

# Online Median Finding

Larry Denenberg

(draft)

## Abstract

The **Online Median** problem requires us to add elements to a set and at any time to find the median of the elements added so far. A straightforward solution using two heaps permits adding elements in  $O(\log N)$  comparisons while keeping the current median always available. We show how to reduce the expected number of comparisons for adding an element to  $2 + o(1)$ .

Who the hell cares how much time it takes?  
If it's too slow for you, buy a faster machine.  
—Don Knuth, *Full Retraction*, §3.16

## 1 Introduction

The median of a set of totally-ordered elements can be found in linear time; the best algorithm to date requires  $2.942N + o(N)$  element comparisons to find the median of  $N$  elements [BF, DZ]. We are interested in the online problem, where elements arrive one by one and we might be asked at any time for the median of the elements added so far. We require two operations:

- **AddElement**( $x$ ), add a new element  $x$  to the current set, and
- **CurrentMedian**(), return the median of the current set of elements.

A straightforward solution uses two priority queues, one with operations **Insert** and **DeleteMin**, and one with “inverted” operations, i.e., with **DeleteMax** in place of **DeleteMin**. The first queue stores elements greater than the current median; the second stores elements smaller than the current median. We keep the queues balanced: At all times, either they have the same number of elements, or one has one more element than the other. A new element is placed in the appropriate queue, and if that queue is then too large we *rebalance* by transferring its extreme element to the other queue.

When the queues have different sizes, the current median is the extreme element of the larger. Otherwise, we take the extreme element of either queue as the median. Call this algorithm the *basic method* (Figure 1).

With priority queues implemented by heaps as in [LD], **AddElement** requires time  $O(\log N)$  in both worst and expected cases. In this note we employ several coding tricks to obtain surprising improvements in the expected time for **AddElement** under the assumption of random inputs.

As usual, we use number of element comparisons as the measure of time, but we eschew techniques where counting comparisons does not reflect true clock time. For example, heap **Insert** can be performed in  $O(\log \log n)$  comparisons by using binary search to find the correct location of the new element on the path from the new leaf to the root, then swapping the new element up to that location without element comparisons. But the clock time is not  $O(\log \log n)$  so this trick won't be used.

We ignore space issues and assume that all heaps grow magically as required. A different approach to online median finding, where very little

```

class OnlineMedianFinder
{
    public int CurrentMedian; // the current median, always available

    MinHeap BigElements = new MinHeap(); // elements > CurrentMedian
    MaxHeap SmallElements = new MaxHeap(); // elements < CurrentMedian
    int balance = 0; // tells which heap (if any) has an extra element

    public void AddElement(int newval) {

        switch (balance) {

        case 0: // the two heaps have the same number of elements
            if (newval <= CurrentMedian) { // "<=" vs "<" is irrelevant
                CurrentMedian = SmallElements.Insert(newval); // new max
                balance = -1;
            } else {
                CurrentMedian = BigElements.Insert(newval); // new min
                balance = +1;
            }
            return;

        case +1: // BigElements has an extra element
            if (newval <= CurrentMedian) { // "<=" vs "<" is significant
                // this Insert brings us back into balance
                SmallElements.Insert(newval);
            } else {
                BigElements.Insert(newval);
                // BigElements now has two extra elements; must rebalance
                SmallElements.Insert(BigElements.DeleteMin());
            }
            balance = 0; // we're always balanced at this point
            return; // note that CurrentMedian doesn't change

        case -1: // SmallElements has an extra element
            // [the code is parallel to the +1 case]
        }
    }
}

```

Figure 1: The basic method, implemented in Java. For convenience, heap Insert is assumed to return the new extreme element.

space is used but which produces only an approximation to the median, has been described by several researchers; see [CH] for more information.

For simplicity, we generally use terminology appropriate for a `MinHeap`, the heap containing larger elements and implementing `DeleteMin`. For example, we say that “rebalance requires a single `DeleteMin`” even though some rebalances use `DeleteMax` instead.

## 2 Analysis of the Basic Method

At each call on `AddElement`, either the two heaps have equal size or one of them has one more element than the other; these two states alternate. So half the time the sizes are equal and we perform a single `Insert`.

When the heaps are unbalanced, the newly-arrived element either belongs in the heap with fewer elements and we regain balance with a single `Insert`, or it belongs in the heap with more elements and we rebalance with two `Inserts` plus a `DeleteMin`. These two cases have equal probability when inputs are random. In summary, with probability 1/4 we perform the three-operation rebalance, and with probability 3/4 we do a single `Insert`.

To `Insert` a random element into a heap is usually cheap, since the new element rarely belongs near the top. In fact, the expected number of comparisons for `Insert` is constant, independent of heap size. We can see this with a little handwaving: Roughly half the heap’s elements, the larger ones, are in the bottom layer, so in half of all `Inserts` the new element comes to rest after a single comparison. Similarly, two comparisons are required about 1/4 of the time, and so forth. Summing, we find that the expected number of comparisons for `Insert` is about 2.

There are several more precise estimates of the expected number of comparisons for `Insert`, depending on specific randomness assumptions[BS, PS]. Our problem seems different, no doubt because of the mixture of `Inserts` and `DeleteMins` and possibly because of the truncated distribution: A new element smaller than the heap’s current minimum will almost certainly go into the other heap! We sidestep this question by simply defining  $\gamma$  as the expected number of comparisons for heap `Insert` of a random element during basic online median finding, recognizing that this value has not been proven to exist. Experimentation suggests that  $\gamma$  is no more than 2.06.

`DeleteMin`, on the other hand, is expensive. As we swap an element downwards we need two comparisons at each level to find the smaller child of the current node. About half the heap is at the bottom, so a random element travels nearly all the way down. Further, the traveling element

is *not* random; it was taken from the bottom level and is therefore even more likely (though not certain) to return all the way down. The expected number of comparisons is thus very close to the worst-case value,  $2 \log_2 N$ , where  $N$  is the number of elements in the heap. (We use this definition of  $N$  throughout; it equals half the number of calls to `AddElement`.)

Finally, the last `Insert` during rebalance is atypically expensive. The element moved from one heap to the other is necessarily an extreme element of both heaps, and when added to the new heap it bubbles all the way to the root with  $\log_2 N$  comparisons (unless there are duplicate minimal elements).

So rebalance requires roughly  $3 \log_2 N + \gamma$  comparisons, and the expected number of comparisons for `AddElement` is about  $(3/4) \log_2 N + \gamma$ . We haven't counted the single comparison that selects which heap will contain the new element. We'll continue to count only comparisons during heap operations, hence `AddElement` always uses one more comparison than we say.

### 3 Heap Operations

Our first step toward further progress is to sharpen the heap operations.

Start with `DeleteMin` and note that we never use this operation without an immediately preceding `Insert` into the same heap. We do well to combine these two: Replace the root with the new element and bubble that element down as far as it goes, returning the original root element as the result. Call this operation `DeleteMinThenInsert`. (The basic algorithm actually does the `DeleteMin` after the `Insert`, not before, but doing the `DeleteMin` first is slightly cheaper and doesn't affect the result since the new element to be inserted can be no smaller than the current heap minimum.)

With `DeleteMinThenInsert` we save an `Insert` completely. Furthermore, `DeleteMinThenInsert` is even cheaper than `DeleteMin` because now the element bubbling down from the root *is* random, and therefore more often comes to rest before reaching the bottom. We ignore this savings and continue to count `DeleteMinThenInsert` as  $2 \log_2 N$  comparisons.

We've already mentioned that the second (and now only) `Insert` during rebalance is especially costly since the new element is no larger than the heap minimum. Call this operation `InsertMin` (or `InsertMax` for a `MaxHeap`). In the next section we provide a special implementation of this operation. Here we point out a foregone opportunity for savings: `InsertMin` can be implemented with zero comparisons by blindly swapping the new element to the root of the heap, making an expensive heap operation absolutely free. In practice this saves almost no time and can slow things down when there

are duplicate elements. In any case we've promised not to cheat in this way.

In summary, we change the rebalancing case of the algorithm from

```
BigElements.Insert(newval);
SmallElements.Insert(BigElements.DeleteMin());
```

to

```
SmallElements.InsertMax(
    BigElements.DeleteMinThenInsert(newval));
```

(with corresponding change in the parallel case) yielding a small savings of  $\gamma/4$  in the expected cost of `AddElement`.

## 4 Slots

Suppose we give each heap a new location called a *slot*. The slot either is empty or contains a single element. If the slot contains an element, that element is the extreme element of the heap, that is, the smallest element in a `MinHeap` or the largest in a `MaxHeap`; we say that the slot sits just above the root of the tree. In accord with our convention we'll always say that an occupied slot holds the smallest heap element.

The three heap operations become:

- `Insert(x)`: Perform normal heap `Insert`. Then, if  $x$  lands at the root, and the slot is occupied, and  $x$  is smaller than the element in the slot, exchange the contents of the root with the contents of the slot. The slot's state does not change.
- `InsertMin(x)`: If the slot is empty, just put  $x$  into the slot. Otherwise, perform `Insert(x)` as above; the new element ends up in the slot unless there are duplicate values. The slot always ends up occupied.
- `DeleteMinThenInsert(x)`: If the slot is occupied, empty it and return its contents after doing heap `Insert(x)` as above. Otherwise, just do standard `DeleteMinThenInsert`. The slot always ends up empty.

These changes add no comparisons to `Insert` unless the new element is the smallest or second-smallest in the heap. But `InsertMin` becomes essentially free when the slot is empty, and expensive `DeleteMinThenInsert` becomes cheap `Insert` when the slot is occupied.

Now, when do we realize gains from favorable state of the slot? Both slots are initially empty. `Insert` does not change the state of the slot, so nothing happens until the first rebalance. At that point one heap executes `InsertMin` and its slot becomes occupied. The other heap executes `DeleteMinThenInsert` and its slot remains empty.

One of two things happens at rebalance when one slot is empty and the other occupied: Either *neither* slot's state is favorable to the heap operations, in which case the rebalance is as expensive as ever and neither slot changes state, or *both* slots are in favorable state, in which case the rebalance requires only a single cheap `Insert` and the slots switch state, the occupied one becoming empty and the empty one occupied. That is, each rebalance is either expensive and leaves the slots alone, or is cheap and switches their states. By symmetry, these two cases occur with equal probability on random input. (It's never the case that both slots are occupied, nor are they ever both empty after the first rebalance.)

The bottom line is that the expensive case of the algorithm now occurs only one time in eight, rather than one in four, and all other cases require only a single `Insert`. The expected number of comparisons in `AddElement` becomes  $(3/8)\log_2 N + (7/8)\gamma$ .

## 5 Burrows

One slot is good. More slots are better.

Let's attach  $k$  slots, called the *burrow*, to our heap implementation. The burrow consists of slots organized as a stack. If the burrow is nonempty, its top element is the heap's minimum element, and its bottom element is the element next smaller than the one in the root of the tree. (The burrow is so named because it lies "under" the root, even though we draw trees upside down with root on top. So the top of the burrow is farthest from the root.)

The heap operations now work like this:

- `Insert(x)`: Perform `Insert(n)` as in a slotless heap. Then, if  $x$  lands at the root, and the burrow is nonempty, compare  $x$  to the bottom element of the burrow. If  $x$  is smaller, swap these two, then compare  $x$  with the next element up the burrow. Continue until  $x$  reaches either a smaller element or the top of the burrow. The number of elements in the burrow is unchanged.
- `InsertMin(x)`: If the burrow is not full, push  $x$  onto the top of the burrow. Otherwise, perform `Insert(x)` as above;  $x$  rises to the top

of the burrow (unless there are duplicate minimal elements) and the burrow remains full.

- **DeleteMinThenInsert**( $x$ ): If the burrow is not empty, pop it, and return its contents after doing **Insert**( $x$ ) as above. If the burrow is empty, do **DeleteMinThenInsert**( $x$ ) in the usual (slotless) way.

How does the burrow affect running time? As before, only rebalancing changes the number of occupied slots in either burrow. And whenever we empty a slot in one heap we fill a slot in the other heap, and vice versa except during an initial “burrow-filling” period. At any time after this initial period, one burrow will have  $i$  occupied slots and the other  $k - i$  occupied slots for some  $0 \leq i \leq k$  (indeed, the two burrows could be stored in a single array of size  $k$ ). Thus there are exactly  $k + 1$  possible states of the burrows. When we need to rebalance, exactly one of these states is unfavorable and yields an expensive rebalance; in any of the other states, the rebalance requires only a single **Insert**.

With random inputs we find ourselves in each state with equal probability. To see this let  $p_i$  be the probability of being in state  $i$  after a rebalance, and note that for  $0 < i < k$  we have  $p_i = (p_{i-1} + p_{i+1})/2$  since from each state we transition to the adjacent states with equal probability. Moreover,  $p_0 = (p_0 + p_1)/2$  and  $p_k = (p_{k-1} + p_k)/2$  since the state doesn’t change after a rebalance when the burrows are in unfavorable state. These  $k + 1$  equations in  $k + 1$  unknowns have unique solution  $p_i = 1/(k + 1)$  for each  $i$ .

Thus the probability that a given rebalance is expensive is  $p_0/2 + p_k/2 = 1/(k + 1)$ . Since we rebalance with probability  $1/4$ , the probability of an expensive rebalance is  $1/4(k + 1)$ . Note that for  $k = 0$  and  $k = 1$  this expression has the correct values as determined in previous sections.

An expensive rebalance occurs with probability  $1/4(k + 1)$  and consists of a **DeleteMinThenInsert** with empty burrow, using  $2 \log_2 N$  comparisons as before, plus an **InsertMin** with a full burrow, using  $\log_2 N + k$  comparisons since the inserted element always comes to the top of the burrow.

In all other cases we do just an **Insert**. The burrow will on average be half full, with  $k/2$  occupied slots. Therefore with probability  $k/2N$  a new element enters the burrow, and if it does it travels on average halfway up the burrow. So elements that enter the burrow use  $k/4$  comparisons in addition to  $\log_2 N$  comparisons required to reach the burrow. Other elements use  $\gamma$  comparisons as before. Thus the expected number of comparisons for **Insert** is  $(1 - k/2N)\gamma + k/2N(\log_2 N + k/4)$ .

Putting it all together, the expected number of heap comparisons for

`AddElement` is at most

$$(1 - 1/4(k + 1))((1 - k/2N)\gamma + (k/2N)(\log_2 N + k/4)) \\ + (1/4(k + 1))(3\log_2 N + k)$$

which is less than

$$(3/4k)\log_2 N + \gamma + 1/4 + (k/2)(\log_2 N)/N + k^2/8N$$

A consequence of this expression is that if the number of slots grows dynamically,  $k = \Omega(\log N)$ , the algorithm runs in expected time  $\Theta(1)$ . It's not hard to let  $k$  grow with  $N$  if we can expand the array of slots; we might check  $N$  at each expensive rebalance and sometimes allocate a new slot.

We also see that we can have many, many slots without asymptotic expected cost. As long as  $k = o(\sqrt{N})$  the expected number of comparisons remains  $\gamma + 1/4 + o(1)$ . If not for this condition we might simply permit the burrow to grow without bound, making all `InsertMins` free. But doing so makes the number of occupied slots a (bounded) random walk, and its maximum extent will be  $\Omega(\sqrt{N})$ .

## 6 $d$ -ary Heaps

Instead of using binary heaps for priority queues, we can use  $d$ -ary heaps, that is, increasing trees in which each node (with possibly a single exception) has either zero or  $d$  children. Such heaps can be kept in an array exactly the same way as binary heaps: the root is at index 1, the children of the node at index  $i$  are at indices  $di - d + 2, di - d + 1, \dots, di + 1$ , and the parent of the node at index  $i$  is at index  $\lfloor (i + d - 2)/d \rfloor$ .

(There is a problematic hidden cost here: With  $d$ -ary heaps we must multiply and divide by  $d$  to traverse the heap, and this integer arithmetic can be expensive. In practice,  $d$  should be a power of 2 so that bit shifts can replace arbitrary multiplication and division.)

A  $d$ -ary heap is shallower than a binary heap by a factor of  $\log d / \log 2$ . With shorter path to the root, `Insert` is cheaper, and if we consider `Insert` alone there's no downside to increasing  $d$  indefinitely—in the limit we have just a set plus distinguished smallest element, and `Insert` requires exactly one comparison. We define  $\gamma_d$  as the expected number of comparisons for `Insert` into a  $d$ -ary heap during basic online median finding (i.e., with no slots). Table 1 has several experimentally-determined approximations to  $\gamma_d$ .

The effect on `DeleteMin` of increased  $d$  is less monotonic. Although there are fewer levels to bubble down, there are  $d$  comparisons at each level

$d$	$\gamma_d$	$d$	$\gamma_d$	$d$	$\gamma_d$
2	2.06	6	1.25	32	1.05
3	1.56	7	1.21	64	1.025
4	1.39	8	1.18	128	1.013
5	1.30	16	1.10	256	1.007

Table 1: Estimated upper bounds on  $\gamma_d$ , which is imprecisely defined and may not exist, for selected  $d$ , chiefly powers of 2.

to swap the travelling element with its smallest child. Assuming as before that bubbling continues to the bottom, the total number of comparisons is  $d \log_d N$ , which has a minimum at  $d = e$ . The best  $d$  for a given application depends on the mix of **Inserts** and **DeleteMins**; as the proportion of **Inserts** increases, the optimal value of  $d$  also increases.

With  $d$ -ary heaps the expected comparison count for **AddElement** is

$$((d + 1)/4k) \log_d N + \gamma_d + 1/4 + (k/2)(\log_d N)/N + k^2/8N$$

and we find ourselves in the following interesting situation: As we add slots, the proportion of expensive rebalances goes down, all other calls on **AddElement** requiring just an **Insert**. This reduction in the number of **DeleteMins** then makes it worth while to increase  $d$ . But doing so shifts comparisons from the **Inserts** to the **DeleteMins**, making it advantageous to add yet more slots! This virtuous cycle terminates only when there are so many slots that they add appreciably to the cost of **Insert**. In the next section we address further the question of the best choice of  $d$  and  $k$ .

## 7 Circular Burrows and Burgeoning Heaps

Although the expected time for adding an element is low, the worst-case time can be very bad. Indeed, one way to choose the number of slots is to consider how bad a worst case is tolerable.

Suppose for example that we have a billion elements (half a billion in each heap) with  $d = 128$  and  $k = 6000$ . Although we expect an expensive rebalance only about every 24000 calls on **AddElement**, that rebalance may take  $129 \lceil \log_{128} 2^{29} \rceil + 6000 = 6645$  comparisons, six thousand times worse than the expected case. Even a simple (non-rebalance) **Insert** in the worst case may require 6004 comparisons. Another coding trick considerably mitigates this problem and further improves the expected time.

We implement the burrow as a deque instead of a stack, permitting manipulation at both ends. The burrow is still physically stored in an array, but that array is now “circular”: the ends of the burrow can be anywhere, and the burrow may wrap around from the high-index end to element 0. The extra index arithmetic that this scheme introduces can be minimized by using sentinels at the ends of the array.

We can now implement `InsertMin` much more efficiently. If the burrow is not full, just push the new element onto the top as before. Otherwise we must free up a slot. We previously did this by putting the new element into a new leaf of the tree and bubbling it to the top of the burrow; during this process the value in the bottom slot of the burrow moves into the root of the tree. Instead, we can take the value from the *root* of the tree, put it into a new leaf, and bubble that value upward; it stops just below the original root. Now the original root location is empty. We move the bottom element of the burrow into that spot, creating a free space in the burrow’s array. Finally, we transfer that free space from the bottom to the top of the burrow by simply adjusting the burrow boundaries within the array, and we put the new element in that free space. (A heap implementation with arbitrary arity and circular burrow can be found at <http://denenberg.com/MinHeap.java>.)

The result of this change is that `InsertMin` requires only the  $\log_d N$  comparisons to bubble an element to the top of the tree;  $k$  comparisons vanish. In the example above, the worst-case number of comparisons for `AddElement` goes from 6645 to 645.

The circular burrow can also be used to improve `Insert`. If a new element bubbles all the way to the root of the tree, we needn’t simply bubble it up the burrow. Instead, we first compare it to the element in the middle of the burrow. If it’s larger than this element, we bubble it up as usual. But if the new element is smaller than the middle element, we start from the *top* of the burrow and bubble it *down* into place, adjusting the boundaries of the burrow to transfer the new free space from the bottom to the top. With this trick we never have to bubble a new element through more than half the burrow; in the example, a worst-case `Insert` takes only 3004 comparisons.

These improvements also improve the expected number of comparisons. The change to `Insert` has small significance since we get its benefit so rarely; in the formula for the expected number of comparisons in `AddElement`, the term  $k^2/8N$  is cut in half to  $k^2/16N$ .

The improvement to `InsertMin` is more important. The  $k$  comparisons that create space in the burrow during an unfavorable rebalance are responsible for the term  $1/4$  in the expected-case number of comparisons ( $k$  comparisons with probability about  $1/4k$ ), and these no longer exist. So the

expected number of comparisons, assuming that the maximum number of slots is both  $\omega(\log N)$  and  $o(\sqrt{N})$ , becomes  $\gamma_d + o(1)$ . To find the optimal  $k$  we differentiate the expression for the number of comparisons, a calculation made simpler by ignoring the small term  $(k/2N)\log_d N$ . The best  $k$  for given  $N$  and  $d$  turns out to be approximately  $\sqrt[3]{2N(d+1)\log_d N}$ .

But for best theoretical performance, not only  $k$  but also  $d$  must increase with  $N$ . The standard implementation of heaps makes this difficult; it is not practical to restructure a heap on fly to increase its arity. Instead, we can use a data structure we might call a **burgeoning heap**, in which the arity of the tree is not constant but increases as the tree becomes higher.

Such a heap can be stored in an array without loss of space in reasonably practical manner: We keep two auxiliary arrays that store, for each level of the tree, the arity at that level plus a “displacement” that permits us to waste no array locations. If these arrays are called **Arity** and **Displacement**, then the index of the leftmost child of the node with index  $i$  is

$$i * \text{Arity}[\text{d}(i)] + \text{Displacement}[\text{d}(i)]$$

where  $\text{d}(i)$  is the depth of node  $i$ . This scheme is reasonable in practice because we always traverse the tree from top to bottom or bottom to top, hence we can keep track of the current level rather than computing or storing it. The rate of burgeoning can be anything we like, since at the start of each new heap level we can pick  $d$  arbitrarily and then adjust the displacement so the next array element used is the next one free. (Code for the burgeoning heap modifications can be found at <http://denenberg.com/MinBurgeoningHeap.java>.)

With burgeoning heaps and growing burrow we can let both  $k$  and  $d$  increase with  $N$ , yielding an algorithm that runs with expected number of comparisons  $1 + o(1)$  (still not counting 1 in the main algorithm).

## 8 Results

Table 2 gives some experimental results for various  $N$ ,  $k$ , and  $d$ . We report the average number of comparisons during heap operations calculated over many runs of the algorithm. Number of comparisons is measured “at the margin”: we count during only the final 10% of the calls on **AddElement**. We compare the basic method as described in Figure 1 (using slotless binary heaps) against the final algorithm with circular buffer and fixed  $k$  and  $d$ .

$N$	mean comps, $d = 2 \ \& \ k = 0$	alternative $d$	alternative $k$	mean comps, this $d \ \& \ k$
100	6.87	8	10	2.10
1000	9.04	32	100	1.70
100000	14.26	32	2500	1.31
1000000	16.50	64	2000	1.23
100000000	21.72	128	3500	1.072

Table 2: Experimental results. “Mean comps” counts comparisons only during the final 10% of calls on `AddElement` and doesn’t count the single comparison that selects the appropriate heap.

## References

- [BF] Blum, Floyd, Pratt, Rivest, and Tarjan, “Time bounds for selection,” *J. Comput. System Sci.* **7** (1973) 448-461.
- [BS] Bollabas and Simon, “Repeated random insertion into a priority queue,” *J. Alg* **6** (1985) 466–477.
- [CH] Cantone and Hofri, “Analysis of An Approximate Median Selection Algorithm,”  
<ftp://ftp.cs.wpi.edu/pub/techreports/pdf/06-17.pdf>
- [DZ] Dor and Zwick, “Selecting the Median,” *SIAM J. Comput.* **28**, **5** (May 1999) 1722–1758.
- [LD] Lewis and Denenberg, *Data Structures and Their Algorithms*, Harper-Collins, 1991, pp 110ff.
- [PS] Porter and Simon, “Random insertion into a priority queue structure,” *IEEE Transactions on Software Engineering* **1** (1975), 292–298.

**Acknowledgement:** The author thanks Google Inc. for posing the problem and for providing ample leisure time to work on the solution.