

Topic Notes: Priority Queues and Heaps

Heaps and Priority Queues

From here, we will look at some ways that trees are used in other structures.

First, we'll consider a structure that seems somewhat like the ordered structures and somewhat like a queue:

A *priority queue* is a structure where the contents are comparable elements and the elements with "small" (or "large") values are removed before elements with "larger" (or "smaller") values.

Waiting at a restaurant can be a priority queue – you get in line but people who are regular customers or who give a tip to the host or hostess may move ahead of you in line.

This could happen at a hospital emergency room. The order in which patients are seen is not strictly based on order of arrival but on the severity of their condition.

Same idea when airports are backed up. Planes will get in line for their turn on the runway, but scheduling concerns or fuel issues or whatever else may cause ground control to give a plane which "got in line" later a higher priority and move them up.

An operating system may be scheduling processes that are competing to use the CPU in a multi-programmed system. The next process may be selected based on the highest priority process that is seeking a turn on the CPU.

Much like stacks and traditional queues, we only need to define a single `add` and `remove` method. We can add any element at any time, but we can only ever examine or remove the highest priority item. What is meant by "highest priority"? It could mean the smallest number, largest number, first string alphabetically, or anything else. In the general case, we'll either need to restrict to `Comparable` elements or require a `Comparator` (we might talk about these, but you'll definitely see them in CSIS 225 if you continue on) that determines the priorities of the items.

One can implement a priority queue sort of like a regular queue, built upon an array, `ArrayList`, but the operations would require more work either to insert or to remove an element.

Option 1: Maintain the values sorted by priority order so that the value to be removed is always at the end. This means insert is $O(n)$, but remove will be $O(1)$.

Option 2: Do not maintain any order among the values, but search for the highest priority on removal. This means insert is $O(1)$, but we'll need an $O(n)$ linear search to find the item to remove.

Similarly, we could use a linked list structure, but these suffer from the same problem: we're either spending $O(n)$ time on insert maintaining the list in priority order to permit an $O(1)$ remove, or

spending $O(n)$ time on remove from an unsorted list.

But... we can do better. Using the observation that we don't need to keep the entire structure ordered – at any time we only need quick access to the smallest element – we can provide a more efficient implementation using a structure called a *heap*.

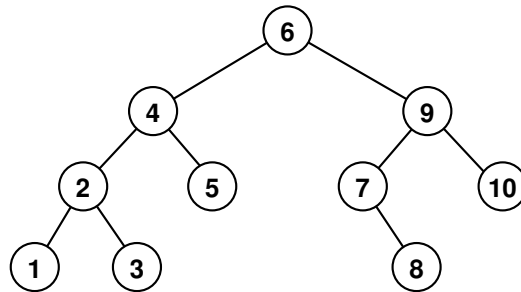
Some Binary Tree Properties

Before we discuss the idea of a heap and how it allows us to build priority queues, recall the following tree terminology and properties:

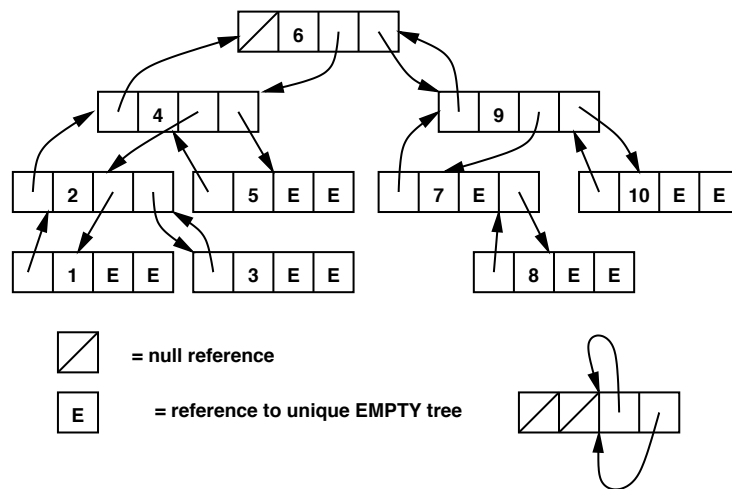
- A *complete* binary tree of height h is obtained from a full binary tree of height h with 0 or more (but not all) of the rightmost leaves at level h removed.
 - We say T is *balanced* if it has the minimum possible height for its number of nodes.
 - **Lemma:** If T is a binary tree, then at level k , T has $\leq 2^k$ nodes.
 - **Theorem:** If T has height h then $n = \text{num nodes in } T \leq 2^{h+1} - 1$. Equivalently, if T has n nodes, then $n - 1 \geq h \geq \log(n + 1) - 1$.
-

Array representations of trees

Our representation of the tree:



actually looks like this:



That’s a lot of extra references to parents and children, and to empty nodes. So to store 10 actual data values, we need space for 40 references plus the 4 that make up the empty tree instances.

The following array contains exactly the same information:

v

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	6	4	9	2	5	7	10	1	3	/	/	/	8	/	/

The array, `data[1..n]`, holds the values to be stored in the tree. It does not contain explicit references to the left or right subtrees or to parents. If we leave slot 0 empty, we can simplify the arithmetic and code for finding children and parents.

Instead the children of node i are stored in positions $2 * i$ and $2 * i + 1$, and therefore the parent of a node j , may be found at $j/2$

This lets us save space for links, but it is possible that there is a significant waste of storage:

Storing a tree of height n requires an array of length $2^{(n+1)} - 1$ (!), even if the tree only has $O(n)$ elements. This makes this representation very expensive if you have a long, skinny tree. However, it is very efficient for holding full or complete trees. For our example, we need 15 references to hold a tree of 10 items, compared to 40 for the fully constructed binary tree.

Heaps

Recall that a complete binary tree is one in which every level is full except possibly the bottom level and that level has all leaves in the leftmost positions. (Note that this is more restrictive than a balanced tree.)

Definition: A *Min-Heap* H is a complete binary tree such that

1. H is empty, or

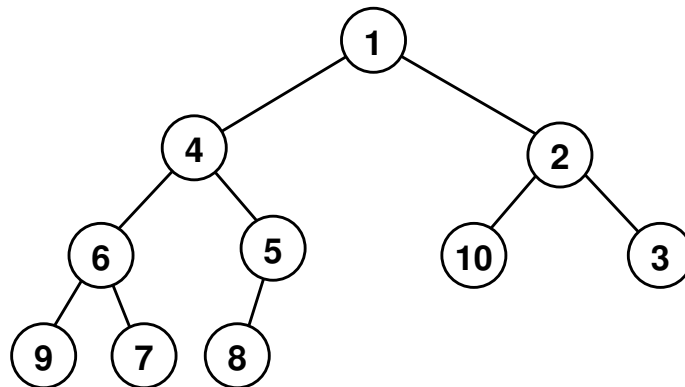
2. (a) The root value is the smallest value in H and
- (b) The left and right subtrees of H are also heaps

This is equivalent to saying that $H[i] \leq H[2*i+1]$, and $H[i] \leq H[2*i+2]$ for all appropriate values of i in the array representation of trees.

We could just as well implement a *Max-Heap* – just reverse all of our comparisons. The text in fact does talk about max heaps.

In either case, it often makes sense to use the array/ArrayList representation of binary trees here, since the tree is also guaranteed to be complete, so there will never be empty slots wasting space.

Another way of looking at Min-Heap is that any path from a leaf to the root is in non-ascending order.



Note that there are lots of possible Min-Heaps that would contain the same set of values. At each node, the subtrees are interchangeable (other than those which have different heights, strictly speaking).

In a Min-heap, we know that the smallest value is at the root, the second smallest is a child of the root, the third smallest is a child of the first or second smallest, and so on.

This turns out to be exactly what is needed to implement a priority queue.

We need to be able to maintain a Min-Heap, and when elements are added or the smallest value is removed, we want to “re-heapify” the structure as efficiently as possible.

Inserting into a Heap

1. Place number to be inserted at the next free position.
2. “Percolate” it up to correct position.

Deleting the Root from a Heap

1. Save value in root element for return (it’s the smallest one).

2. Move last element to root
3. Push down (or “sift”) the element now in the root position (it was formerly the last element) to its correct position by repeatedly swapping it with the smaller of its two children.

Notice how these heap operations implement a priority queue.

When you add a new element in a priority queue, copy it into the next free position of the heap and “percolate” it up into its proper position.

```
insert (value){
    if the array is full {
        //size of heap = array.length - 1
        double the size of the array
    }

    increase the size of the heap by 1
    pos = size

    //Percolate up
    while (pos > 1 and value < array[pos / 2]) {
        array[pos] = array[pos/2]
        pos = pos / 2
    }

    array[pos] = x;
```

When you remove the next element from the priority queue, remove the element from the root of heap (first element, since it has lowest number for priority), move the last element up to the first slot, and then sift it down.

How expensive are these insert and remove operations?

Each is $O(\log n)$, as we can, at worst, traverse the height of the tree. Since the tree is always complete, we know its height is always at most $\log n$. This is much better than storing the priority queue as regular queue and inserting new elements into the right position in the queue and removing them from the front.

Sorting with a Heap (*HeapSort*)

The priority queue suggests an approach to sorting data. If we have n values to sort, we can add them all to the priority queue, then remove them all, and they come out in order. We’re done.

What is the cost of this? If we use the naive priority queue implementations (completely sorted or completely unsorted internal data, making either add or remove $O(n)$ and the other $O(1)$), we

need, at some point, to do a $O(n)$ operation for each of n elements, making an overall sorting procedure of $O(n^2)$. That's not very exciting – we had that with our naive sorting algorithms.

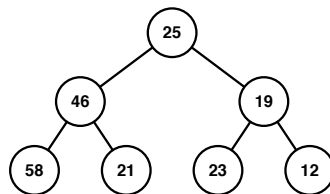
But what about using the heap-based priority queues?

We can build a heap from a collection of objects by adding them one after the other. Each takes at most $O(\log n)$ to insert and “percolate up” for a total time of $O(n \log n)$ to “heapify” a collection of numbers. That's actually the cost of the entire sorting procedure for merge sort and quicksort, and here, we've only achieved a heap, not a sorted structure. But we continue..

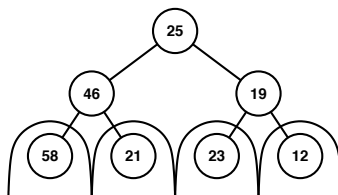
Once the heap is established, we remove elements one at a time, putting smallest at end, second smallest next to end, etc. This is again n steps, each of which is at most an $O(\log n)$ operation.

So we have an overall cost of $O(n \log n)$, just like our other good sorting procedures.

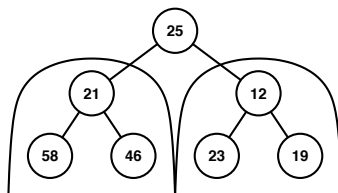
We can actually do a little better on the “heapify” part. Consider this example, which I will draw as a tree, but we should remember that it will really just all be in an array or an `ArrayList`.



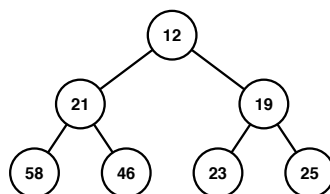
We want to heapify. Note that we already have 4 heaps – the leaves:



We can make this 2 heaps by “sifting down” the two level 1 nodes:



Then finally, we sift down the root to get a single heap:



How expensive was this operation, really?

We only needed to do the “push down root” operation on about half of the elements. But that’s still $O(n)$ operations, each costing $O(\log n)$.

The key is that we only perform “push down root” on the first half of the elements of the array.

That is, no “push down root” operations are needed corresponding to leaves of the tree (corresponding to $\frac{n}{2}$ of the elements).

For those elements sitting just above the leaves ($\frac{n}{4}$ of the elements), we only go through the loop once (and thus we make only two comparisons of priorities).

For those in the next layer ($\frac{n}{8}$ of the elements) we only go through the loop twice (4 comparisons), and so on.

Thus we make

$$2 * \left(\frac{n}{4}\right) + 4 * \left(\frac{n}{8}\right) + 6 * \left(\frac{n}{16}\right) + \dots + 2 * (\log n) * (1)$$

total comparisons.

Since $2^{\log n} = n$, we can rewrite the last term to fit in nicely:

$$2 * \left(\frac{n}{4}\right) + 4 * \left(\frac{n}{8}\right) + 6 * \left(\frac{n}{16}\right) + \dots + 2 * (\log n) * \left(\frac{n}{2^{\log n}}\right)$$

We can factor out the n , multiply in the 2 (to reduce each denominator by 2) and put things into a more suggestive format:

$$n * \left(\frac{1}{2^1} + \frac{2}{2^2} + \frac{3}{2^3} + \dots + \frac{\log n}{2^{\log n}} \right)$$

The sum inside the parentheses can be rewritten as

$$\sum_{i=1}^{\log n} \frac{i}{2^i}$$

This is clearly bounded above by the infinite sum,

$$\sum_{i=1}^{\infty} \frac{i}{2^i}$$

Let’s see if we can evaluate this infinite sum:

$$\frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \dots$$

We can rewrite this in a triangular form to be able to use some tricks to get the sum:

$$\begin{array}{rcl}
 \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots & = & 1 \\
 \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \cdots & = & \frac{1}{2} \\
 \frac{1}{8} + \frac{1}{16} + \cdots & = & \frac{1}{4} \\
 \frac{1}{16} + \cdots & = & \frac{1}{8} \\
 \cdots & = & \frac{1}{16} \\
 & & \ddots \\
 & & 2
 \end{array}$$

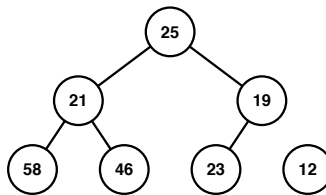
Thus

$$n * \left(\frac{1}{2^1} + \frac{2}{2^2} + \frac{3}{2^3} + \cdots + \frac{\log n}{2^{\log n}} \right) \leq 2n$$

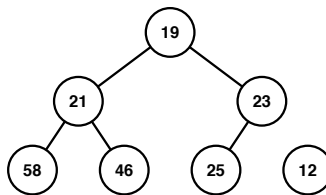
and hence the time to heapify an array or `ArrayList`, in place, is $O(n)$.

The second phase, removing each successive element, still requires n removes, each of which will involve a $O(\log n)$ heapify.

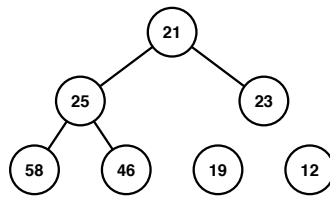
We can, however, do this in place in our array or `ArrayList`, by swapping each element removed from the heap into the last position in the heap and calling the heap one item smaller at each step.



And then we do a sift down of the root (which just came up from the last position):



Then the next item (19) comes out of the heap by swapping into the last position, and we sift down the 25:



And the process continues until there is no more heap.

The entire process of extracting elements in sorted order is $O(n \log n)$.

Therefore, the total time is $O(n \log n)$ in the worst case.

Plus, there's no extra space needed! It's an in-place sort, which is a benefit over our other $O(n \log n)$ sorting algorithm: merge sort.

Note that using a Min-Heap, we end up with the array sorted in descending order. If we want to sort in ascending order, we will need a Max-Heap.

So we have an efficient sorting procedure to add to our arsenal.