

Algorithm for Computing the Median

The **median** of a set X of n distinct elements is the $\lceil \frac{n}{2} \rceil$ largest element in the set.

If $n = 2k + 1$, the median element is the $k + 1$ -th element in the sorted order.

Easily computed through sorting in $O(n \log n)$ time. There exists a complicated $O(n)$ deterministic algorithm.

Randomized Median Algorithm

Input: A set of $n = 2k + 1$ elements from a totally ordered universe.

Output: The $k + 1$ -th largest element in the set.

1. Pick a (multi)-set R of $s = n^{3/4}$ elements in S , chosen independently and uniformly at random with replacement. Sort the set R .
2. Let d be the $(\frac{1}{2}n^{3/4} - \sqrt{n})$ th smallest element in the sorted set R .
3. Let u be the $(\frac{1}{2}n^{3/4} + \sqrt{n})$ th smallest element in the sorted set R .
4. By comparing every element in S to d and u compute the set $C = \{x \in S : d \leq x \leq u\}$, and the numbers $\ell_d = |\{x \in S : x < d\}|$ and $\ell_u = |\{x \in S : x > u\}|$.
5. If $\ell_d > n/2$ or $\ell_u > n/2$ then FAIL.
6. If $|C| \leq 4n^{3/4}$ then sort the set C , otherwise FAIL.
7. Output the $(\lfloor \frac{n}{2} \rfloor - \ell_d + 1)$ st element in the sorted order of C .

Theorem

Theorem 1. *With probability at least $1 - O(n^{-1/4})$, the above algorithm finds the median. The running time of the algorithm is $2n + o(n)$.*

Intuition

- We can sort sets of size $o(n)$ in linear time.
- The sample of R elements are spaced “more or less” evenly among the elements of X .
- W.h.p. more than $\frac{1}{2}n^{3/4} - \sqrt{n}$ samples are smaller than the median.
- W.h.p. more than $\frac{1}{2}n^{3/4} - \sqrt{n}$ samples are larger than the median.
- W.h.p. the median is in the set C , and $|C| \leq 4n^{3/4}$.

Let Y_1 be the number of samples below or equal to the median.

Let Y_2 be the number of samples above or equal to the median.

The algorithm computes the median in $O(n)$ time if all the following three events hold:

1. $E_1 : Y_1 \geq \frac{1}{2}n^{3/4} - \sqrt{n}$.
2. $E_2 : Y_2 \geq \frac{1}{2}n^{3/4} - \sqrt{n}$.
3. $E_3 : |C| \leq 4n^{3/4}$.

What is the probability that the three random variables Y_1, Y_2 and $|C|$ are all within the required ranges?.

The sample space in execution of this algorithm is the set of all possible choices of $n^{3/4}$ elements from n , with repetitions. (The sample space has $n^{n^{3/4}}$ points.)

Each point in the sample space defines values for Y_1 , Y_2 and $|C|$.

Computing the probabilities directly is too complicated, instead we use bounds on deviation from the expectation.

Y_1 = the number of samples below or equal the median.

What is the probability that $Y_1 < \frac{1}{2}n^{3/4} - \sqrt{n}$

Viewing Y_1 as the sum of $n^{3/4}$ independent 0-1 random variable, each with expectation $1/2$ and variance $1/4$ we prove:

$$E[Y_1] > \frac{1}{2}n^{3/4}.$$

$$Var[Y_1] < \frac{1}{4}n^{3/4}.$$

Applying Chebyshev Inequality we get:

$$\Pr(\bar{E}_1 : Y_1 < \frac{1}{2}n^{3/4} - \sqrt{n}) \leq \Pr(|Y_1 - E[Y_1]| > \sqrt{n}) \leq$$

$$\frac{\text{Var}[Y_1]}{n} = \frac{n^{3/4}/4}{n} = \frac{1}{4}n^{-1/4}.$$

Similarly

$$\Pr(\bar{E}_2 : Y_2 < \frac{1}{2}n^{3/4} - \sqrt{n}) \leq \frac{1}{4}n^{-1/4}.$$

$$\Pr(\bar{E}_1 \cup \bar{E}_2) \leq \frac{2}{4}n^{-1/4}.$$

Recall: $E_3 : |C| \leq 4n^{3/4}$.

Lemma 1.

$$\Pr(\bar{E}_3) \leq \frac{1}{2}n^{-1/4}.$$

Define the following two events:

1. $\mathcal{E}_{3,1}$: at least $2n^{3/4}$ elements of C are greater than the median;
2. $\mathcal{E}_{3,2}$: at least $2n^{3/4}$ elements of C are smaller than the median.

If $|C| > 4n^{3/4}$, then at least one of the above two events occurs.

We bound $\mathcal{E}_{3,1}$: at least $2n^{3/4}$ elements of C are greater than the median;

At least $2n^{3/4}$ elements of C above the median \Rightarrow

u is at least the $\frac{1}{2}n + 2n^{3/4}$ smallest in $S \Rightarrow$

R had at least $\frac{1}{2}n^{3/4} - \sqrt{n}$ samples among the $\frac{1}{2}n - 2n^{3/4}$ largest elements in S .

Let X be the number of samples among the $\frac{1}{2}n - 2n^{3/4}$ largest elements in S . Let $X = \sum_{i=1}^{n^{3/4}} X_i$ where

$$X_i = \begin{cases} 1 & \text{the } i\text{-th sample in } \frac{1}{2}n - 2n^{3/4} \\ & \text{largest elements in } S \\ 0 & \text{otherwise.} \end{cases}$$

$$E[X_i] = E[(X_i)^2] = \frac{1}{2} - 2n^{-1/4}$$

$$\text{Var}[X_i] = E[(X_i)^2] - (E[X_i])^2 \leq \frac{1}{4}.$$

$$E[X] = \frac{1}{2}n^{3/4} - 2\sqrt{n}$$

$$\text{Var}[X] \leq \frac{1}{4}n^{3/4}$$

Applying Chebyshev's Inequality yields

$$\begin{aligned} \Pr(\mathcal{E}_{3,1}) &= \Pr(X \geq \frac{1}{2}n^{3/4} - \sqrt{n}) \\ &\leq \Pr(|X - E[X]| \geq \sqrt{n}) \\ &\leq \frac{\text{Var}[X]}{n} = \frac{\frac{n^{3/4}}{4}}{n} = \frac{1}{4}n^{-1/4}. \end{aligned}$$

Similarly,

$$\Pr(\mathcal{E}_{3,2}) \leq \frac{1}{4}n^{-\frac{1}{4}},$$

and

$$\Pr(\bar{E}_3) \leq \Pr(\mathcal{E}_{3,1}) + \Pr(\mathcal{E}_{3,2}) \leq \frac{1}{2}n^{-\frac{1}{4}}.$$

The probability that the algorithm succeeds is

$$\geq 1 - (\Pr(\bar{E}_1) + \Pr(\bar{E}_2) + \Pr(\bar{E}_3)) \geq 1 - \frac{1}{n^{1/4}}.$$