Randomized Median Finding[1]

- In this lecture we will see a randomized algorithm which takes an array $A[1:n]$ (any array, worst-case) and finds the median of A (the $n/2$th largest/smallest entry) with high probability. This probability will tend to 0 as $n \to \infty$. The number of queries made by the algorithm will be roughly $1.5n$. Apart from being a simple algorithm and a nice example of how sampling helps, this result is also significant because the best *deterministic* algorithm for median finding makes $2.95n$ comparisons[2], and it is known that *any* deterministic algorithm must make at least $2.01n$ comparisons[3].

- *Idea.* The idea is simple. One samples a small random subset R of the elements of A of s elements. Think of $s \approx n^{3/4}$, and thus this can be sorted in $o(n)$ time. The first thought that perhaps comes to ones mind is : the median of R should be the median of A. This is very unlikely. However, the median of R and median of A can't be "too far" in the sorted order of A. More precisely, one looks at two elements a and b, where a is the $(\frac{s}{2}-t)$th element in R and b is the $(\frac{s}{2}+t)$th element in R where $t \ll s$ is chosen such that with high probability, (a) the median of A lies between a and b, and (b) not too many (only $o(n)$) elements of A lie between a and b.

  If this is indeed the case, then the algorithm is clear: find a and b in R by sorting R, find the rank of a in A taking one linear scan of A, find all elements Q between a and b in A with another linear scan, and since Q is guaranteed to contain the median and have $o(n)$ elements, sort Q to find what you want. The rest of these notes give details of this simple idea. This algorithm is due[4] to Robert Floyd and Ronald Rivest.

- Let us first do some calculations. Throughout we will have integer parameters s and t which we will set at the very end. As we go along, we will keep remarking what are the conditions we need on s and t, and thus, obtain the final values. Let R be a random sample of A where every element of A is picked in R independently with probability $p = \frac{s}{n}$. Thus, we expect $\mathbf{Exp}[|R|] = s$ and Chernoff bound gives us the following.

  Claim 1. $\mathbf{Pr}[|R| \geq 2s] \leq e^{-s/2}$

  Remark: *Indeed, the probability $|R|$ is more than $s + C\sqrt{s}$ is $\approx e^{-C^2}$, and the probability $|R| \geq s + s^{2/3}$ itself would be $o(n)$. We just chose $2s$ for convenience.*

- Next, let S be the subset of "small" elements of A which are at most the median, and thus $|S| = \frac{n}{2}$. Let B be the subset of "big" elements A which are larger than the median. Define $X := |S \cap R|$, that is, the number of small elements picked in R. Once again, $\mathbf{Exp}[X] = \frac{s}{2}$ since each of the $n/2$ elements are picked with probability $\frac{s}{n}$, and Chernoff gives us

[1] *Lecture notes by Deeparnab Chakrabarty. Last modified : 31st March, 2021*
These have not gone through scrutiny and may contain errors. If you find any, or have any other comments, please email me at deeparnab@dartmouth.edu. Highly appreciated!

[2] "*Selecting the Median*", D. Dor and U. Zwick, SIAM J. Computing, 28(5), 1722–1758

[3] "*On Lower Bounds for Selecting the Median*", D. Dor, J. Hastad, S. Ulfberg, and U. Zwick. SIAM J. Disc. Math., 14(3), 299–311

[4] "*Expected time bounds for selection.*", R. W. Floyd and R. L. Rivest, Communications of the ACM, 18: 165–172, 1975

Claim 2. $\mathbf{Pr}\left[X \leq \frac{s}{2} - t\right] \leq e^{-t^2/s}$ and $\mathbf{Pr}\left[X \geq \frac{s}{2} + t\right] \leq e^{-2t^2/3s}$

Proof. X can be written as $\sum_{i \in S} X_i$ where $X_i = 1$ if $i \in R$ and $X_i = 0$ otherwise. X is thus a sum of independent random Bernoullis with $\mathbf{Exp}[X] = \frac{s}{2}$. Next, set[5] $\varepsilon = \frac{2t}{s}$ and apply Chernoff bound to get

$$\mathbf{Pr}\left[X \leq (1-\varepsilon)\frac{s}{2}\right] \leq e^{-\varepsilon^2 \mathbf{Exp}[X]/2} \leq e^{-t^2/s}$$

The other inequality follows using the upper tail bound. □

Here is the important consequence of the above fact.

Claim 3. Let a be the $(\frac{s}{2} - t)$th ranked element in R and b be the $(\frac{s}{2} + t + 1)$th ranked element of R. Then with probability $\geq \left(1 - 2e^{-2t^2/3s}\right)$, the median of A lies between a and b.

Proof. Note that if $X \geq \frac{s}{2} - t$, that is, we pick more than $\frac{s}{2} - t$ small elements in R, then the $(\frac{s}{2} - t)$th element of R, that is a, must be small. Similarly, if $X \leq \frac{s}{2} + t$, that is, we pick at most $\frac{s}{2} + t$ small elements in R, then the $(\frac{s}{2} + t + 1)$th element of R, that is b, must be big. □

Remark: *Since we want to succeed with high probability, we must have $2t^2 \ll 3s$. At this point $\frac{2t^2}{3s} \leq \ln(1/\delta)$ would give $(1-\delta)$ as failure probability.*

- Given an element $x \in A$, let $\mathsf{rank}_A(x)$ be its rank in A; the minimum has rank 1, the median rank $n/2$, the maximum rank n. We have established that whp $\mathsf{rank}_A(a) \leq \frac{n}{2}$ and $\mathsf{rank}_A(b) > \frac{n}{2}$. Next we show that whp these ranks, $\mathsf{rank}_A(a)$ and $\mathsf{rank}_B(a)$ are indeed actually pretty close to $\frac{n}{2}$.

Claim 4. With probability $\geq 1 - e^{-\frac{2t^2}{3(s-4t)}}$, we have $\mathsf{rank}_A(a) \geq \frac{n}{2} - \frac{2tn}{s}$.

Proof. To prove this, we define T to be the smallest $\frac{n}{2} - \frac{2tn}{s}$ items of A. Let's call them the "tiny" elements of A. Note that elements which are not tiny have rank bigger than the RHS in the claim. So, $T \subseteq S$. Let $Z = |R \cap T|$ be the number of tiny elements we pick in R. Note that if $Z < \frac{s}{2} - t$, then a which is the $(\frac{s}{2} - t)$th ranked items in R *must* not be tiny. Which, in turn, would imply $\mathsf{rank}_A(a) \geq \frac{n}{2} - \frac{2tn}{s}$. Thus,

$$\mathbf{Pr}\left[\mathsf{rank}_A(a) \geq \frac{n}{2} - \frac{2tn}{s}\right] \geq \mathbf{Pr}\left[Z < \frac{s}{2} - t\right]$$

To show that the RHS is big, we argue that its complement is small. To this end, for every $i \in T$, define $Z_i = 1$ if $i \in R$ and 0 otherwise. $Z = \sum_{i \in T} Z_i$, and thus, $\mathbf{Exp}[Z] = \frac{s}{n} \cdot |T| = \frac{s}{2} - 2t$. Therefore,

$$\mathbf{Pr}\left[Z \geq \frac{s}{2} - t\right] = \mathbf{Pr}\left[Z \geq \underbrace{\left(\frac{s}{2} - 2t\right)}_{\mathbf{Exp}[Z]} \cdot (1+\varepsilon)\right] \underbrace{\leq}_{\text{Chernoff LT}} e^{-\varepsilon^2 \mathbf{Exp}[Z]/3} \tag{1}$$

[5] so, $2t \leq s$, but we will get a stronger req soon

where, ε is chosen such that $\left(\frac{s}{2} - 2t\right)(1+\varepsilon) = \frac{s}{2} - t$. A little manipulation shows that one can choose

$$\varepsilon = \frac{2t}{s - 4t}$$

which, when substituted in (1) gives the desired claim. □

A similar calculation (which I urge you to do) gives the following claim on the $\mathsf{rank}_A(b)$.

Claim 5. With probability $\geq 1 - e^{-\frac{t^2}{(s+4t)}}$, we have $\mathsf{rank}_A(b) \leq \frac{n}{2} + \frac{2tn}{s}$.

Remark: *I have been careful with constants till now just to show that one can. For aesthetic reasons, and also to stress the important points, I am now going to move to Big-Oh notation. Let's start by noticing that we will choose $t^2 \ll s$ so that the probability of bad events are low. Therefore, $s - 4t$ and $s + 4t$ are all roughly s. This means the probabilities of the claimed events not occurring in the above three claims are $e^{-O(t^2/s)}$.*

The above claim imply the following lemma which will be used in the analysis of the algorithm.

Lemma 1. Let R be a sample of A where every element is picked with probability $\frac{s}{n}$. Let a be the $\left(\frac{s}{2} - t\right)$th element in R and let b be the $\left(\frac{s}{2} + t + 1\right)$th element in R, where $t \ll s$. Then with probability $1 - e^{-O(t^2/s)}$, we have the following:

$$\frac{n}{2} - \frac{2tn}{s} \leq \mathsf{rank}_A(a) \leq \frac{n}{2} \leq \mathsf{rank}_A(b) \leq \frac{n}{2} + \frac{2tn}{s}$$

- ***Algorithm and Analysis.***

procedure RANDOMIZED-MEDIAN(A):
  Set parameters $s = \lceil n^{2/3} \rceil$ and $t = \lceil 10 n^{1/3} \log n \rceil$.
  Sample R by picking each element of A independently with probability $\frac{s}{n}$.
  if $|R| > 2s$ **then**:
    return ABORT.
  Sort R. ▷ *Takes $O(s \log s)$ comparisons.*
  $a \leftarrow \left(\frac{s}{2} - t\right)$th element in R, and $b \leftarrow \left(\frac{s}{2} + t + 1\right)$th element of R. ▷ *Using sorted order of R takes $O(1)$ time.*
  Find $\mathsf{rank}_A(a)$ using $(A_1, A_2) \leftarrow$ PIVOT(A, a). ▷ *This takes n comparisons.*
  if $\mathsf{rank}_A(a) \notin [\frac{n}{2} - \frac{2tn}{s}, \frac{n}{2}]$ **then**:
    return ABORT.
  Find $\mathsf{rank}_A(b)$ using $(Q, A_3) \leftarrow$ PIVOT(A_2, b). ▷ *This takes $n - \mathsf{rank}_A(a) \leq 0.5n + \frac{2tn}{s}$ comparisons. $\mathsf{rank}_A(b) = |Q| + \mathsf{rank}_A(a)$.*
  if $\mathsf{rank}_A(b) < \frac{n}{2}$ or $|Q| > 4ts/n$ **then**:
    return ABORT.
  Sort Q and return the $\frac{n}{2} - \mathsf{rank}_A(a)$th item in Q. ▷ *Sorting Q takes $O(\frac{nt}{s} \log(\frac{nt}{s}))$ comparisons.*

Theorem 1. RANDOMIZED-MEDIAN(A) makes $\leq 1.5n + o(n)$ comparisons and either returns the median or ABORTs. The probability of ABORT is $\leq n^{-O(1)}$.

Proof. The probability of the ABORTs in Line 4, Line 9, and Line 12 is at most $e^{-O(t^2/s)}$ by Lemma 1 and Claim 1. If $t^2 = \Omega(s \log n)$, then this probability is $\leq \frac{1}{n^{O(1)}}$. This explains the relation between s and t in Line 2 If no ABORTs occur, then the median of A lies in Q, and then Line 14 returns the correct median.

The number of comparisons in Line 8 and Line 11 are $\leq 1.5n + \frac{2tn}{s}$. The other comparisons are in Line 6 and Line 14. These take $O(s \log s) + O(\frac{nt}{s} \log(nt/s))$. Now we see how the parameters are chosen. We need (a) $t^2 \gg s$, (b) $s \log s \ll n$, and (c) $nt/s \ll n$. If we balance the $s \log s \approx (nt/s) \log(nt/s)$, and using $t^2 \approx s$, we get $s \approx n^{2/3}$. Using the parameters chosen in Line 2, the theorem follows. □

Learning Tidbits:

- Algorithm Design: *To solve a problem on "big data", many times you can take a small sample, solve the problem (or related problem), and then try to port back. For the median, we didn't solve the median on the sample (that didn't work), but found stuff close to the median of the sample, and ported back.*
- Analysis: *Once again, Chernoff bounds allowed us to argue that what we do in the sample, ports back "as we expected" (or close to that). The other thing to digest is that the same random sample* simultaneously *is good (ie, behaves as "expected") for multiple sets. In this case, think : "small" elements* and *"tiny" elements.*