

Markov, Chebyshev, and Sampling-Based Median!

CS265/CME309, Class 4

1 Warm-Up

Suppose you flip a p -biased coin n times. What does Markov's inequality tell you about the probability that you see more than $2pn$ heads? What does Chebyshev's inequality tell you about that probability?

Announcements

- HW1 due Friday!
- HW2 out now! (You have until Friday 1/31)

- OH Schedule:
 - Mary: Mondays 3-5pm, Gates 168
 - Nevin: Tuesdays 11am-1pm, Huang Basement
 - Joey: Wednesdays 3-5pm, Huang Basement
 - Luna: Thursday 1:30pm-3:30pm, Huang Basement

Recap

- Markov's inequality: If $X \geq 0$, $\Pr[X \geq \alpha] \leq \frac{E[X]}{\alpha}$
- Chebyshev's inequality: $\Pr \left[|X - E[X]| \geq c\sqrt{\text{Var}(x)} \right] \leq \frac{1}{c^2}$

Questions?

Quiz, Videos, Warm-up?

1 Warm-Up

Suppose you flip a p -biased coin n times. What does Markov's inequality tell you about the probability that you see more than $2pn$ heads? What does Chebyshev's inequality tell you about that probability?

Warm Up

- Markov says: $1/2$

- Chebyshev says: $\frac{1-p}{np}$

Sampling-based Median

Finding the median of n things

- You may have seen an $O(n)$ time algorithm in CS161.
 - It was pretty complicated.
- Today: a simpler randomized algorithm!

Array S of n distinct numbers:

9	5	34	1	2	33	12	4	15	3	6	8	10	18	0
---	---	----	---	---	----	----	---	----	---	---	---	----	----	---

$n = 15$
here.

Array S of n distinct numbers:

9	5	34	1	2	33	12	4	15	3	6	8	10	18	0
---	---	----	---	---	----	----	---	----	---	---	---	----	----	---

$n = 15$
here.

Choose a set R of size $n^{3/4}$ by drawing that many things uniformly at random, independently.

5

12

15

5

10

3

33

Array S of n distinct numbers:

9	5	34	1	2	33	12	4	15	3	6	8	10	18	0
---	---	----	---	---	----	----	---	----	---	---	---	----	----	---

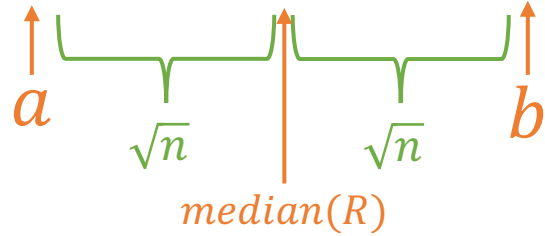
$n = 15$
here.

Choose a set R of size $n^{3/4}$ by drawing that many things uniformly at random, independently.

5	12	15	5	10	3	33
---	----	----	---	----	---	----

Sort R :

3	5	5	10	12	15	33
---	---	---	----	----	----	----



Array S of n distinct numbers:

9	5	34	1	2	33	12	4	15	3	6	8	10	18	0
---	---	----	---	---	----	----	---	----	---	---	---	----	----	---

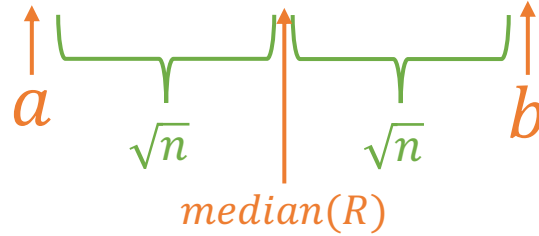
$n = 15$
here.

Choose a set R of size $n^{3/4}$ by drawing that many things uniformly at random, independently.

5	12	15	5	10	3	33
---	----	----	---	----	---	----

Sort R :

3	5	5	10	12	15	33
---	---	---	----	----	----	----



Find all the things in S between a and b (time $O(n)$), to form a list T :

9	5	12	15	6	8	10
---	---	----	----	---	---	----

Array S of n distinct numbers:

9	5	34	1	2	33	12	4	15	3	6	8	10	18	0
---	---	----	---	---	----	----	---	----	---	---	---	----	----	---

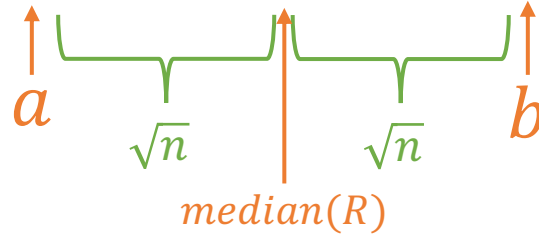
$n = 15$
here.

Choose a set R of size $n^{3/4}$ by drawing that many things uniformly at random, independently.

5	12	15	5	10	3	33
---	----	----	---	----	---	----

Sort R :

3	5	5	10	12	15	33
---	---	---	----	----	----	----



Find all the things in S between a and b (time $O(n)$), to form a list T :

9	5	12	15	6	8	10
---	---	----	----	---	---	----

If $|T| < 4n^{3/4}$, sort T :
(otherwise output FAIL)

5	6	8	9	10	12	15
---	---	---	---	----	----	----

Array S of n distinct numbers:

9	5	34	1	2	33	12	4	15	3	6	8	10	18	0
---	---	----	---	---	----	----	---	----	---	---	---	----	----	---

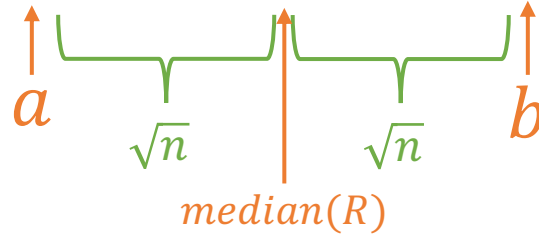
$n = 15$
here.

Choose a set R of size $n^{3/4}$ by drawing that many things uniformly at random, independently.

5	12	15	5	10	3	33
---	----	----	---	----	---	----

Sort R :

3	5	5	10	12	15	33
---	---	---	----	----	----	----



- We can see in time $O(n)$ that there are 5 things in S less than a , and 3 things in S larger than b .

Find all the things in S between a and b (time $O(n)$), to form a list T :

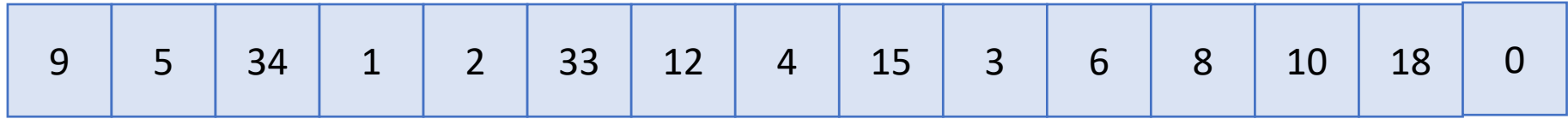
9	5	12	15	6	8	10
---	---	----	----	---	---	----

If $|T| < 4n^{3/4}$, sort T :
(otherwise output FAIL)

5	6	8	9	10	12	15
---	---	---	---	----	----	----

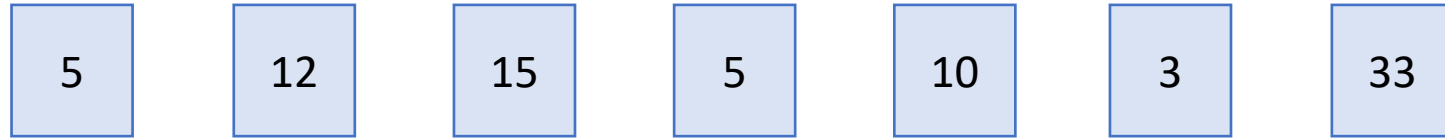
					5	6	8	9	10	12	15			
--	--	--	--	--	---	---	---	---	----	----	----	--	--	--

Array S of n distinct numbers:

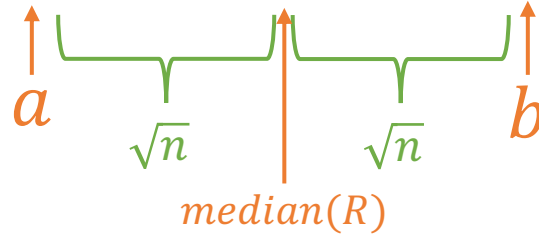
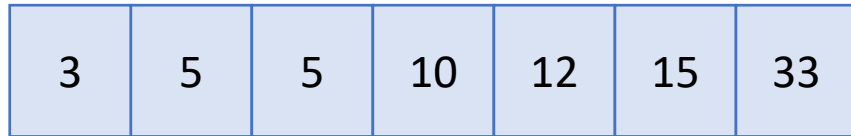


$n = 15$ here.

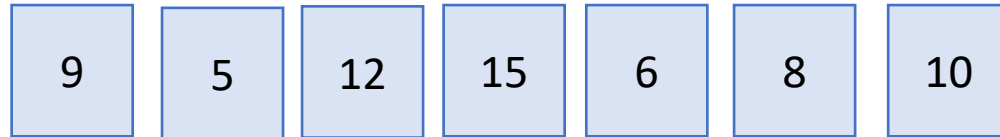
Choose a set R of size $n^{3/4}$ by drawing that many things uniformly at random, independently.



Sort R :



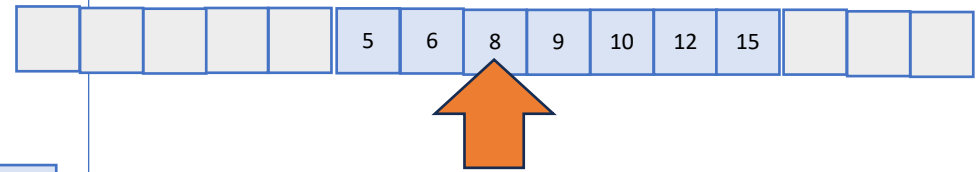
Find all the things in S between a and b (time $O(n)$), to form a list T :



If $|T| < 4n^{3/4}$, sort T :
(otherwise output FAIL)



- We can see in time $O(n)$ that there are 5 things in S less than a , and 3 things in S larger than b .



- The median is the 8'th smallest thing in S , which is the $8 - 5 = 3$ 'rd smallest thing in T .

Array S of n distinct numbers:

9	5	34	1	2	33	12	4	15	3	6	8	10	18	0
---	---	----	---	---	----	----	---	----	---	---	---	----	----	---

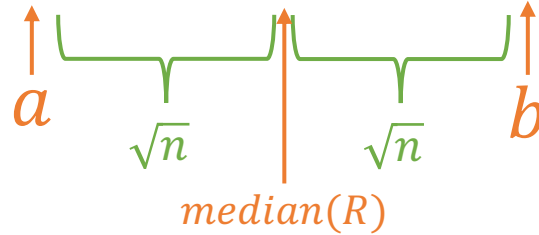
$n = 15$
here.

Choose a set R of size $n^{3/4}$ by drawing that many things uniformly at random, independently.

5	12	15	5	10	3	33
---	----	----	---	----	---	----

Sort R :

3	5	5	10	12	15	33
---	---	---	----	----	----	----



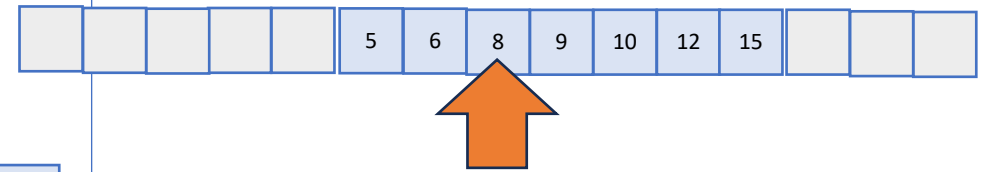
Find all the things in S between a and b (time $O(n)$), to form a list T :

9	5	12	15	6	8	10
---	---	----	----	---	---	----

If $|T| < 4n^{3/4}$, sort T :
(otherwise output FAIL)

5	6	8	9	10	12	15
---	---	---	---	----	----	----

- We can see in time $O(n)$ that there are 5 things in S less than a , and 3 things in S larger than b .



- The median is the 8'th smallest thing in S , which is the $8 - 5 = 3$ 'rd smallest thing in T .

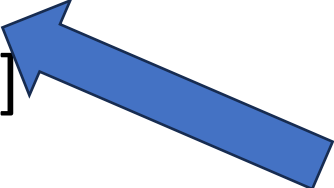
- Return it!

8

If this calculation shows that the median is not in T , output FAIL.

Group work!

1. Make sure you understand the alg.
2. Suppose that whp, $\text{median}(S) \in T$, and whp, $|T| < 4t$. Explain why alg is correct whp (and if it's not correct it says FAIL).
3. What is the running time of the algorithm?
4. Show that $\text{median}(S) \in T$ whp. [See agenda for outline]
5. Show $|T| < 4t$ whp. [See agenda for outline]
6. Put it all together
7. (Optional) Play around with the argument



The plan is to regroup partway through class to go over 1-4. (But keep working if you finish 4 earlier).

Solutions to group work

2. Suppose that:

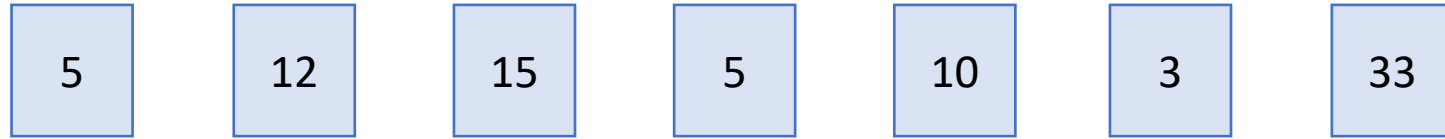
- With probability at least 0.9, the median of S is in T .
 - With probability at least 0.9, $|T| < 4t$.
- Show that the algorithm returns the correct answer with probability at least 0.8, and otherwise returns FAIL.

Array S of n distinct numbers:

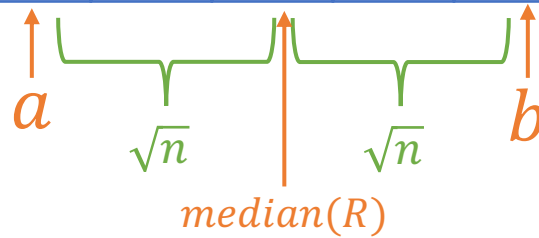
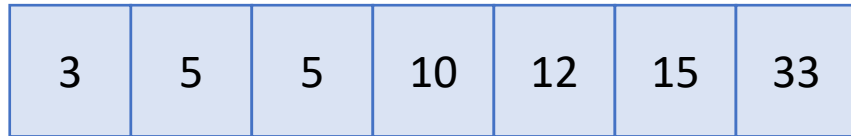


$n = 15$ here.

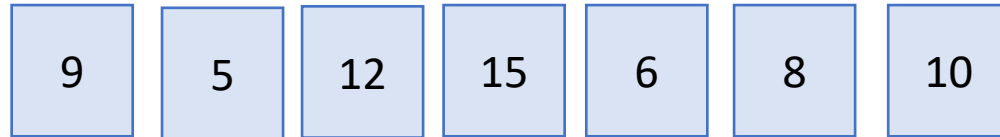
Choose a set R of size $n^{3/4}$ by drawing that many things uniformly at random, independently.



Sort R :



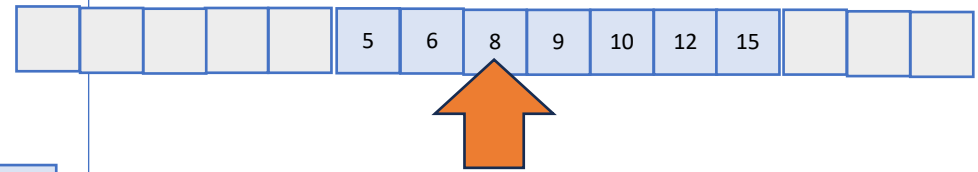
Find all the things in S between a and b (time $O(n)$), to form a list T :



If $|T| < 4n^{3/4}$, sort T :
(otherwise output FAIL)



- We can see in time $O(n)$ that there are 5 things in S less than a , and 3 things in S larger than b .



- The median is the 8'th smallest thing in S , which is the $8 - 5 = 3$ 'rd smallest thing in T .

- Return it!



If this calculation shows that the median is not in T , output FAIL.

Solutions to group work

2. Suppose that:

- With probability at least 0.9, the median of S is in T .
 - With probability at least 0.9, $|T| < 4t$.
- Show that the algorithm returns the correct answer with probability at least 0.8, and otherwise returns FAIL.
- If both events happen, then the algorithm never returns FAIL.
 - $\Pr[\text{either happens}] \leq \Pr[\text{first happens}] + \Pr[\text{second happens}] \leq 0.1 + 0.1 = 0.2$
 - If it doesn't return FAIL, then it returns the right answer by construction.

Solutions to group work

3. What's the running time?

What's the running time, asymptotically?

$O(n)$

0%

$O(n \log n)$

0%

The expected running time is $O(n)$, but the worst-case running time may be worse.

0%

$o(n)$

0%

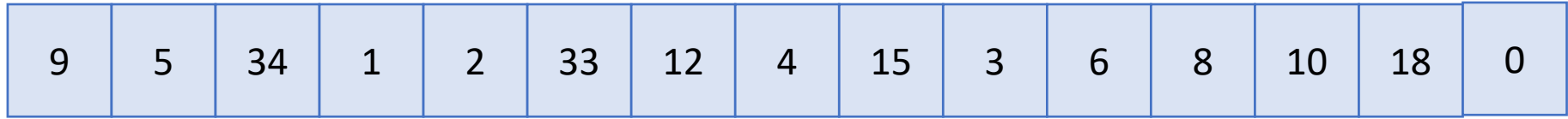
None of the above

0%

What is the leading constant in the running time?

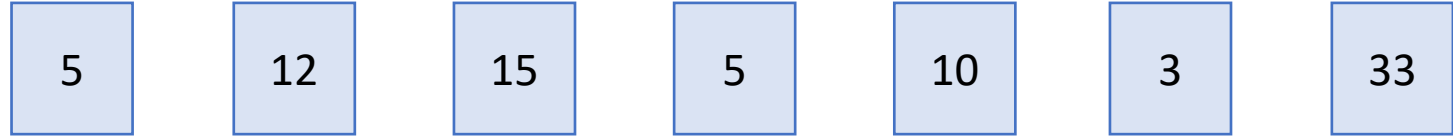


Array S of n distinct numbers:



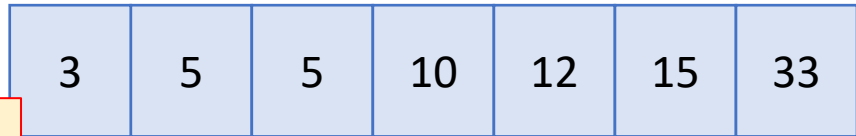
$n = 15$ here.

Choose a set R of size $n^{3/4}$ by drawing that many things uniformly at random, independently.

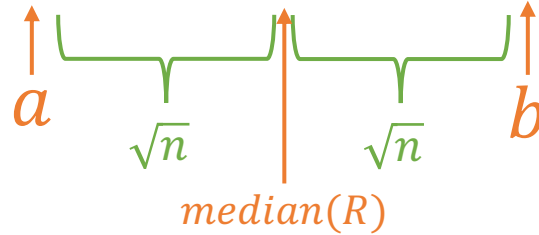


$O(n^{3/4}) = o(n)$ operations.

Sort R :



$O\left(n^{\frac{3}{4}} \log\left(n^{\frac{3}{4}}\right)\right) = o(n)$ operations.

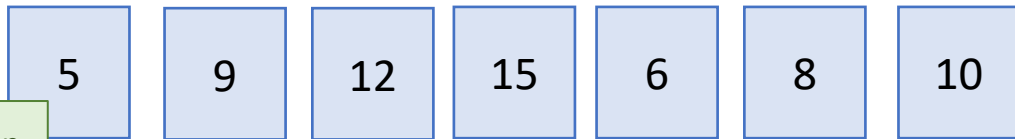


- We can see in time $O(n)$ that there are 5 things in S less than a , and 3 things in S larger than b .

$O(n)$

Do this in the same scan as before; no extra time.

Find all the things in S between a and b , to form a list T :



$O(n)$

$\frac{3}{2}n + o(n)$ in expectation
Or $2n + o(n)$ worst-case

- The median is the 8'th smallest thing in S , which is the $8 - 5 = 3$ 'rd smallest thing in T .

$O(1)$

If $|T| < 4n^{3/4}$, sort T :
(otherwise output FAIL)



$O\left(n^{\frac{3}{4}} \log\left(n^{\frac{3}{4}}\right)\right) = o(n)$ operations.

- Return



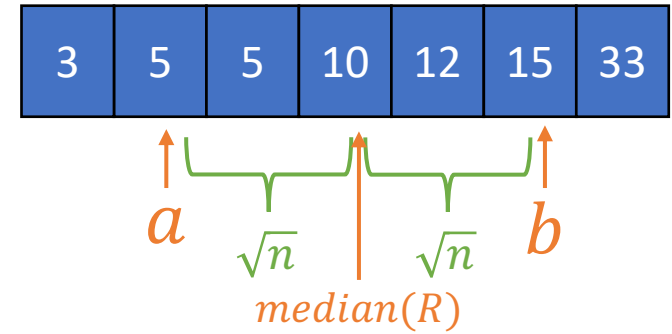
If this calculation shows that the median is not in T , output FAIL.

Solutions to group work

- Question 4: want to show that $\text{median}(S) \in T$ w.h.p.

Solutions to group work

Sorted version of R:

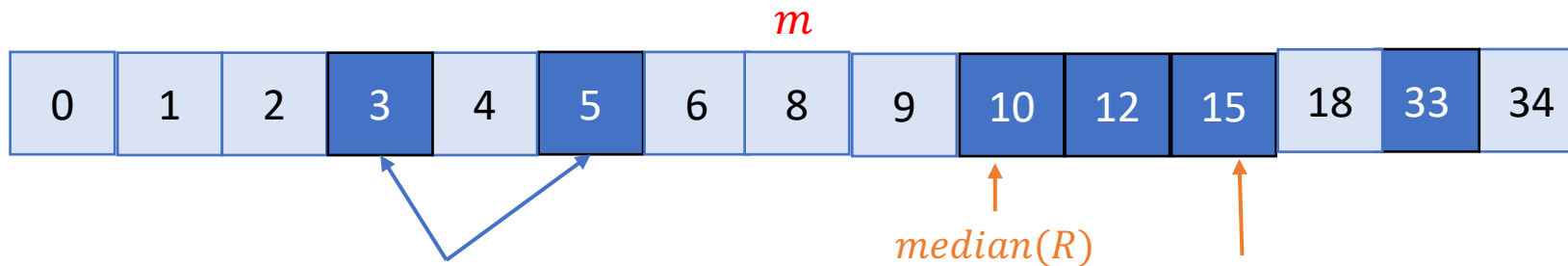


4a. Consider two events:

$$|\{r_i \in R: r_i < m\}| < \frac{t}{2} + \sqrt{n}$$

$$|\{r_i \in R: r_i > m\}| < \frac{t}{2} + \sqrt{n}$$

Sorted version of S:



$$|\{r_i \in R: r_i < m\}| < \frac{t}{2} + \sqrt{n}$$

$$\Rightarrow b \geq m$$

b
 $\frac{t}{2} + \sqrt{n}$ 'th smallest
 thing in R .

Solutions to group work

4a. Consider two events:

$$|\{r_i \in R: r_i < m\}| < \frac{t}{2} + \sqrt{n}$$

$$\Rightarrow b \geq m$$

$$|\{r_i \in R: r_i > m\}| < \frac{t}{2} + \sqrt{n}$$

$$\Rightarrow a \leq m$$

- Then $a \leq m \leq b$, aka $m \in T$

4(b) What's the (best) probability that the first event doesn't happen?

$O(1/n)$

0%

$O(1/\sqrt{n})$

0%

$O(1/n^{1/4})$

0%

At most 0.01

0%

Not sure (didn't solve it yet)

0%

None of the above

0%

Solutions to group work

- 4b. What's the probability that $|\{r_i \in R: r_i < m\}| \geq \frac{t}{2} + \sqrt{n}$?

Let $X = |\{r_i \in R: r_i < m\}|$

- Then $X = \sum_i X_i$ where $X_i = 1$ iff $r_i < m$ and 0 otherwise, for $i = 1, \dots, t$
- $\mathbf{E}[X_i] = \Pr[r_i < m] \leq \frac{1}{2}$,
- $\text{Var}[X_i] \leq \frac{1}{4}$

$$\Pr \left[\sum_i X_i \geq \frac{t}{2} + \sqrt{n} \right] \leq \Pr \left[\sum_i (X_i - \mathbf{E}X_i) \geq \sqrt{n} \right] \leq \frac{t/4}{n} = \frac{1}{4n^{1/4}} = o(1)$$

Solutions to group work

- 4c. Bound the probability that $m \in T$.

$$|\{r_i \in R: r_i < m\}| < \frac{t}{2} + \sqrt{n}$$

$$\Rightarrow b \geq m$$

$$|\{r_i \in R: r_i > m\}| < \frac{t}{2} + \sqrt{n}$$

$$\Rightarrow a \leq m$$

- Then $a \leq m \leq b$, aka $m \in T$

Both have probability at least $1 - O(n^{-1/4})$

$$\Pr[m \in T] \geq 1 - O(n^{-1/4})$$

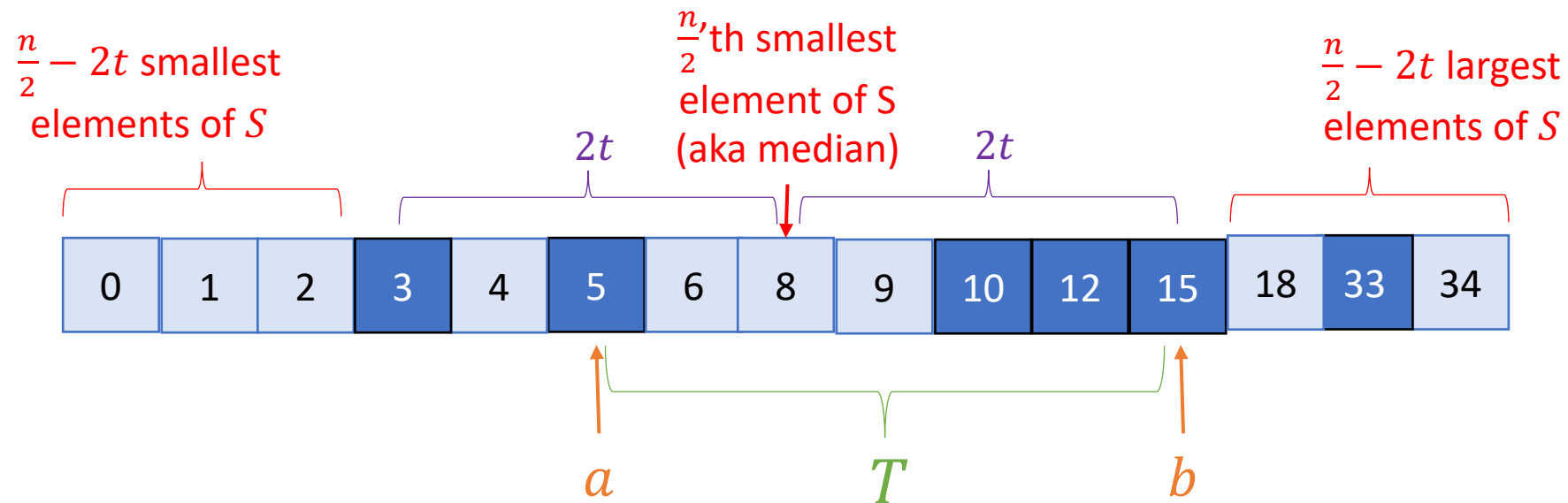
Solutions to group work

- Question 5: Show that $|T| < 4t$ w.h.p.

Solutions to group work

So what's the probability of this?

- 5(a). Say that a is not one of the $\frac{n}{2} - 2t$ smallest elements of S
- Say that b is not one of the $\frac{n}{2} - 2t$ largest elements of S



- Then $|T| < 4t$

5(b) What's the (best bound on) the probability that a is one of the smallest $(n/2-2t)$ items in S ?

$O(1/n)$

0%

$O(1/\sqrt{n})$

0%

$O(1/n^{1/4})$

0%

At most 0.01

0%

Not sure (didn't solve it yet)

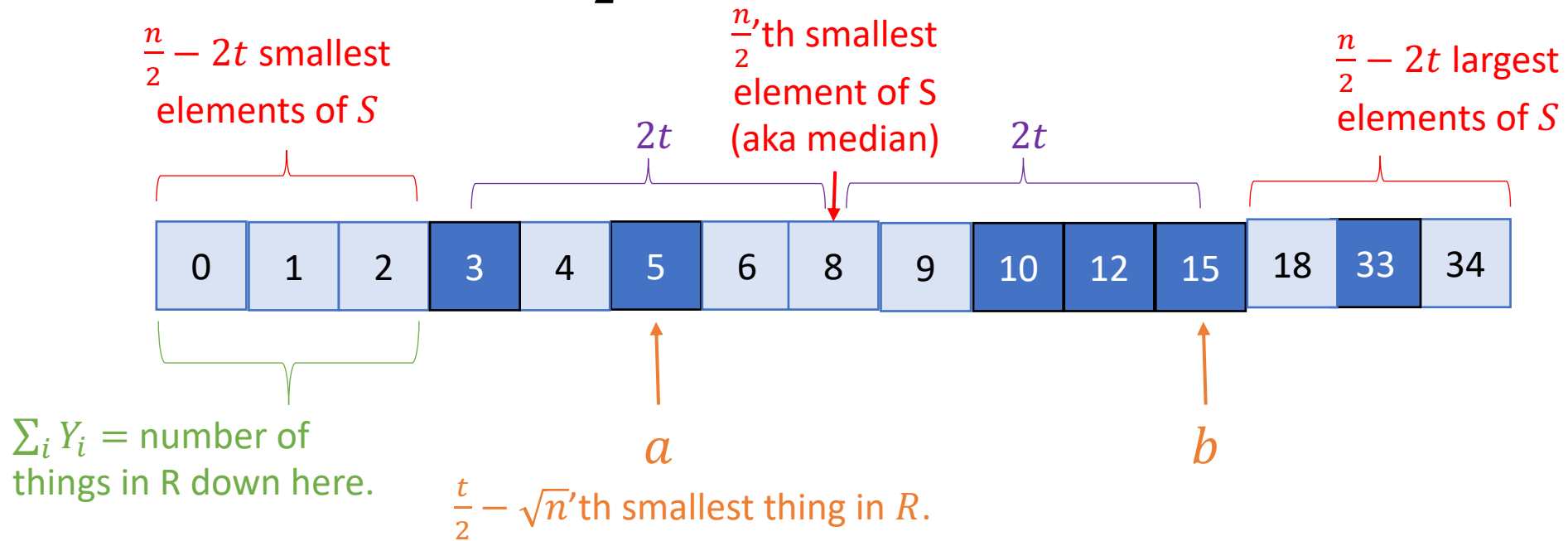
0%

None of the above

0%

Solutions to group work

- 5(b) What's the prob. a is not one of the $\frac{n}{2} - 2t$ smallest elements of S ?
- Let $Y_i = 1$ iff r_i is in the $\frac{n}{2} - 2t$ smallest elements of S , 0 else



- Claim:
- $\sum_i Y_i \geq \frac{t}{2} - \sqrt{n} \iff a$ is among the $\frac{n}{2} - 2t$ smallest elements of S

- $\sum_i Y_i \geq \frac{t}{2} - \sqrt{n} \Leftrightarrow a$ is among the $\frac{n}{2} - 2t$ smallest elements of S

Solutions to group work

- 5(b) What's the prob. a is not one of the $\frac{n}{2} - 2t$ smallest elements of S ?

- Let $Y_i = 1$ iff r_i is in the $\frac{n}{2} - 2t$ smallest elements of S , 0 else

- $\mathbf{E}Y_i = \frac{1}{2} - \frac{2t}{n} = \frac{1}{2} - \frac{2}{n^{1/4}} \quad \text{Var}[\sum_i Y_i] = \sum_i \text{Var}[Y_i] \leq \frac{t}{4}.$

- $\Pr \left[\sum_i Y_i \geq \frac{t}{2} - \sqrt{n} \right] \leq \Pr \left[\sum_i (Y_i - \mathbf{E}Y_i) \geq \frac{2t}{n^{1/4}} - \sqrt{n} \right]$

- $= \Pr \left[\sum_i (Y_i - \mathbf{E}Y_i) \geq \sqrt{n} \right] \quad t = n^{3/4}$

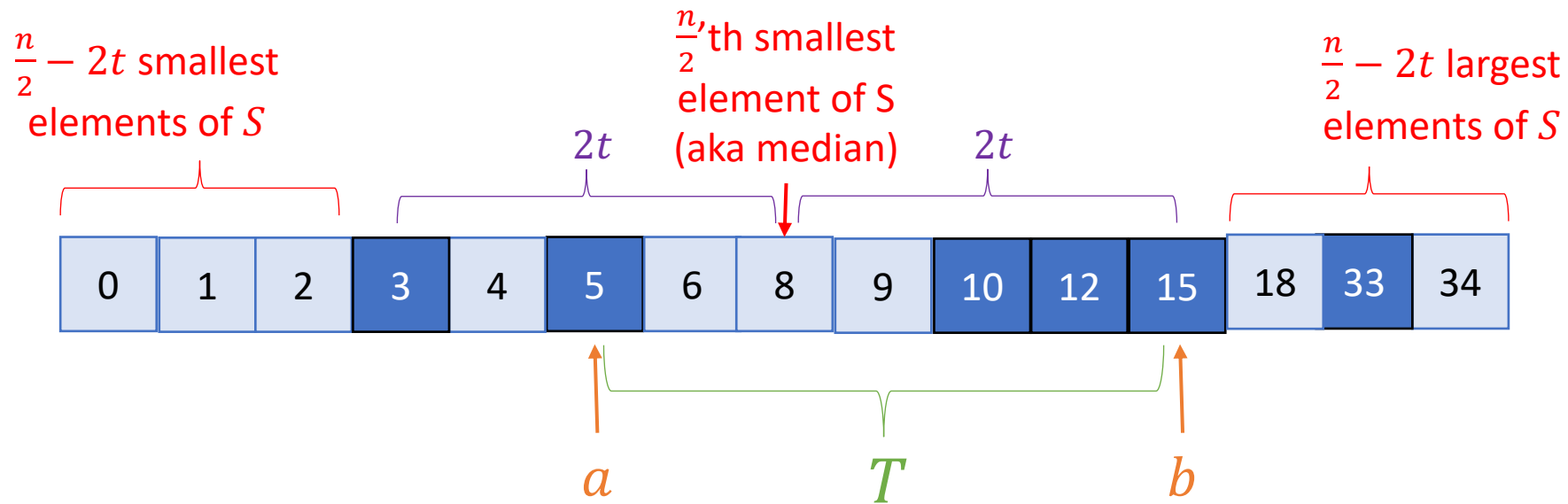
- $\leq \frac{\text{Var}[\sum_i Y_i]}{n}$

- $\leq \frac{t}{4n} = \frac{1}{4n^{1/4}} = o(1)$

Solutions to group work

Both have probability at least $1 - O(n^{-1/4})$

- 5(c). What's the probability that $|T| < 4t$?
 - Say that a is not one of the $\frac{n}{2} - 2t$ smallest elements of S
 - Say that b is not one of the $\frac{n}{2} - 2t$ largest elements of S



- Then $|T| < 4t$

$$\Rightarrow \Pr[|T| < 4t] \geq 1 - O(n^{-1/4})$$

6. All together:

- Question 2: To show that this algorithm works whp, it's enough to show that :
 - whp, $\text{median}(S) \in T$
 - whp, $|T| < 4t$
- Question 4: whp, $\text{median}(S) \in T$
- Question 5: whp, $|T| < 4t$
- (And Question 3: it runs in time $O(n)$).

