CS5314
Randomized Algorithms

Lecture 9: Moments and Deviation
(Randomized Median)

Objectives

- Compute the median of n numbers
- In fact, there is a deterministic algorithm, which runs in optimal $O(n)$ time … [so, how can we improve this??]
- We will see a simpler randomized algorithm which also runs in $O(n)$ time, but with a smaller hidden constant

Computing the Median

Definition: Let S be a set of n numbers. If x is the jth smallest number in S, we say the rank of x is j

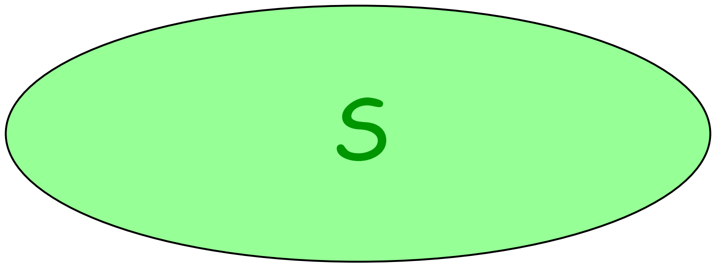
Definition: In a set of n numbers, median is the number whose rank is $\lceil n/2 \rceil$

Ex: $S = \{ 1, 3, 4, 6, 8, 13, 15, 22 \}$

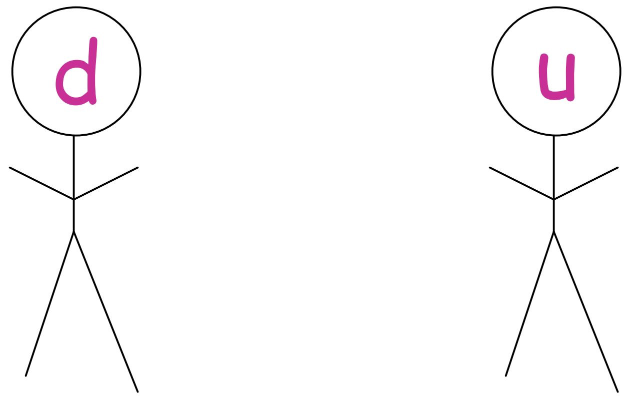
Median = 6

Randomized Median (idea)

Input:



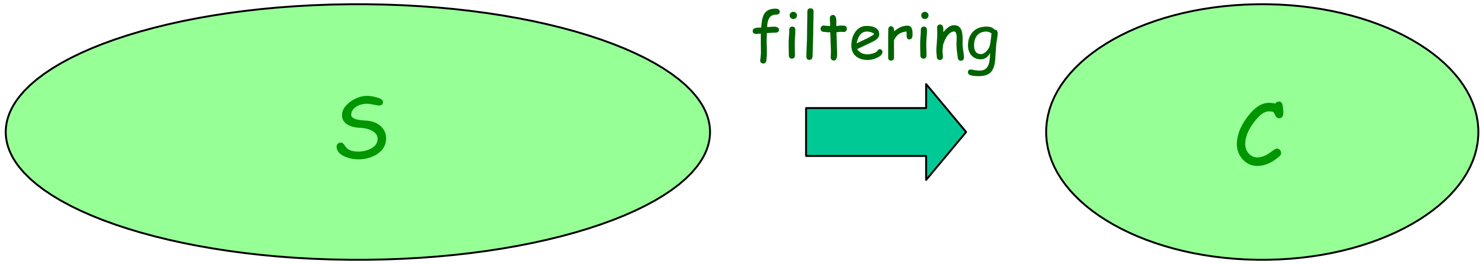
Step 1: Find two good “guards” d and u, so as to enclose the median m



$\leq$ m $\geq$ m

Randomized Median (idea)

Step 2: Use d and u to filter out those numbers which cannot be m



Candidate median after filtering

Step 3: If C is small enough, find m by brute-force

Details of the Algorithm

Step 1: Finding Guards
(the only step with randomization)

(i) Randomly pick $\lceil n^{3/4} \rceil$ numbers from S, independently and uniformly (with replacement)

(ii) Let R = multi-set of such numbers

(iii) Sort R

(iv) Set d = number in R whose rank is $\lfloor n^{3/4}/2 - n^{1/2} \rfloor$

Details of the Algorithm

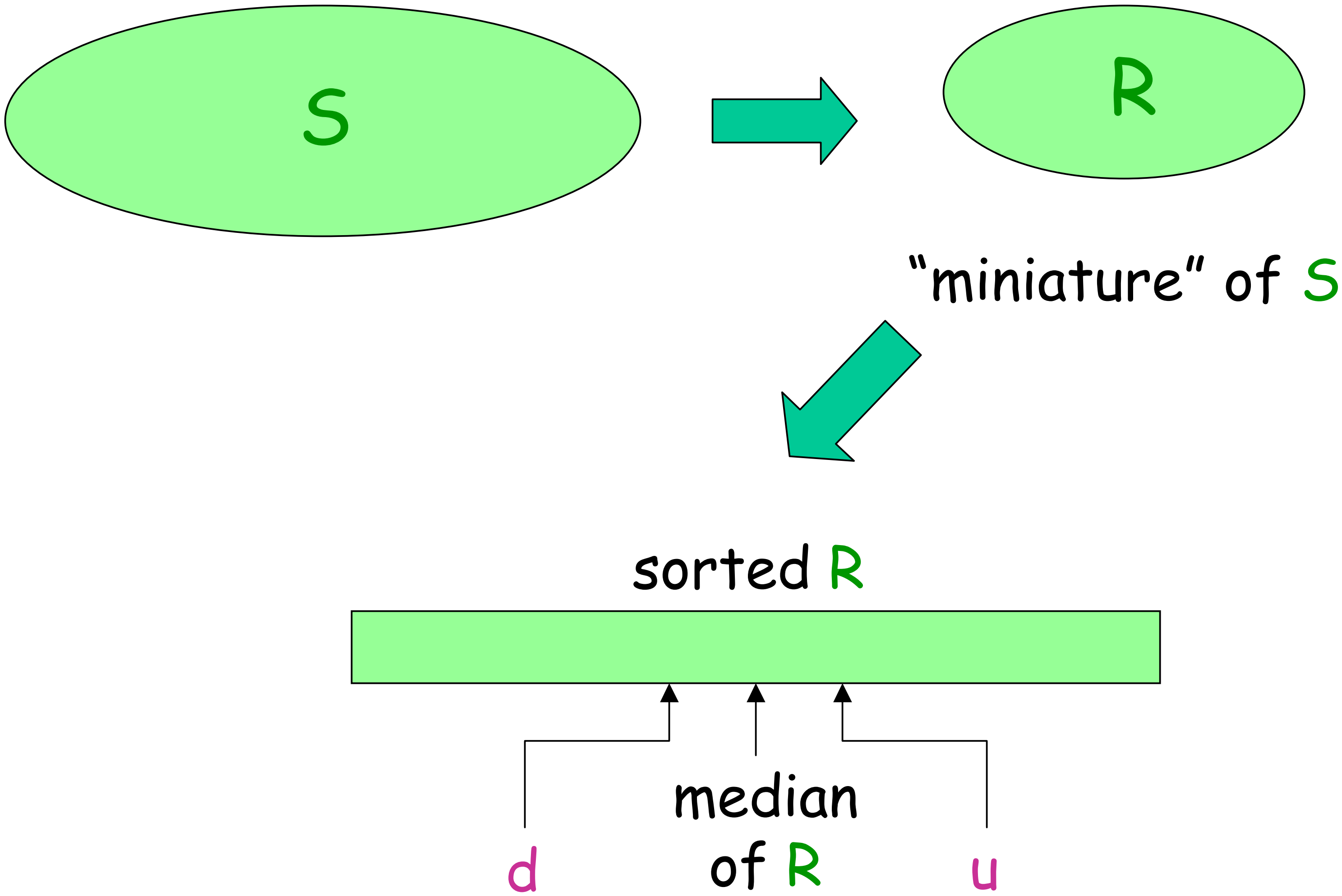
Step 1: Finding Guards [cont.]

(v) Set u = number in R whose rank is $\lfloor n^{3/4}/2 + n^{1/2} \rfloor$

(vi) Scan S to check if d and u encloses m

- if so, proceed to Step 2;
- if not, output FAIL immediately

What happens in Step 1 ?



Details of the Algorithm

Step 2: Filtering

(i) Scan S

(ii) Let C = set of numbers in S between d and u

(iii) Check if C is "small" enough

- if C has at most $4n^{3/4}$ numbers, proceed to Step 3;
- else, output FAIL immediately

Details of the Algorithm

Step 3: Finding median by brute force

(i) Let p = #numbers in S less than d (obtained in Step 1)

(ii) Sort C

(iii) Output the number in C whose rank is $\lfloor n/2 - p \rfloor$ ➔ which must be the median

Time and Correctness

Lemma: The randomized median always terminate in O(n) time. If it does not output FAIL, then the output number is the correct median

Proof:

(Time) Each step takes O(n) time

(Correctness) If it does not output FAIL, C always contains the median

Failure Probability

The algorithm will FAIL if and only if one of the following events occurs:

E_1 : $d >$ median m

E_2 : $u <$ median m

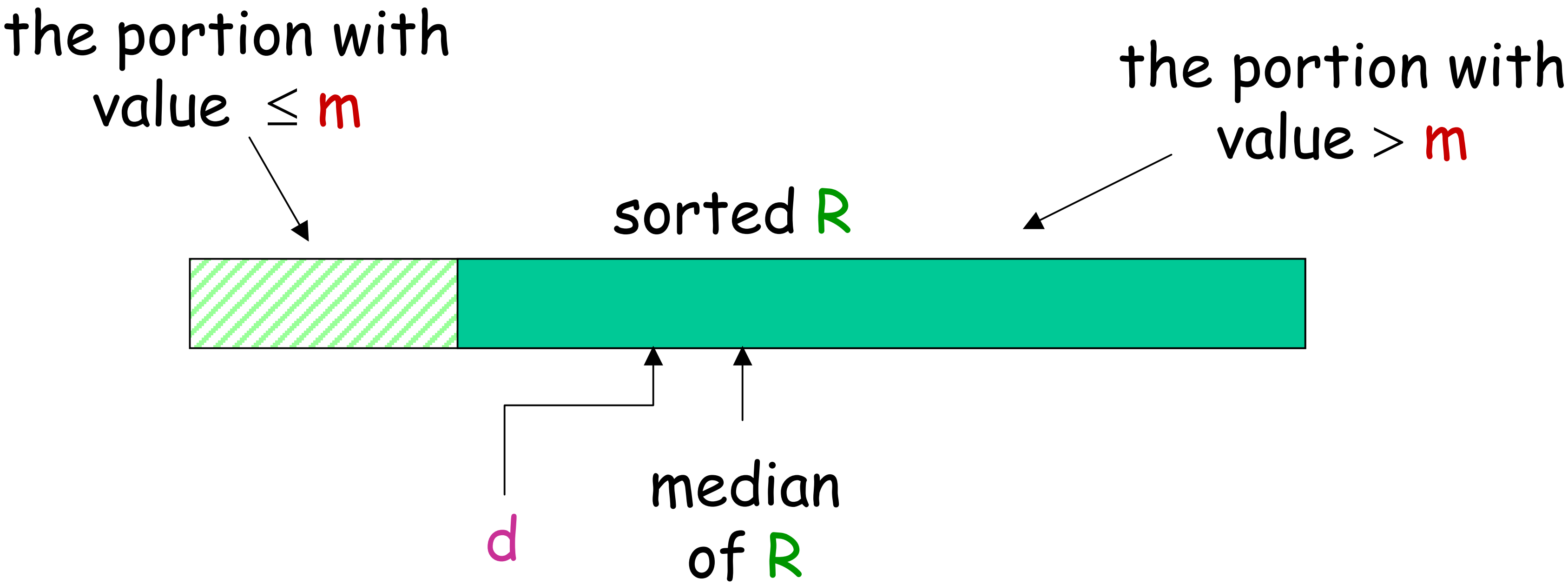
E_3 : C has more than $4n^{3/4}$ numbers

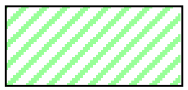
Thus, $\Pr(\text{FAIL}) = \Pr(E_1 \cup E_2 \cup E_3)$

$\leq \Pr(E_1) + \Pr(E_2) + \Pr(E_3)$

Bounding $\Pr(E_1)$

Question: When will d > median m ?



Answer: if and only if  has less than $\lfloor n^{3/4}/2 - n^{1/2} \rfloor$ numbers

Bounding $Pr(E_1)$

Let Y = size of

$$Pr(E_1) = Pr(Y < \lfloor n^{3/4}/2 - n^{1/2} \rfloor)$$

Let Y_j be an indicator that :

$Y_j = 1$ if j^{th} number of $R \le$ median

$Y_j = 0$ otherwise

$\Rightarrow$ $Y = Y_1 + Y_2 + \ldots + Y_{\lceil n^{3/4} \rceil}$

Bounding $Pr(E_1)$

If we can find $E[Y]$ and $Var[Y]$, then we can use Chebyshev inequality to bound $Pr(E_1)$

Note: $E[Y_1] = Pr(Y_1 = 1)$

$\geq 1/2$

$\rightarrow E[Y] \geq \lceil n^{3/4} \rceil /2$

$Var[Y] \leq \lceil n^{3/4} \rceil /4$

Bounding $\Pr(E_1)$

Thus,

$$\begin{aligned}
\Pr(E_1) &= \Pr(Y < \lfloor n^{3/4}/2 - n^{1/2} \rfloor) \\
&\leq \Pr(Y < n^{3/4}/2 - n^{1/2}) \\
&\leq \Pr(Y < E[Y] - n^{1/2}) \\
&\leq \Pr(\,|Y - E[Y]\,| \geq n^{1/2}) \\
&\leq \mathrm{Var}[Y] \,/\, (n^{1/2})^2 \\
&= O(1 / n^{1/4})
\end{aligned}$$

... which is small for large n

Similarly, $\Pr(E_2) = O(1 / n^{1/4})$

Bounding $\Pr(E_3)$

Question: When will C has more than $4n^{3/4}$ numbers?

Answer: Either one of the following events occurs:

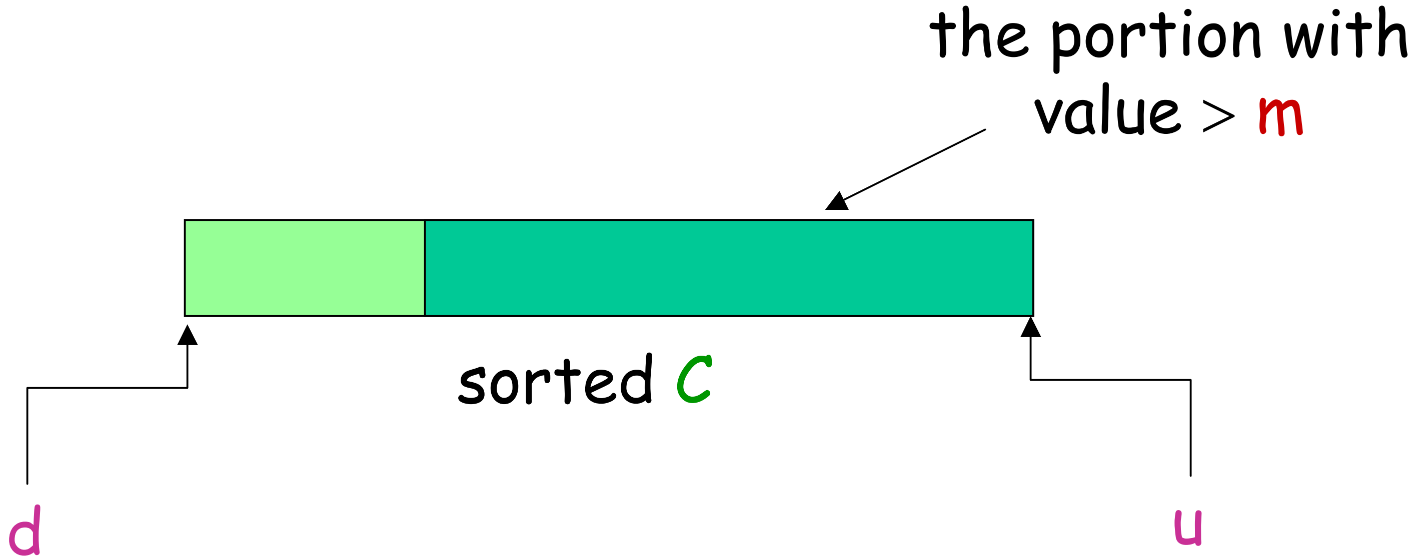
A : more than $2n^{3/4}$ numbers in C has value $>$ median m

B : more than $2n^{3/4}$ numbers in C has value $<$ median m

➔ $\Pr(E_3) \leq \Pr(A \cup B) \leq \Pr(A) + \Pr(B)$

Bounding Pr(*A*)

When *A* happens :

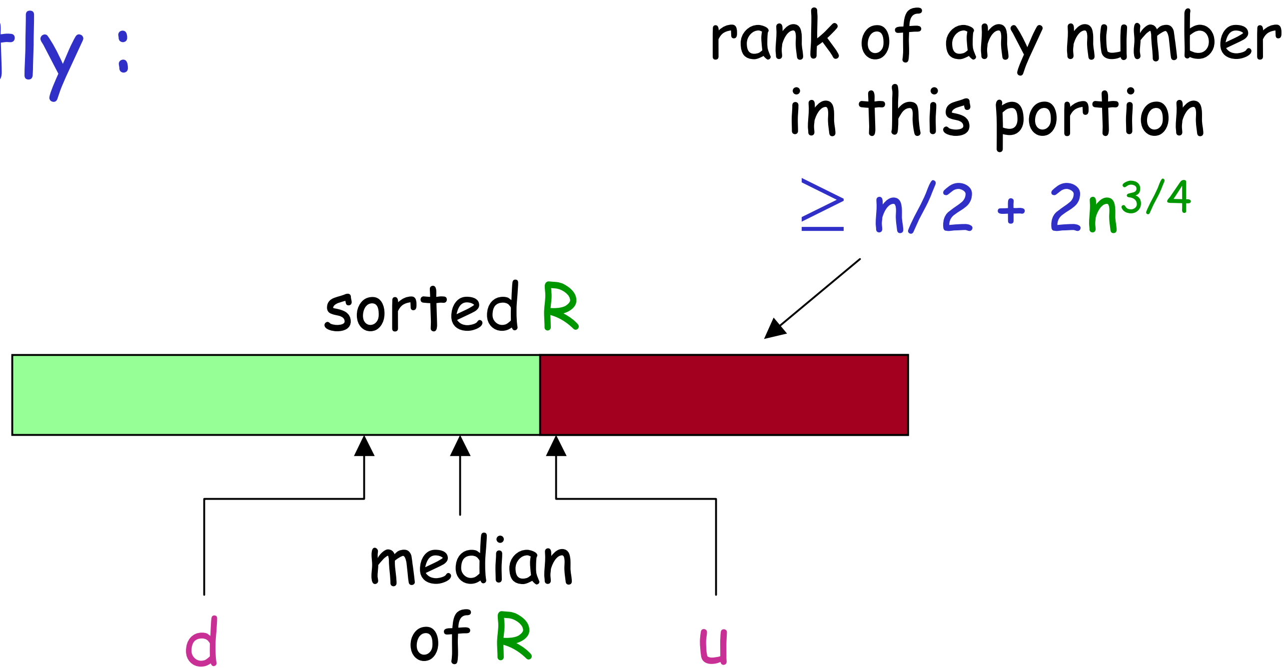


the rank of u (in S) = n/2 + size of 

$\geq n/2 + 2n^{3/4}$

Bounding Pr(A)

Consequently :



and we will soon show that this is unlikely

Bounding Pr(A)

Let Z = # chosen numbers in R whose rank is at least $n/2 + 2n^{3/4}$ = size of

So, $\Pr(A) \leq \Pr(Z \geq \lfloor n^{3/4}/2 - n^{1/2} \rfloor)$

Let Z_j be an indicator that :

$Z_j = 1$ if j^{th} number of R is in

$Z_j = 0$ otherwise

➔ $Z = Z_1 + Z_2 + \ldots + Z_{\lceil n^{3/4} \rceil}$

Bounding Pr(A)

Next, we want to find E[Z] and Var[Z], so that we can use Chebyshev inequality to bound Pr(A)

It is easy to check that

$$E[Z_1] = Pr(Z_1 = 1)$$
$$= 1/2 - 2/n^{1/4} + 1/n$$

$$\rightarrow E[Z] = \lceil n^{3/4} \rceil /2 - 2n^{1/2} + n^{1/4}$$
$$Var[Z] \leq \lceil n^{3/4} \rceil /4$$

Bounding Pr(A)

Thus,

$$
\begin{aligned}
\Pr(A) &\le \Pr(Z \ge \lfloor n^{3/4}/2 - n^{1/2} \rfloor) \\
&\le \Pr(Z \ge n^{3/4}/2 - n^{1/2} - 1) \\
&\le \Pr(Z \ge E[Z] + n^{1/2} - 1 - n^{1/4}) \\
&\le \Pr(\,|Z - E[Z]\,| \ge n^{1/2} - 1 - n^{1/4}) \\
&\le \mathrm{Var}[Z] \,/\, (n^{1/2} - 1 - n^{1/4})^2 \\
&= O(\,1 \,/\, n^{1/4}\,)
\end{aligned}
$$

... which is small for large n

Similarly, $\Pr(B) = O(\,1 \,/\, n^{1/4}\,)$

Bounding Pr(FAIL)

Thus,

$$\Pr(E_3) \le \Pr(A) + \Pr(B) = O(1 / n^{1/4})$$

Conclusion:

$$\Pr(\text{FAIL}) = \Pr(E_1 \cup E_2 \cup E_3)$$
$$\le \Pr(E_1) + \Pr(E_2) + \Pr(E_3)$$
$$= O(1 / n^{1/4})$$

➔ Algorithm succeeds with high probability