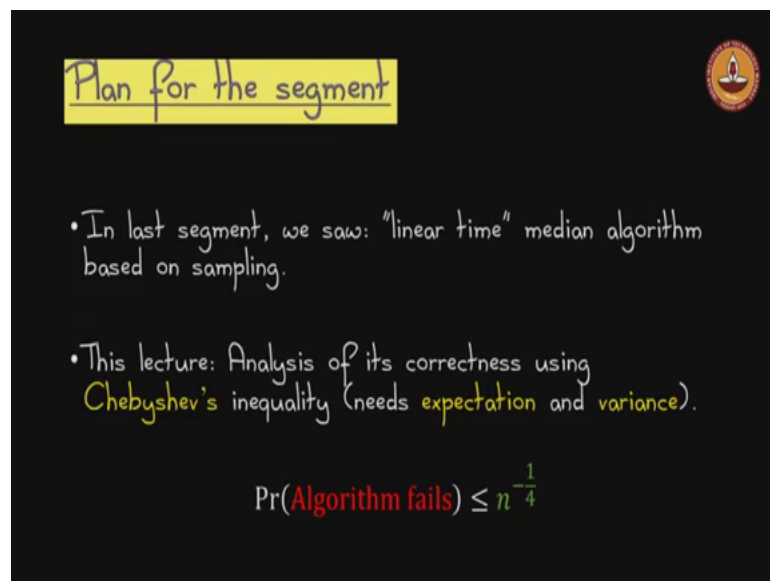


Probability & Computing
Prof. John Augustine
Department of Computer Science and Engineering
Indian Institute of Technology Madras

Module – 03
Tails Bounds I
Lecture - 18
Segment 4: Median via Sampling – Analysis

We are now in segment four of module three we have already seen a median algorithm based on sampling. We just saw how the algorithm works, but we and we it is also a reasonably clear, that it is a linear time algorithm; what is not clear is why it is correct and how to state how to do how to analyse it.

(Refer Slide Time: 00:42)



Plan for the segment

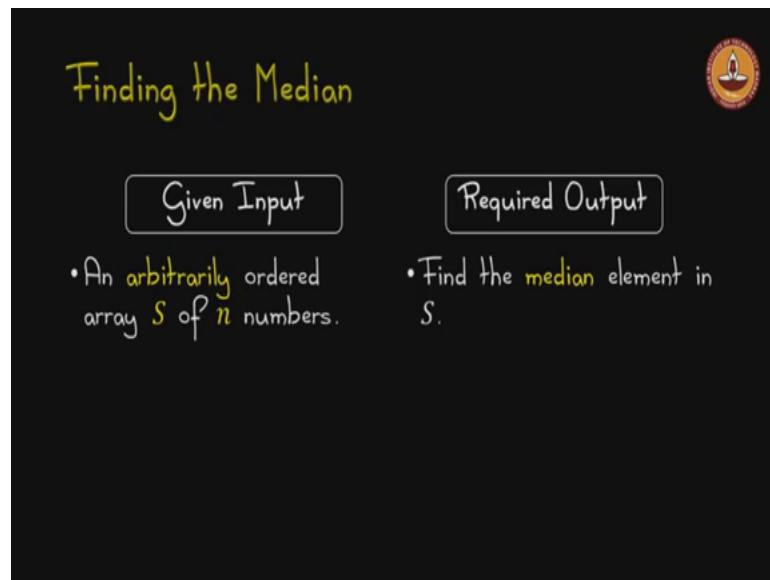
- In last segment, we saw: "linear time" median algorithm based on sampling.
- This lecture: Analysis of its correctness using Chebyshev's inequality (needs expectation and variance).

$$\Pr(\text{Algorithm fails}) \leq \frac{1}{n^4}$$

And that is going to be the plan for today; we are going to focus on the analysis of correctness of this median algorithm. And we will be using Chebyshev's inequality in particular we are going to prove the following claim, that the algorithm fails with probability at most $\frac{1}{4n}$.

So, we already alluded to this claim. We are going to see how this is proved. And if you recall Chebyshev's inequality requires you to know the expectation and the variance of the random variable. So, with these two pieces of information you will be able to bound random variables that show up in the analysis ok.

(Refer Slide Time: 01:25)



Finding the Median

Given Input

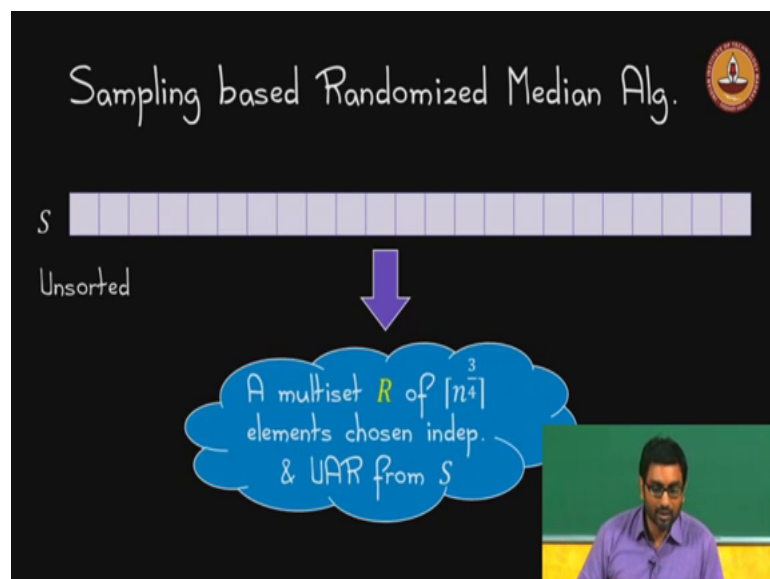
- An arbitrarily ordered array S of n numbers.

Required Output

- Find the median element in S .

So, I will first recall the median problem the algorithm and some of the details there ok. So, of course, you have a set S arbitrarily ordered and you need to find the median that is the median problem ah.

(Refer Slide Time: 01:42)



Sampling based Randomized Median Alg.

S [array of 15 cells]

Unsorted

↓

A multiset R of $\lceil n^3 \rceil$ elements chosen indep. & UAR from S

[Instructor's video feed]

And let us recall the algorithm and keep in mind that there are three major objects in this algorithm that interplay ok. There is the set S which is the; or the array S if you will, it has the elements from which you need to find the median that is the input set ok.

And; then the first thing that happens is you sample multiset R that is the second object ok; you have a multiset R and it is of size n to the three-fourths these are elements chosen uniformly and independently at random from the set S ok.

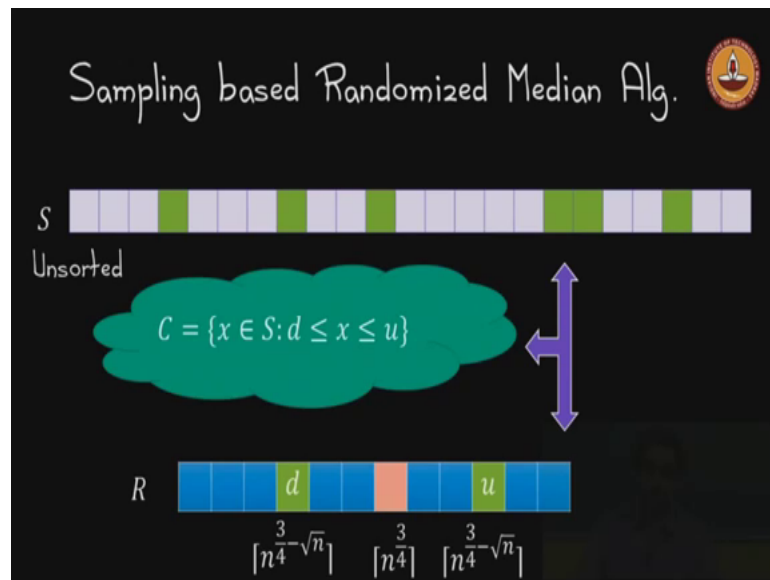
(Refer Slide Time: 02:24)

The slide is titled "Sampling based Randomized Median Alg." and features a logo in the top right corner. It contains the following elements:

- A blue cloud-shaped box containing the text: "A multiset R of $\lceil n^3 \rceil$ elements chosen indep. & UAR from S ".
- A purple arrow pointing downwards with the text "Sort into array" next to it.
- A horizontal array of colored blocks representing a sorted array. From left to right: a green block labeled d , a red block, and a green block labeled u . Below the array, three intervals are marked: $[\frac{3}{n^4} - \sqrt{n}]$ under the d block, $[\frac{3}{n^4}]$ under the red block, and $[\frac{3}{n^4} + \sqrt{n}]$ under the u block.
- A purple rounded rectangle containing the text "Intuition:" followed by two bullet points:
 - d is less than but close to the median
 - u is greater than but close to the median
- A small video inset in the bottom right corner showing a man speaking.

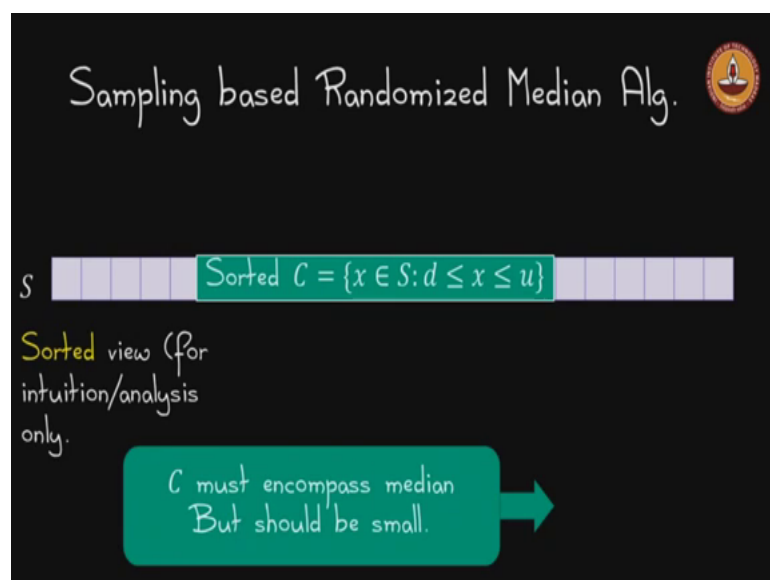
This is the sample and the third object; well before we go to the third object what we do is we take this second object R we sort it and we pick out two elements d and u from this array. And our intuition is that d should be to the should be less than the median, but close to the median it should not be too far from the median. Similarly u should be greater than the median, but again close to the median. So, these are this is the intuition based on which the algorithm works of course, in our analysis today we will be proving this intuition formally.

(Refer Slide Time: 03:02)



And now comes a third object. So, based on these two elements d and u ; you pick out you go back to set S and you pick out all those elements that are in S that lie within the range d and u that S the third object C . So, these three objects interplay quite a bit and how these three objects interplay is very important? it is it is crucial to understanding how to analyse this algorithm ok.

(Refer Slide Time: 03:28)

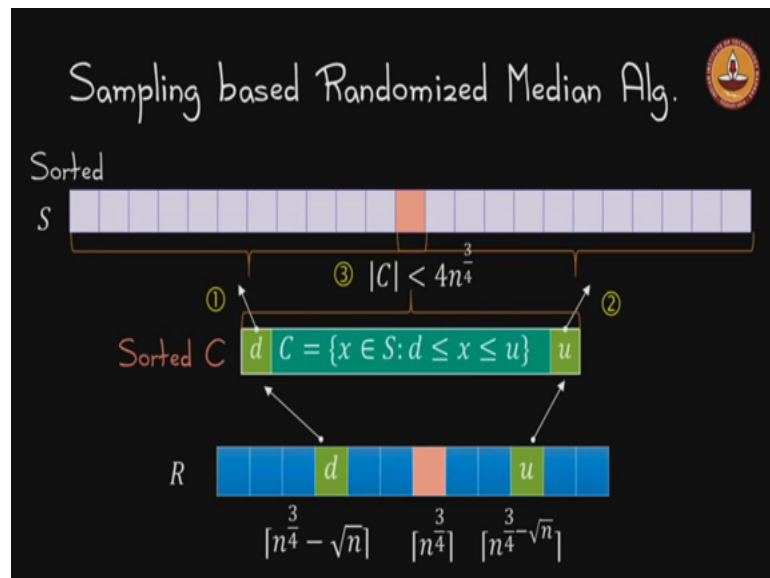


So, what is the intuition here; this third object remember; d is to the left of the median is as our hope use to the right of the median. So, when you take the; now when we are

trying to understand the analysis sometimes we look at S , but we look at it in the sorted view ok. The sorted view is not available to the algorithm. Remember the analysis is one thing; the algorithm is another thing ok. The algorithm works on an unsorted list, but the analysis for understanding how the algorithm works you can view the sorted array and play with it and that is what we are doing.

So, in this sorted array; what you do is you can superimpose the set the array C and remember array C is something that we actually sought in the algorithm as well. And in this array C , what we do is well the d the smallest element in this array we hope is less than the median in S and u we hope is to the right of the median. So, this sorted set spans the median. So, it must encompass the median, but it should not be too large it should be something like into the three-fourth. So, which means that because it spans the median you can spot the median of S within the array C , this is the intuition that I hope you already have. Any questions on how this algorithm works; any issues at all? Because that is understanding this algorithm is absolutely imperative to understanding the analysis.

(Refer Slide Time: 05:07)

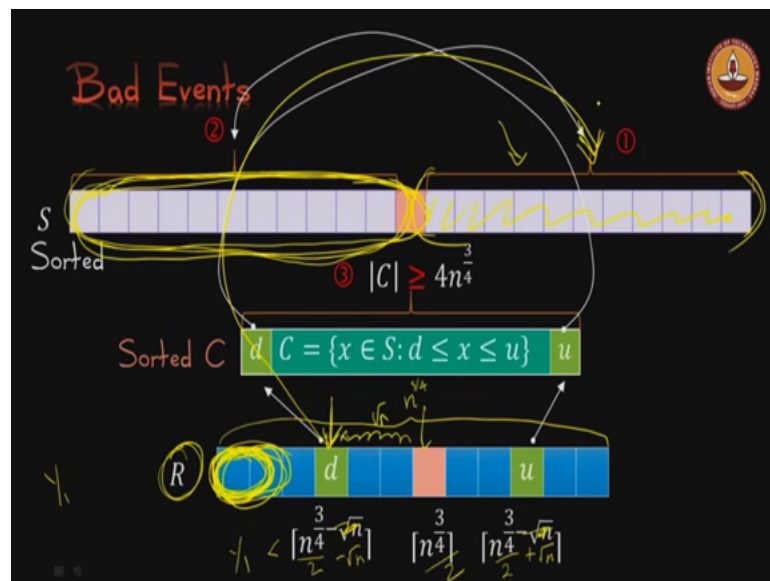


So here is the three objects ok. Now what are the recall that we had some we hope some good things will happen. What was the first good thing that we hoped would happen? The d the element d is to the left of the median ok. And let us ask ourselves ok; how does that happen? If you think about it where is the right event should occur right way even

when you sample the elements R ok. What is this element d ? You go back to the fact that you first sample the set of elements this multiset of elements R . You sorted it found the median, you walked left square root of n steps and then you found the median and from here that ended up being the leftmost element in C and that has to fall in this left region of S ok.

Similarly, the other good thing that must happen is you are u there this is this is a symmetric good event; u should be from the right part of the array S ok. And the third good thing you want is that; the set C should not be too large it should be no more than some four times into the three-fourths and here is an important principle. So, there are these good events that you want and you want to make sure that these good events happen with high probability ok, but often in that that is hard to prove. So, what do we do? We focus on the corresponding bad events, because those are is often easier to bound ok.

(Refer Slide Time: 07:09)



So, let us take these good events and convert them into the appropriate bad events ok. So, what is the bad event? So, the good event was that d falls from the left side; the bad event is that the d falls on the right side that is one bad event. I am going to call that bad event number one ok; I am going to denote that by a one n close to the circle and of course, the mirror image is also easy to see the other bad event is that the u falls on the

left side; u should rightfully fall on the right side to the right of the median, but falling on the left side is the bad event ok.

And you want to C to be less than four times n to the power three-fourths the bad event is just the opposite. We will understand the third bad event more carefully and more slowly later on ok, but hopefully the first two events the back the corresponding bad events are easy to visualize and what is going on ok.

(Refer Slide Time: 08:05)

The slide contains the following text:

Formal Definitions of Bad Events.

① : $Y_1 \triangleq |\{r \in R: r \leq m\}| < \frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}$

② : $Y_2 \triangleq |\{r \in R: r \geq m\}| < \frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}$

③ : $|C| > 4n^{\frac{3}{4}}$

Union Bound:

$$\Pr(\textcircled{1} \cup \textcircled{2} \cup \textcircled{3}) \leq \Pr(\textcircled{1}) + \Pr(\textcircled{2}) + \Pr(\textcircled{3})$$

The slide also features a small video inset of a man in a purple shirt speaking in front of a green chalkboard.

So, let us formally write down the bad events; because, now what we have here is a pictorial representation ah. This d over here this d falling on the right side is the first bad event. Let us try to write it out formally and let us make sure we understand how this works and so for this we first define a random variable Y 1 ok. Y 1 is simply all those elements in R remember R capital R is the set of elements that we sampled ok. So, original elements that we sampled; out of which we are picking out all the elements that are less than or equal to the median; m here represents the median and we are asking how large is that set ok. And this is the; this is this is the random variable Y one; the cardinality of the set of elements in r that r less than or equal to the median and the bad situation is when this Y 1 is less than this quantity ok. So, this will need this will require us to understand exactly why that is the case ok.

So, Y 1 let us go back to this picture. Y 1 is all those elements that are less than the median. So, Y 1 is you are counting the elements in this R ok; please pay attention to this

part. You are counting the elements in R ok, but those elements that are falling to the left of the median, alright. So, let us let us see what; so, now, why is that a bad situation? So, and in particular we are we are asking whether the particular bad event is when that quantity is less than half n to the three-fourth minus square root of n ok.

So, let us see why that is the case can somebody what I would suggest that you do is that you stare at this picture a little bit and convince yourself that that this bad event is actually captured by this by this formal statement ok.

So, what happens if the number of elements sampled from this region to the left is less than n to the three-fourths ah; so this by the way has ok. So, I think there is a small typo one more typo here this has to be this median element has to be n to the three-fourths divided by 2 ok. So, this median element here let me correct that over here small. So, this whole thing is of length n to the 3 by 4. So, this is; this element is actually not n to the 3 by 4 it is by 2 ok.

So, in apologies for the typo; so here also we have that and it is minus square root of n ; this is divided by 2 plus square root of n ok. Now; what is Y_1 . Y_1 is all those elements from R that are falling in this range on the left side ok. If this Y_1 is too few; where will d fall. In other words; if Y_1 those elements that are from the left of the median are just a small set of elements over here where will d fall?

Student: (Refer Time: 12:41).

Ha so, that is this is the intuition here. When this Y_1 is too small when it is particularly lesser than this quantity it is just going to occupy a small part on the left of R , which means that this d will fall in this region ok.

Student: So why that specific quantity.

The specific quantity Y_1 oh;

Student: No, no. How does it come less than n to the power 3 by 4 by 2 minus m ? Why Y_1 less than m ?

Ok, so, we are trying to capture a bad event right. The bad event is that d falls in this region ok.

Student: Yeah.

And now we perform ally capture that bad event and we want to capture it in such a way that we can apply Chebeshev's inequality ok. We want to define this random variable Y_1 which can for which we can find the expectation in variance right and be able to ensure that we can say that Y_1 will have will be small only with small probability that is the bad event when Y_1 is too small the d falls on the bad region.

Student: but the d will still be left of the median.

No when the Y_1 is small what happens think about it. What is Y_1 ? Y_1 is all is is the count of the elements in R that are sampled from the left side ok. If that is a very small part of this set R d is positioned over here if the if the elements corresponding to Y_1 are coming from this small set even inside R ; that will naturally push d to the right of the median. So, what is the precise definition here you define Y_1 and you want that to be less than half n to the three-fourth minus square root of n . What does that mean? When it is; when Y_1 is less than this quantity all the elements that were sampled left of median are only occupying the portion to the left of d ok.

They are only occupying portion to the left of d which means the d th element d th element is clearly based on a position within R . So, you are finding the middle element and you are walking back square root of n steps ok. So, d th position d position is fixed, but all the elements that correspond to Y_1 are to the left which means it if you think about it d will be pushed into the bad region this is important make sure you understand this because this is how the analysis a lot of setting up the variables to work out the analysis is captured in this step ok. Shall we proceed?.

If you want I can give a minute to stare for you to stare at this picture. That is a very good point thanks for pointing that out. So, Bharath's recommendation is that just think of the extreme situation where Y_1 is simply all is equal to 0; basically means that no element was sampled at all from this region. So, even the smallest element is over here which means that d is also going to be in this part. So, that is an extreme situation.


It is not like the procedure magically knows what the median is and keeps the d at that point right. The algorithm is just running without knowledge of what the median is right;

there is no magic about this. So, you it is a proper is just you are just sampling elements. So, you don't know where d lies really I mean where the median lies all right ok.

so let us move on. So, this is the formal way to represent the bad event ok. If the first event is captured this way the second event is going to be a mirror image ok. Here it is going to be another variable Y_2 ; where you are looking for all the elements greater than or equal to median in the sample set R and you are asking how will that will you know the bad event would be that that cardinality of that set that is Y_2 is fewer than n to the three-fourth by 2 minus square root of n it is too small ok. This is just a mirror image so not too tricky. And the third one I am going to defer the details till later, but will maintain we call it 3. So, we have listed the three bad events and we have given them names 1, 2 and 3 ok. 1 and 2 are essentially the same. So, we are going only going to bound 1 and the same bound will hold for 2 as well ok.

What we are going to do is apply a union bound? This is a very very important thing simple, but extremely useful; if you have a few bad things that can go wrong. In this case there are three bad things that can happen; and if you can ensure that each one of those bad things happens with very small probability then the probability that any one of those bad things happens is just at most the sum of the individual probabilities this is just coming from inclusion of inclusion exclusion principle and throwing away weight terms that we do not.

(Refer Slide Time: 18:12)



Pr(①)

①: $Y_1 \triangleq |\{r \in R: r \leq m\}| < \frac{1}{2}n^{\frac{3}{2}} - \sqrt{n}$


Let

$$X_i = \begin{cases} 1, & i^{\text{th}} \text{ sample} \leq \text{median} \\ 0, & \text{otherwise} \end{cases}$$

Clearly, $E[X_i] \approx \frac{1}{2}$, and

$$Y_1 = \sum_i X_i.$$

Therefore $E[Y_1] = \sum_i E[X_i] \approx n^{\frac{3}{2}} \left(\frac{1}{2}\right)$



So let us look at this first bad event ok; and recall that this is how we formally defined it. So, Y_1 is the cardinality of this set and the bad case is when it is too small ok. So, how do we bound this? Remember we even we carefully define this random variable. So, we can apply Chebyshev's which means that we need to be able to find the expectation and the variance ok. So, we want to break Y_1 into smaller pieces. What is Y_1 ? Is basically the count of the set of elements that are less than or equal to the median ok.

So, go back to the sampling algorithm your sampling elements into this set R X_i equal to 1; if the i th sample is less than or equal to the median. So, remember this is you have this less than or equal to here right that is what is defining your set which gives you whose cardinality is Y_1 ok. So, now, what you do is? You basically just define you break Y_1 into individual exercise X_i equal to 1, if the i th sample is at most the median value 0 otherwise, which simply means.

So, now the nice thing is X_i is very easy to work with because expectation of X_i is nothing, but a half I mean close to a half if you look at the textbook they are very careful and precise about computing the median I am being a little slack here, because then it depends on how you define the median is it the middle element if it is odd you have a middle element if you say even you do not have exactly a middle element. So, you look at the left of the middle in position what not we are ignoring that and saying look the expectation of X_i is half approximately close enough ok


And, when you have the X_i ; so, you can very easily compute the Y_1 ; Y_1 is simply the summation of the X_i which means now you can compute the expectation on the Y Y_1 because again linearity of expectation. And so the expectation of Y_1 is n to the three-fourth; why because? Y_1 the set R has cardinality n to the three-fourths and each time you sample the expected contribution towards Y_1 is half right. So, that is into the three-fourth times half that is expectation of course, that is only half the story you need the other half the variances as well.

(Refer Slide Time: 20:43)

Pr(①)

$\text{Var}[X_i] = p(1-p) \approx \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

By independence,

$$\text{Var}[Y_1] = \sum_i \text{Var}[X_i] \approx \sum_i \frac{1}{4} = n^{\frac{3}{4}} \left(\frac{1}{4} \right).$$


Again it is easy to compute the variance for the individual X_i ; it is p times 1 minus p if you recall this is just a Bernoulli random variable. And again the p s and the 1 minus p s are roughly a half ok. So, the variance is roughly a fourth and this is where independence helps and I recall somebody saying why did we sample with why not sample without replace with without replacement if you do it with replacement you get independence and so now, you can compute the variance of Y_1 simply by summing the individual variances ok. So, remember the individual variances are a quarter. So, sum over all i and remember there are i is running from 1 to n to the three-fourths. So, you get n to the three-fourths times 1 ok.

(Refer Slide Time: 21:45)

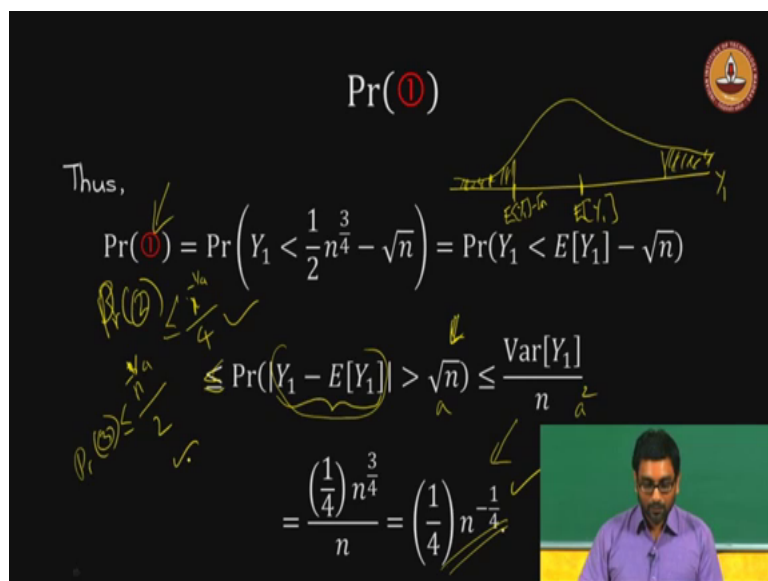

Pr(①)

Thus,

$\text{Pr}(①) = \Pr\left(Y_1 < \frac{1}{2}n^{\frac{3}{4}} - \sqrt{n}\right) = \Pr(Y_1 < E[Y_1] - \sqrt{n})$

$\leq \Pr(|Y_1 - E[Y_1]| > \sqrt{n}) \leq \frac{\text{Var}[Y_1]}{n}$

$= \frac{\left(\frac{1}{4}\right)n^{\frac{3}{4}}}{n} = \left(\frac{1}{4}\right)n^{-\frac{1}{4}}$

So, we have the two pieces the expectation and the variance. So, let us proceed from here. Let us go back to stating the bad event. Probability of the first bad event; if you recall is that; Y_1 is less than n to the three-fourths by 2 minus square root of n ok, but what is n to the three-fourths divided by 2 that is nothing, but the expectation of Y_1 ok. So, that is what and what we have what I am going to do now is rearrange the terms just a little bit.

So, let us make sure we get a picture of what is going on often it is in you know while working with these notations we can lose track of the intuition of what is going on what is this asking. So, let us say we have this Y_1 this is the let us say the expectation of Y_1 and this is expectation of Y_1 minus square root of n and we are asking what is the probability and Y_1 has some distribution we are asking; what is this probability? This is the tail bound right. We are asking; what is the probability? That Y_1 is going to be less than this quantity ok.

And if you recall those Chebyshev's had bounds on both sides. So, it is not just this part it also has this part ok. So, we are going to you know make life simple for us we are going to convert this equality into a less than or equal to and just bound both of them together this is only adding probability to a bad event and that is and if the upper bound still is within what we care about then we are fine right. So, this probability how are we going to capture this basically you are looking at this Y_1 minus the expectation of Y_1 and how do you fall in these two extreme when this when the absolute value of this quantity exceeds the square root of n then you are falling into either this region or this region that is what is captured in this event ok.

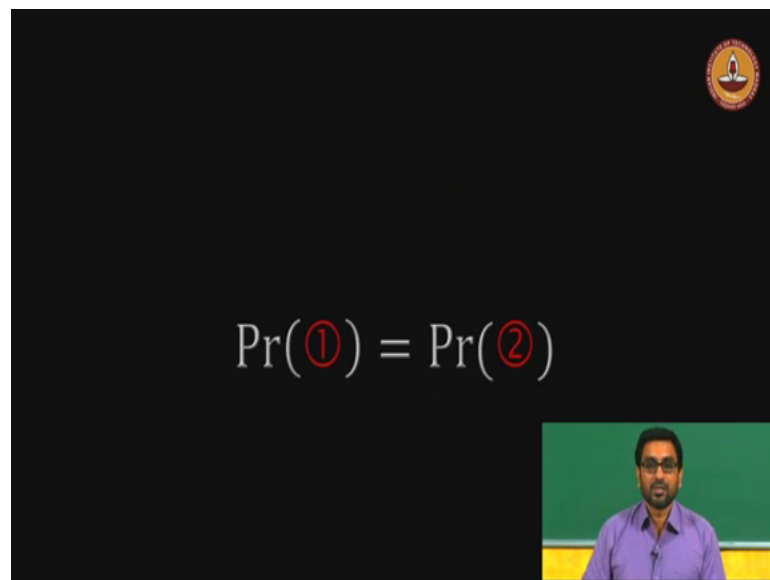
And this form is now exactly the way we want for applying Chernoff bounds ok. The probability of a random variable minus absolute value of a random variable minus its expectation exceeding some quantity square root of n in this case is at most the variance of Y_1 divided by the square of this quantity right. So, if it was a here it will be a squared here right. So, it is square root of n here.

So, it is n over here and if you work it out variance of Y_1 is n to the three-fourth divided by 4 divided by n ; it works out to one-fourth times into the minus one-fourth ok. This is one of some three different bad events right and it is contributing one-fourth of n to the minus one-fourth. So, if you plug it into the union bound there will be other. So, for

example, this is the first bad event. You can do the exact same thing and argue the same thing for probability of the second bad event as well; that is also going to be less than or equal to n to the minus 1 by 4 divided by 4 ok.

And then what we are going to do now is? We are going to show that the third event is going to be less than or equal to n to the 3 by 4 sorry minus 1 by 4 divided by 2. When you add up these 4 probabilities one here, one here and one here it adds up to n to the minus 1 by 4 and that is that is where we completed our analysis ok.

(Refer Slide Time: 25:37)



So; so, that is that is what I am stating here probability of the first bad event is equal to the second bad event. So, we are not we are going to skip the second bad event ok.

(Refer Slide Time: 25:44)

$\Pr(3)$

Top $2n^{\frac{3}{4}}$ elements in C are greater than median

Bottom $2n^{\frac{3}{4}}$ elements in C are lesser than median

$|C| > 4n^{\frac{3}{4}}$

Now, we are looking at the third bad event we need to be again a bit careful about, how to approach this bad event? Ok. This is the bad event that this C this third object in our algorithm C is too large ok. In particular you do not want to be larger than 4 times into the three-fourths ok.

So, let us say the bad event occurs you can you if which means now your C is cardinality is more than 4 times n to the three-fourths. So, then you can you can look at the bottom two times n to the three-fourths and at top two times n to the three-fourths and they are going to be disjoint right because it is the overall width is more than 4 times n to the three-fourths ok. So, you are going to notice that if one of few bad things will happen in this case.

If you notice this top to end to the three-fourth elements in C are all greater than the medium. The other possibility is that, but if you go back here, this bottom set is actually spanning going across the median ok, but at least one of them is fully to the right fully on one side of the median. The other thing that could happen is both of them could be one could be fully to the left and one could be fully to the right. This is also a possibility; the top two n to the three-fourth elements in C a greater than the median the bottom elements this bottom red portion is lesser than the median.

The third possibility is that the right part overlaps, but the left part; the bottom two n to the three-fourth elements in C are less than the median.

(Refer Slide Time: 27:37)

$\Pr(\textcircled{3})$

OR

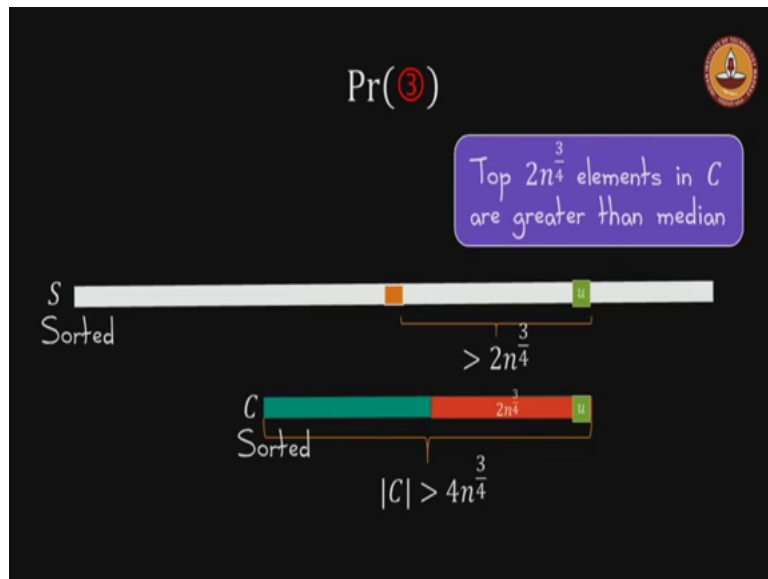
Bottom $2n^{3/4}$ elements in C are lesser than median

So, there are these three possibilities, but the best way to capture it is at least one of these two things will happen means the bad event happens. What is the either the bottom part is completely to the left the top part is completely to the right ok. So, basically the third event can be broken into two smaller bad events and these are again symmetric bad events. So, what we will do is we will just focus on one of them.

if we show that and just to be clear this is a bad event, but this if this bad event occurs it does not necessarily mean that the third bad event occurs ok. What this means is that? It is the other way around; if the third bad event occurs, then at least one of these two bad events occurs. These bad the smaller bad events occur. So, what we are going to do is really if you were to look at it from a Venn diagrams point of view let us say that this is the third bad event ok. So, if you what we what we can say is that if the if the we cannot say that if this were to happen that implies 3 that we cannot say, but what we can say is if 3 were to occur then one of these two occurs.

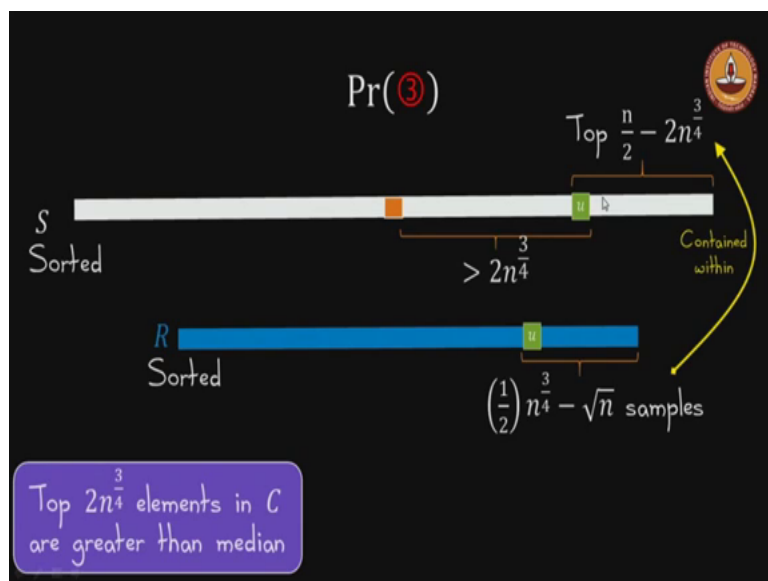
So, this what we are really doing is we are taking a larger event this is either this or this larger event is the or of these two events and what we are going to do is bound this larger event and for that what we are going to do is focus on the two separately and just focus on the first of these two bad events they might be. So, let us see not clear right. So, both can happen and neither can happen it is a little bit more complicated. So, let us not worry about that because it is I think both can also happen or neither can also happen ok.

(Refer Slide Time: 30:14)



These are all possibilities here, but let us just focus on this one bad event; basically what we are focusing is that the top $2n$ to the three-fourth elements in C are greater than the median ok. So, how will this happens basically, what is the rightmost element in this C that is this element u and what is happening is that element u is what is intuition here the element u occurring too far away from the median element this portion is greater than two n to the three-fourths that is why this red portion which is of size two n to the three fourth is fully to the right of the median ok.

(Refer Slide Time: 30:58)



So, now for this to have occurred; actually the problem started way back when we sampled in R , so that is why we are going back to R .

So, this is our u over here in R this is the original sample and somehow this u is where it is falling in the sorted set R ; it is falling at some position greater than $2n$ to the three-fourths away from the median element ok. How would that happen? Again it is a slightly counterintuitive way we are going to see how it is going to happen. How does u get pushed so much to the right? You look at all those elements to the rightmost portion of this set R ; that is basically this portion that is where is u it is it is in the middle half n to the three-fourths sorry this yeah minus minus square root of n this is correct because you are looking at these elements over here to the to the right ok.

All of these where should where should they have fallen where should these samples have come from they should have all been contained within the top n to the n half minus $2n$ to the three-fourth that is this portion all of these elements should have been contained over here that is; that is how this u can get pushed to the right ok. I will let you stare at this picture for just a second ok. Why is this happening? u gets pushed to the right it is its more than $2n$ to the three-fourth from the median how is this happening if you look at the sample set R and you look at the top portion top half n to the three-fourths minus square root of n samples all of them are drawn from this region. So, that is how u gets pushed so much to the right ok.

(Refer Slide Time: 32:56)

$\text{Pr}(\textcircled{3})$

Let X be the number of samples among top $\frac{n}{2} - 2n^{\frac{3}{4}}$ elements in S .

Let

$$X_i = \begin{cases} 1, & \text{if } i^{\text{th}} \text{ sample in top } \frac{n}{2} - 2n^{\frac{3}{4}} \text{ elements in } S \\ 0, & \text{otherwise} \end{cases}$$

Thus, $X = \sum_i X_i$ and

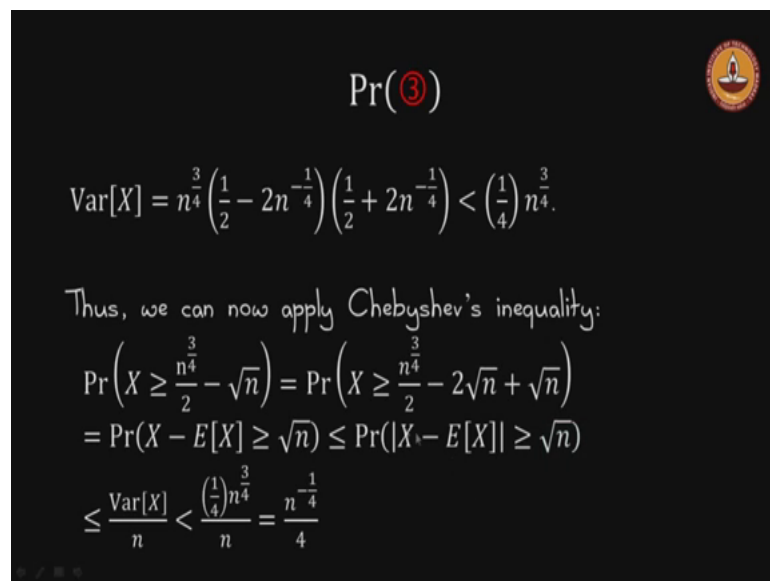
$$E[X] = \sum_{i=1}^{\frac{n}{2} - 2n^{\frac{3}{4}}} \frac{1}{n} = \frac{\frac{n}{2} - 2n^{\frac{3}{4}}}{n} = \frac{n^{\frac{3}{4}}}{2} - 2\sqrt{n}.$$

So, now we can write out this bad event ok. So, now, we for that we are going to define this random variable X ; it is the number of samples in this region in this top n by 2 minus

into the three-fourths. It is the number of samples among in R among the top into the samples in R, but drawn from the top n by 2 minus $2n$ to the three-fourth elements in S ok. So, now, again you define a random variable you break it into its components. So, X_i is one if the i th sample is meeting the requirement of X; otherwise it is 0.

So, this part should be obvious to you or capital X is simply summation of the individual X_i s. So, your expectation of x can you know is simply the summation of the individual expectations, but here we need to be careful what is the expectation of the X_i it is the probability that a sample will fall in the top n by 2 minus this $2n$ to the three-fourth region right and that; what is the probability? Is n by 2 minus $2n$ to the three-fourth divided by n and if you work it out in it you get this quantity into the three-fourths by 2 minus 2 square root of n ok.

(Refer Slide Time: 34:24)



$\Pr(\textcircled{3})$

$$\text{Var}[X] = n^{\frac{3}{4}} \left(\frac{1}{2} - 2n^{-\frac{1}{4}} \right) \left(\frac{1}{2} + 2n^{-\frac{1}{4}} \right) < \left(\frac{1}{4} \right) n^{\frac{3}{4}}.$$

Thus, we can now apply Chebyshev's inequality:

$$\begin{aligned} \Pr \left(X \geq \frac{n^{\frac{3}{4}}}{2} - \sqrt{n} \right) &= \Pr \left(X \geq \frac{n^{\frac{3}{4}}}{2} - 2\sqrt{n} + \sqrt{n} \right) \\ &= \Pr(X - E[X] \geq \sqrt{n}) \leq \Pr(|X - E[X]| \geq \sqrt{n}) \\ &\leq \frac{\text{Var}[X]}{n} < \frac{\left(\frac{1}{4} \right) n^{\frac{3}{4}}}{n} = \frac{n^{-\frac{1}{4}}}{4} \end{aligned}$$

Now, the variance similarly can be computed. So, you have the individual probabilities and what here what we are doing is any times you have p times $1 - p$ and p value is less than 1; it is quantity it is going to be at most of fourth ok. So, variance is going to be the summation of n to the three-fourths terms of the form p times $1 - p$. So, it works out to be one-fourth times n to the three-fourths these you can verify ok. So, simply now we apply Chebyshev's inequality probability that ok.

So, what is the bad event? That this the set of elements that we sampled from the top end to the this top into the n to the half n by 2 minus $2n$ to the three-fourths is greater than

this quantity n to the three-fourths and that is greater than this quantity n to the three-fourths by 2 minus square root of n . And, now set this up carefully because now you have this event X greater than this quantity you want the expectation is n to the three-fourths.

So, if you look at this n to the three-fourths by 2 minus 2 square root of n . So, I'm just going to subtract a minus square root of n and add a square root of n ; so, I get it of this form. So, so that I can write it as probability of X minus E of X greater than square root of n and that is less than or equal to; now I again do the same trick of not just bounding the left tail, but also the right tail and that is going to be X minus probability of the absolute value of X minus E of X greater than square root of n .

And this is now exactly the form that we want. So, we can apply Chebychev's inequality it is nothing, but the variance of X divided by n and it works out to be n to the minus 1 by 4 the whole divided by 4 ok. All of the bad events had the same type of probability. So, the first bad event had the same probability, second bad event by symmetry has the same probability upper bound on the probabilities.

Rather, the third bad event was broken into two pieces and one of them has this probability n to the minus 1 by 4 divided by 4 which means that the other mirror image will also have the same upper bound. So, all the bad events when you add them up there are 4 of them if you if you consider the third one broken into 2 there are four bad events each of the same n to the minus 1 by 4 divided by 4 .

(Refer Slide Time: 37:12)

Pr(③)

Top $2n^{\frac{3}{4}}$ elements in C
are greater than median

$\left(\frac{1}{4}\right)n^{\frac{1}{4}}$

OR

Bottom $2n^{\frac{3}{4}}$ elements in
 C are lesser than median

$\left(\frac{1}{4}\right)n^{\frac{1}{4}}$

So, this is that is what I am showing over here. So, you have these two mirror I mean symmetric bad events for corresponding to 3.

(Refer Slide Time: 37:18)

Theorem: $Pr(\text{Median Algo Fails}) \leq n^{-\frac{1}{4}}$

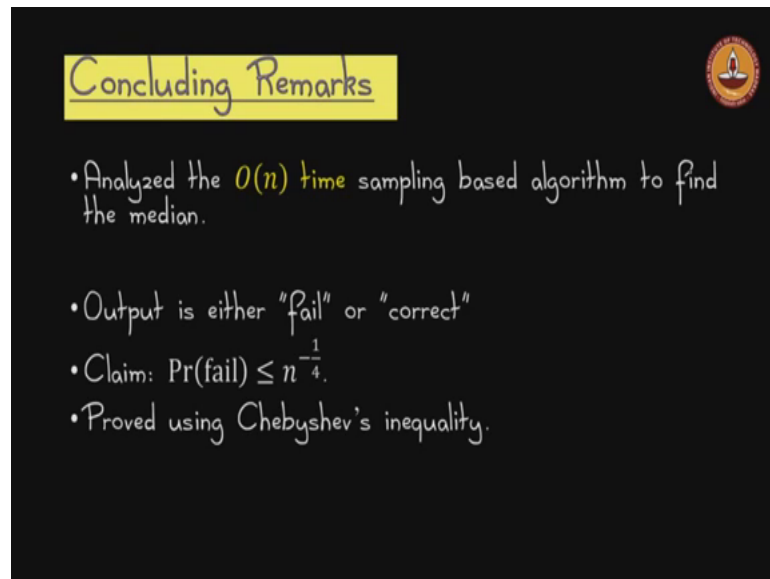
$$Pr(\textcircled{1} \cup \textcircled{2} \cup \textcircled{3}) \leq Pr(\textcircled{1}) + Pr(\textcircled{2}) + Pr(\textcircled{3})$$

$$\leq \frac{n^{-\frac{1}{4}}}{4} + \frac{n^{-\frac{1}{4}}}{4} + 2\left(\frac{n^{-\frac{1}{4}}}{4}\right)$$

$$= n^{-\frac{1}{4}} \quad \blacksquare$$

So, you put all them together you get the bad the upper bound and the probability of bad events is at most $n^{-\frac{1}{4}}$ and this is exactly what we wanted.

(Refer Slide Time: 37:29)



Concluding Remarks

- Analyzed the $O(n)$ time sampling based algorithm to find the median.
- Output is either "fail" or "correct"
- Claim: $\Pr(\text{fail}) \leq n^{-\frac{1}{4}}$.
- Proved using Chebyshev's inequality.

So, with that we can conclude. We basically I hope you got an appreciation for how Chebyshev's was applied in this context; we were able to as you rightfully pointed out it is defining the right variables setting it up the right way so that you can connect your intuition with the formula analysis ok. And the reason we one reason we do this is tuition is often great, but it in tuition can be also deceptive ok. So, when I started teaching probability in computing I thought that the previous median algorithm would also work with high probability that was my intuition and I was proven wrong and I actually was able to prove myself wrong by saying look you cannot actually not prove that with high probability you need this algorithm to prove this high probability bound on the median on finding the median ok.

So, the intuition is great to get you started, but often intuition can be misleading. So, you do need to learn how to analyse your algorithms more carefully ok.

(Refer Slide Time: 38:43)



So, that is the hopefully you are getting an appreciation for that. So, with that we will conclude we will next segment we will actually look at Chernoff bounds which is a little bit you know more powerful than chebyshevs, but in some sense restricted ok. So, that is it.

Thank you.