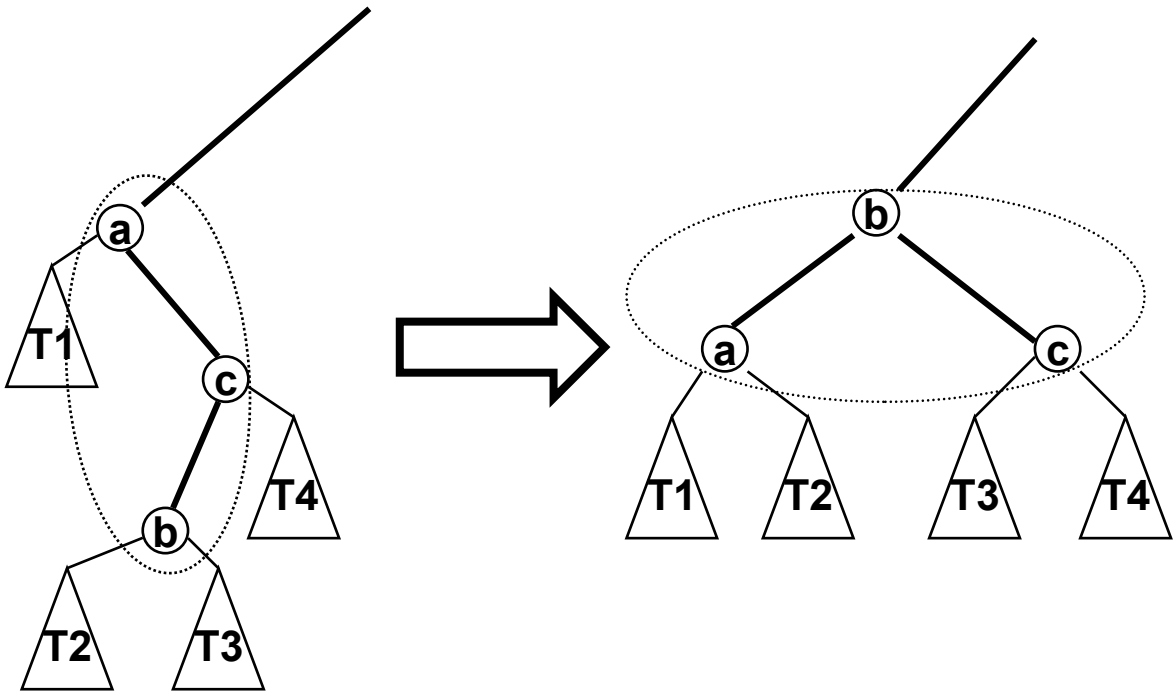
3 moves:

Zig-zig: abc (similarly cba)

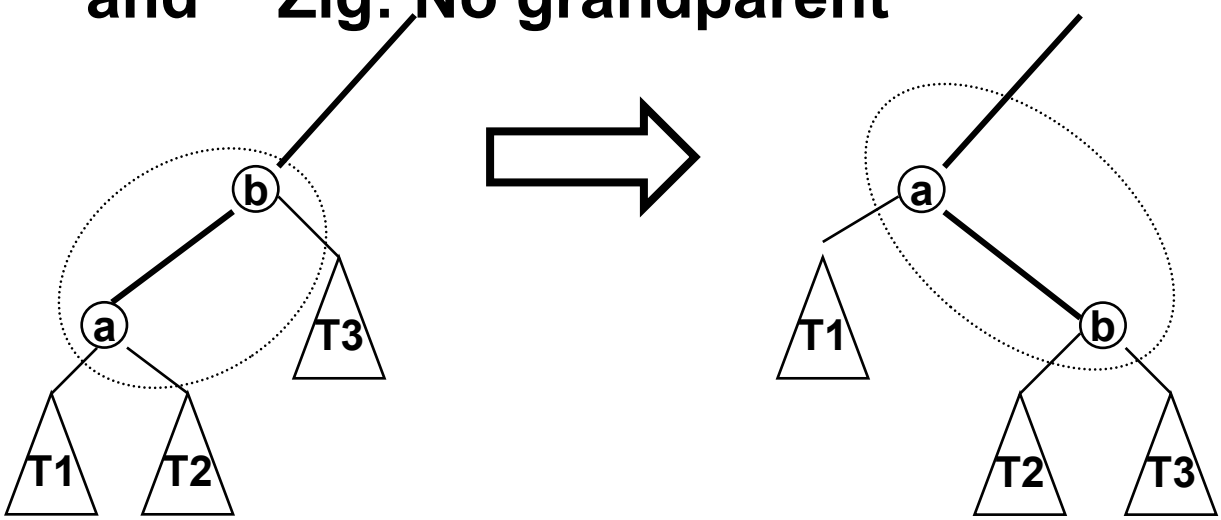


And the other two forms

Zig zag: bca (or bac)



and Zig: No grandparent



Splay trees: The results

Thm: On any long enough sequence of searches, the length of the search path on a spay tree is at most twice that of the optimal tree for that sequence.

Hence:

Thm: The amortized cost of searching in a splay tree is at most 2H + O(1).

Proofs: Omitted (see G&T)