

Algorithms for Big Data
Prof. John Ebenezer Augustine
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Lecture - 36
Estimating Frequency Moments

Our topic for today is going to be estimating frequency moments.

(Refer Slide Time: 00:21)

The slide is titled "FREQUENCY MOMENTS" and contains the following handwritten text:

- Universe $U = \{1, 2, \dots\}$
- Stream of data arrives one item at a time
- $\forall i, x_i \in U$
- $x_1, x_2, x_3, \dots, x_m$
- $\forall j \in U, f_j = \# \text{ of occurrences of } j \text{ in the stream}$
- k^{th} freq. moment $F_k = \sum_{j \in U} f_j^k$

A yellow sticky note on the slide says "Captures many of our characteristics". There is also a small video inset of the professor in the top right corner and an NPTEL logo in the bottom left corner.

So, let us setup the problem. We have here universe U and that ranges values that is saved our lots of generality 1, 2 and so on and we are receiving a stream of data items and each data item is drawn from this universe and, the first items shows up that is x_1 , second items shows up x_2 , third item shows up x_3 and so on up to x_n and there can be repetitions within these items such show up in the data stream.

So, 1 way to think of this would be, for example, a data packets that go through node of the network and these x_i 's could indicate the IP addresses, the source destination pairs of each IP address and this might be able to give you some information about the flow of information or data packets from the network and I show 1 example there are variety of ways in which we can interpret this stream of data and we are just interested in understanding the stream of data.

In particular, we are interested in understanding what is called what is broadly called the

kth frequency moments. So, let us pick an item j in the universe, f_j denotes the number of occurrences of the item j in the stream throughout the stream from all the way from x_1 through x_m any time j occurs f_j is incremented. Let say let us think of it that way now the k th frequency moment is defined as f_k which denoted f_k and it is defined as the summation over all items in the universe, hence use the presentation j , it represents the items in the universe summation over $j f_j$ raise to the power k as it turns out this notion of k th frequency in moment captures a large number of nice characteristics. Now, notice that k is the prime rate and as the prime rate varies we get variation insights about the data string.

(Refer Slide Time: 03:08)

INTERPRETING FREQUENCY MOMENTS

What happens as $k \rightarrow \infty$?

- Most frequent element dominates
- For convenience $F_\infty = \max_j f_j$

What happens when $k=0$?

- Begs the question: what is 0^0 ? 0 by convention
- F_0 counts the # of distinct elements

NPTEL

Let us try to get the sense for what is this k th moment capture. We are not going to be completely thorough in our treatment; we are going to provide a broad burst stroke of the landscape of problems and ideas. So, what if what happens as k tends to infinity notice that the k th frequency moment, when k tends to infinity will be dominated by the most frequent term because the f_j corresponding to the most frequent terms in j will be raise to the power infinity and that is going to be larger than that quantity is may be larger than the other frequencies raise to the infinity. Well, what about k as it goes towards infinity and so, based on this intuition what are we going to do, we just going to conveniently redefine f_∞ , the infinity frequency node, the infinity frequency moment as just the frequency of the most frequent item.

Let us look at the other extreme, what happens when k equals 0? Now, this begs the question, what is 0 raise to the 0 because what happens if there is a positive frequency? Well that would be raise to the power 0 and we know that that value is 1, but if it is 0 raise to 0 well which is going to by convention use the value 0 for 0 raise to the 0.

This is quite nice because now f_0 because 0th frequency moment counts the number of distinct elements because any element that has a non-zero frequency that item f_j will be raise to the power 0 and will count as 1 and all items that have 0 frequency. On the other hand will not be counted, f_0 will discount the number of distinct elements. So, you see that f_{∞} has a nice interpretation f_0 also has a nice interpretation.

(Refer Slide Time: 05:59)

INTERPRETING FREQUENCY MOMENTS

$k=1$ simply counts the stream length
(recall Morris' Algorithm)

What about $k=2$? $F_2 = \sum_j f_j^2$

Intuitively measures how much the f_j 's vary

Perhaps Most Interesting & Important

Databases Measure size of self joins

NPTEL

14

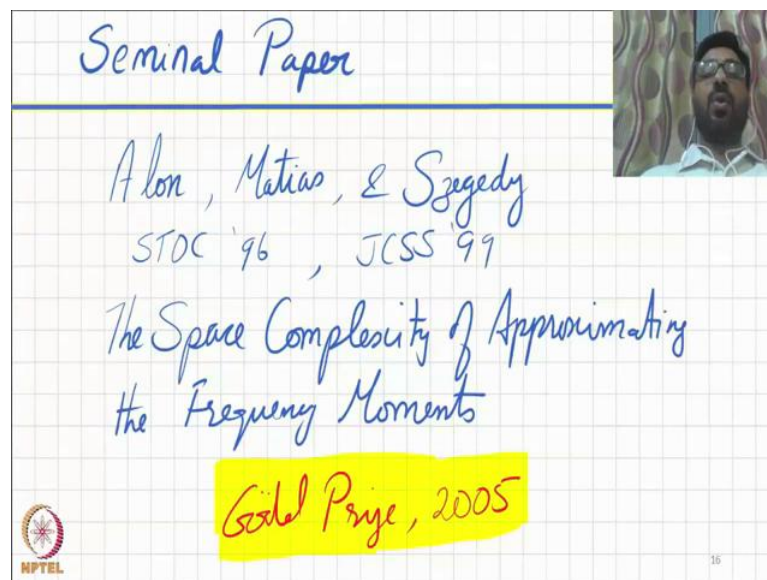
Let us look at k equals 1 and this simply counts the stream length the number of items that appeared in the number of times various items appeared in the stream because every time. So, think about this each item if it have appears f_j times it will account f_j towards the summation and so, when you sum over all the items these f_j 's will add up to the length of the stream.

Now, we are going to talk about k equal 2 which is very interesting and it is perhaps the most interesting and important frequency moment just to be clear what talking about f_2 which is summation over all items in the universe $\sum_j f_j^2$ square if you think about it this f_j^2 square intuitively measures, how much the f_j 's vary and is a. In fact, if you want to compute the variance of these f_j values then this second I mean the second frequency

moment is useful, but that is not all other some other surprising ways in which this will be useful.

For example, in the day in databases if you have a table that you join with itself is the self joint and let say you join the table with itself based on a particular attribute the frequency of values that the attribute takes let say represent f_j 's then the size of the such joint is going to be given by this second frequency moment. So, think about why that would be the case. So, here we again see 1 more application. So, as you can imagine f_2 shows up in a variety of contexts and so, it is just very interesting and important quantity to compute and as it turns out.

(Refer Slide Time: 08:46)



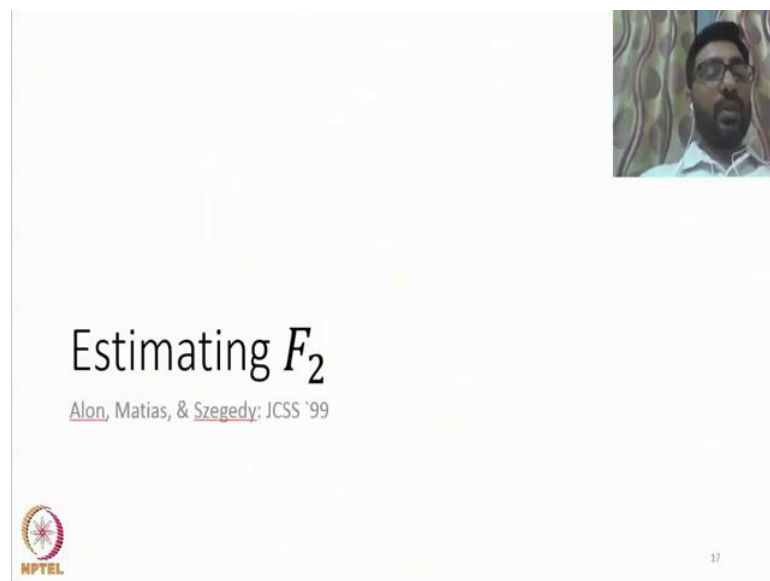
The first paper that talked about these frequency moments in the context of streaming sort of became a very seminal paper 1 that kind of opened up the floodgates for a number of different papers that appeared in this area.

So, this is the paper by Alon, now Alon, Matias and Szegedy appeared. I believe in STOC 96 later as a in general version and JCSS 99 title being the space complexity of approximating the frequency moments. So, this is a big deal because this paper ended up getting the Gödel prize in 2005 and this is a price given to the authors of papers that end up having a lot of influence over a period of time.

So, another example of Gödel prize winning paper is the prime it is in p paper by for

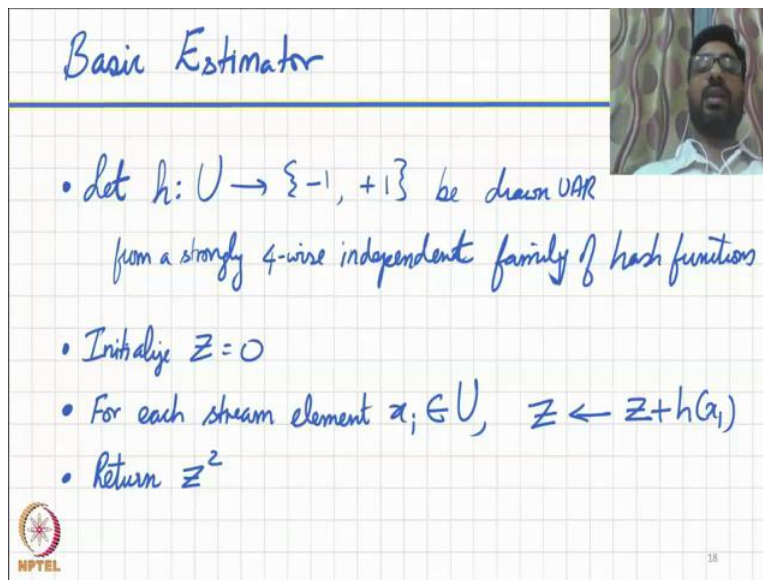
professor and students. So, you can imagine that this prize is given to papers that have had significant impact both at the time that they were in publish as well as over a period of time because they usually are papers that have a lot of impact over a period of time and in this case the whole field of stream in algorithms was opened up and studied in depth in large part because of the beauty of this paper and the beauty of the algorithm that they have presented so. In fact, that is 1 of the algorithms that they presented is what we are going to talk about today in particular, we are going to whom right into the most interesting frequency moment the second frequency moment.

(Refer Slide Time: 11:07)



So, without further adieu, let us get to the second frequency moment.

(Refer Slide Time: 11:09)



Basic Estimator

- Let $h: U \rightarrow \{-1, +1\}$ be chosen UAR from a strongly 4-wise independent family of hash functions
- Initialize $Z = 0$
- For each stream element $x_i \in U$, $Z \leftarrow Z + h(x_i)$
- Return Z^2

NPTEL

Let us begin by an estimator that is basic estimator. This will not be the final estimator that we arrive at, but it will just help us get a sense what is going on and just remind ourselves, we are estimating f_2 which is the second frequency moment. So, here is how the algorithm goes, now h is a hash function that takes elements from the universe and maps it to either minus 1 or plus 1 and such random variables that either take the value minus 1 or plus 1 are often called random occur with variables, so that is in a slide. So, we are going to assume that this h is drawn uniformly at random from a strongly four wise independent family of hash functions. Now, having drawn that hash function we initialize the counters z equal to 0 and now let the stream start.

Now, the elements stream 1 by 1 and for each element. So, z goes from x_1 to x_2 to x_3 and so on and let say we have element x_i at our hand. Right now what do we do with that we hashed that element using h . So, we get h of x_i and then we simply increment z with h of x_i now keep in mind h of x_i can be either positive or negative and. So, some of the elements are going to be incremented, some of the elements are going to be decremented, but keep in mind if there is an element that increments it, in other words h of x_i is plus then h of x_i will continue to continue to be plus 1 throughout the execution of this algorithm and vice versa, if h of x_i that being negative then it is going to be negative throughout.

Now, let the stream run through and we update z as and when a new item comes in and at

the end, we simply returned z square as the required estimate of f^2 . So, this is pretty amazing because at first sight, it is not clear at all as to why this algorithm works the only intuition that I can get at least is the following that z that starts off at 0 and then some elements randomly uniformly at random pull z to on the positive direction and other items pull that on the negative directions.

So, there is the sort of a tug of war and in this tug of war if the frequency of items is about the same for all items then the tug of war is going to be even whereas, if some items are going to be having higher frequency then the tug of wars going to move either in the positive direction or in the negative direction. So, remember that f^2 is a measure that captures how much the frequencies vary and this tug of war also does some about the same thing, but apart from that we really need to analyze the bit more carefully to really see what is going on. So, the first thing we need to do is to convince ourselves that is in fact, an estimator. In fact, we need ourselves is this even an unbiased estimator and what do we mean by that?

(Refer Slide Time: 14:44)

Basic Estimator is Unbiased

$$E[z^2] = E\left[\left(\sum_{j \in U} h(j) f_j\right)^2\right]$$

road map

- $\text{Var}[z^2]$ Not low enough
- Repeat in $||l$
- Take average Reduce Variance
- Use Chebyshev's to get (ϵ, δ) -approximation

$$= E\left[\sum_{j \in U} h(j)^2 f_j^2 + \sum_{j < l} h(j) \cdot h(l) \cdot f_j \cdot f_l\right]$$

$$= \sum_{j \in U} f_j^2 + \sum_{j < l} f_j \cdot f_l \cdot E[h(j) \cdot h(l)]$$

$$= F_2 + \sum_{j < l} f_j \cdot f_l \cdot E[h(j)] \cdot E[h(l)] = 0$$

By (q-wise) independence

Is this z square that we are claiming is an estimator, what is this expected value, does that matching the expected value of f^2 ? If it does then it would be an unbiased estimator and it still would not necessarily be a good estimator because it depends on the variance of z square and how close it is to f^2 on averages and with high probability and so on and so forth. We are not even going there yet what we want for now is to establish that z squared

is an unbiased estimator, in other words the expectations of z^2 is actually f^2 .

So, let us establish that to begin with, the expectation z^2 is well and summation. So, let us just expand out z^2 that is nothing as z is the summation over all items in the universe h_j times f_j that is because each item pulls it either in the positive direction or the negative direction that depending on h_j value for f_j number of times that quantity that is z and the whole squared gives you z^2 , and now let us expand this out we will get two's 2 types of terms in the first type of terms, we will have h_j^2 squared and in this if you notice h_j^2 whether h_j being plus 1 or minus 1 this squared term squared value is going to be a plus 1. So, in this case the h_j 's can be essentially h_j^2 can be replaced by 1. The other type of the other such terms are those which are of the form that that have that have not h_j^2 , but h_j times h_l , but j and l are 2 different elements in the universe and as you can see when you apply the linearity of expectation and also apply the fact that h_j^2 is simply just plus 1. The first type of terms give you this summation overall items in the universe f_j^2 and remember this is simply our f^2 , but we want to estimate and let us look at the second term.

Now, the expectation has been brought and based on due to linearity of expectation and f_j and f_l are not random variables. So, they can be brought out h_j and h_l are on the other hand random variables, but here is the interesting thing we remember, we have a four wise strongly four wise independent hash function here and h_j and h_l are independent of each other. So, the expectations the expectation of their product can be written as the product of their expectations, but each h_j on expectation is 0 because it is plus 1 with probability half and minus 1 with probability half and as a result the expectation of h_j or h_l for that matter they are all going to be equal to 0 and this product ends up being a 0 which means that this entire term vanishes and we are left with just f^2 .

So, this is pretty neat, the estimator that we have as an unbiased estimator, but that is, obviously, not all what we really want is a good epsilon delta approximation. What do we mean by that, for these particular epsilon and delta values we want to ensure that our estimate is within an epsilon fraction of the true value f^2 with probability at least $1 - \delta$ and for this we need to be a bit more careful, but does this is from now on these are techniques we have seen before. So, what is our road map? Now, we simply repeat the basic estimator in parallel several times and we take the average and this will mean that we can by this means; we can reduce the variance and then use Chebyshev's

inequality to get the required epsilon delta approximation.

(Refer Slide Time: 21:31)

Claim $\text{Var}[Z^2] \leq 2F_2^2$

$\text{Var}[Z^2] = E[Z^4] - (E[Z^2])^2$

what is this?

$E[Z^4] = E\left[\left(\sum_{j \in U} h(j) f_j\right)^4\right]$

$= \dots + E[h_{j_1} h_{j_2} h_{j_3} h_{j_4} f_{j_1} f_{j_2} f_{j_3} f_{j_4}] + \dots$

So, for that first let us figure out what the variance is of z squared just a basic estimator because we need to know how much we need to reduce it. So, for that we need to know how much we already have just in the basic estimator the variance of z squared has these 2 terms and the second term is e of z squared the whole square.

What is the e of z squared? Now, we already know what that is that just f 2. So, the second term is really f 2 square, so that once already done our worry now is about e of z to the 4, what is this e of z to the 4? Well, that is let us expand it out it is going to be the expectation over the summation over items in the universe we use j denote these items and it is summation over h summation over j h j f j that is z, but then this whole summation raise to the power 4 will be our z to the 4 and if we have the patience to expand this out we are going to get a long series of terms.

(Refer Slide Time: 22:58)

Claim $\text{Var}[Z^e] \leq 2F_2^2$

$\text{Var}[Z^e] = E[Z^e] - (E[Z^e])^2$

$E[Z^e] = E\left[\left(\sum_{j \in U} h(j) f_j\right)^e\right]$ |U|^e terms

$= \dots + E\left[h_{j_1} h_{j_2} h_{j_3} h_{j_4} f_{j_1} f_{j_2} f_{j_3} f_{j_4}\right] + \dots$

$\sum_{j \in U} h(j)^4 f_j^4$ quad

$\sum_{j_1 < j_2} h(j_1)^2 h(j_2)^2 f_{j_1}^2 f_{j_2}^2$ pairs

Case 1: \exists an index that is different from all others. $\Rightarrow 0$ [by principle of deferred decisions]

Case 2: All indices are pairs or quadruples

In particular we are going to get some cardinality of e raised to the power 4 terms. So, that is pretty long and boring in some sense, but let see there is some pattern and emerges.

Let the terms can have 1 of 2 forms. Now, there could be there is a possibility there is some of the terms are going to have 1 of the indices. Now, notice that these terms are of the form $h_{j_1} h_{j_2} h_{j_3} h_{j_4} f_{j_1} f_{j_2} f_{j_3} f_{j_4}$ and these indices j_1, j_2, j_3 and j_4 can take on any value from U cross. Now, if there is an index value that is appearing that is different from the other three index values then we have something nice happening the entire terms simply vanishes, why because when we take the expectation of the product we can apply the fact that itself, we have a four wise independent hash function.

So, it ends up being e of h_{j_1} times e of h_{j_2} times e of h_{j_3} times e of h_{j_4} and let say j_4 without loss of generality is the index that is different from all others that e of h_{j_4} will end up being 0 because as we just mentioned a little while ago the expectation of any h_{j_4} is going to be 0 because it is either plus 1 or probably half minus 1 probably half.

So, all of those terms and which there is at least 1 index that is different from all others are going to vanish. So, what are we going to be left with we are going to be left with the cases where the indices are either pairs or quadruples meaning. So, either of the form $h_{j_1}^2 h_{j_2}^2 f_{j_1}^2 f_{j_2}^2$ or $h_{j_1}^4 f_{j_1}^4$. So, I am calling $h_{j_1}^2 h_{j_2}^2$ to

the power 4 as quads and h of j squared times h of l squared as pairs, but of course, we know that h of j squared is going to be plus 1 h of l square is going to be plus 1 and similar way h of j raise to the power 4 is also going to be plus 1.

(Refer Slide Time: 26:10)

Claim $\text{Var}[Z^2] \leq 2F_2^2$

$\text{Var}[Z^2] = E[Z^4] - (E[Z^2])^2$

$E[Z^4] = E\left[\left(\sum_{j \in U} h(j) f_j\right)^4\right]$

$= \dots + E[h_{j_1} h_{j_2} h_{j_3} h_{j_4} f_{j_1} f_{j_2} f_{j_3} f_{j_4}] + \dots$

$= \sum_{j \in U} f_j^4 + 6 \sum_{j < l} f_j^2 f_l^2$

$\leq 3F_2^2$

But $F_2^2 = \left(\sum_{j \in U} f_j^2\right)^2$

$= \sum_{j \in U} f_j^2 + 2 \sum_{j < l} f_j^2 f_l^2$

$\text{Var}[Z^2] \leq 3F_2^2 - F_2^2 = 2F_2^2$

So, we can simply write this e of z to the power 4 as summation over j f j to the power 4 plus 6 times summation over all pairs j and l f j square f l square and where is this 6 come from, we get the 6 from the following reasoning.

Now, we have 4 indices, out of these 2 of them are j's the 2 smaller ones and then 2 of them are l's now given a particular j and l we have 4, choose 2 locations or pairs of locations in which the j values can fall and that is why we have the 4 choose 2 over there and as a result we get 6 coefficient here, but let us recall what f 2 is. So, remember that we have reached certain value certain expression for e of z to the 4 in order to proceed just let us look at what f 2 squared is what is f 2 square is just the summation over all j f j squared the whole square and of course, when we expand that out we get summation over j f j squared plus twice the summation over all j and l f j squared f l square this means that 3 times f 2 squared is going to be certainly larger than what we have over here.

So, we apply that bound to get required bound on e of z to the 4. So, e of z to the 4 is at most 3 times f 2 square and now going back to the variance of z squared, we can apply our bound on e of z to the 4 and that is going to be 3 f 2 squared and we already know

that e of z squared the whole square is f^2 square also. So, that is going to this variance of z square ends up being at most twice f^2 squared.

(Refer Slide Time: 28:52)

The slide is titled "Recall Roadmap" in blue cursive. It contains a list of handwritten notes in red ink:

- Road map
- $\text{Var}[z^2]$ Not low enough
- Repeat in $||l$
- Take average
↓
Reduce Variance
- Use Chebyshev's

There is a green checkmark to the right of the first two items. A small video inset in the top right corner shows a man speaking. The NPTEL logo is in the bottom left corner, and the number 35 is in the bottom right corner.

So, just reminding ourselves, the road map we now have a handle on the variance of z squared for the basic estimator. What we need to do is figure out how to repeat it in parallel and remember this variance is too much, we need to be able to reduce the variance a little bit and for that we will repeat this basic estimator and in repeated and there therefore, reduce the variance to a convenient value, so that we can then apply Chebyshev's inequality in order to get the epsilon delta estimator that we want.

(Refer Slide Time: 29:36)

FINAL ALGORITHM

- Choose $h_1, h_2, \dots, h_t: U \rightarrow \{-1, +1\}$
draw UAR from a strongly 4-wise indep family → *To be fixed later*
- Initialize $Z_i = 0 \quad \forall 1 \leq i \leq t$
- For each new item j in the data stream
 - For each $i = 1, 2, \dots, t: Z_i \leftarrow Z_i + h_i(j)$
- Return $Y = \frac{1}{t} (Z_1^2 + Z_2^2 + \dots + Z_t^2)$

NPTEL

So, here is our final algorithms, we are ready to state the final an algorithm and as you can imagine this is simply the basic estimator just repeated in parallel several times. In particular, we are repeating it t times and what is their exact value of t that is going to be fixed later and recall these hash functions h_1, h_2 up to h_t are all drawn uniformly at random from a strongly 4 wise independent family and these hash functions map the universe to either plus 1 or minus 1. So, now, instead of 1 counter that we have previously we had just z . Now, we have the z_i 's where i ranges from 1 to t and as each allow the stream to pass by as each item in the stream arrives you look at the item and you hash the item using each of the of the t hash functions.

So, for example, when you hash using h_i you get h_i of j you add that to z_i that is the i th counter and you get the new update of z_i and you can see do this for all the i and after the stream as gone by we are left with t counters with various values and then we simply square each of the z_i 's and take the average of those squared values and that is going to be our estimator. This is our final estimator and we need to show that this final estimator it has all the properties that we need, for example, we already know hat this is an unbiased estimator, but we need to be able to prove that this is an epsilon delta approximation of f^2 .

(Refer Slide Time: 31:46)

FINAL QUESTION: $t = ?$

$$\text{Var}[Y] = \text{Var}\left[\frac{z_1^2 + z_2^2 + \dots + z_t^2}{t}\right]$$

$$= \frac{1}{t^2} \text{Var}[z_1^2 + z_2^2 + \dots + z_t^2] = \frac{1}{t} \text{Var}[z^2]$$

$$\leq \frac{2f_2^2}{t}$$

Now applying Chebyshev's inequality

$$Pr(|Y - E[Y]| \geq \epsilon f_2) \leq \frac{\text{Var}[Y]}{\epsilon^2 f_2^2} = \frac{2f_2^2}{t\epsilon^2 f_2^2} = \frac{2}{t\epsilon^2} \rightarrow \delta$$

$t \geq \frac{2}{\epsilon^2 \delta}$

Basic Estimator

NPTEL

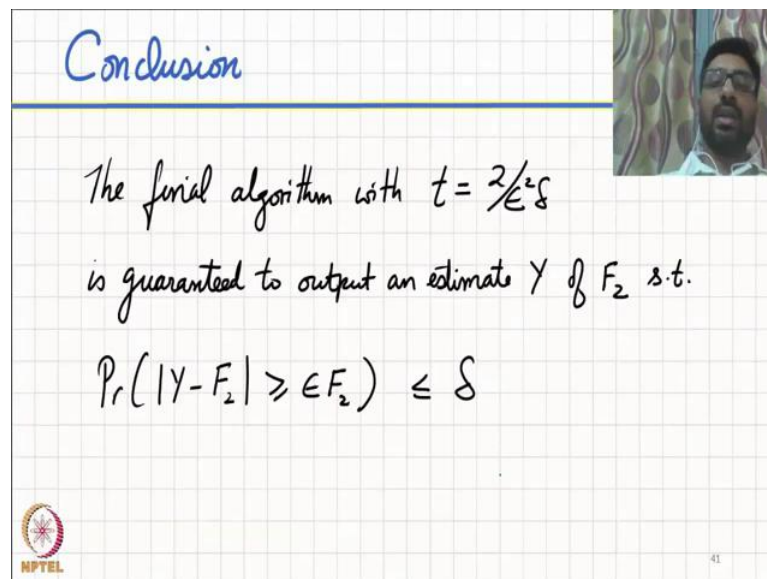
It is a matter of figuring out what the appropriate t value is this? Once, if we repeat an appropriate number of times the variance is going to come down sufficiently. So, really that is the only question at hand. So, what is the variance of y and that is just the variance remember y is just z_1 square plus z_2 square and so on up to z_t squared divided by t this is the average of the sum of squares and. So, you get the t outside you get 1 over t squared times the variance of the summation of squares, but keep in mind that each 1 of these z ones z_1 z_1 square z_2 square, they are all essentially the basic estimator. So, their variances are the same and so, you can simply think of it as t times the variance of z square.

So, there will be a t in the numerator and t square in the denominator. So, we end up with 1 over t variance of z square and we already know that variance of z squared is at most twice f_2 squared. So, variance of y is it most twice f_2 squared divided by t . So, what you can see is as you increase t and the variance of y keeps decreasing. So, that is useful for us. Now, we are ready to apply our Chebyshev's inequality. So, let us apply Chebyshev's inequality, what is the probability that y minus the expectation of y is greater than an epsilon fraction of f_2 ? Remember expectation of y is this f_2 .

So, it is really what we are asking is what is the probability that y deviates from f_2 by more than epsilon times f_2 and direct application of Chebyshev's inequality is going to say that it is variance at most variance of y divided by epsilon f_2 and up applying the

fact that variance of y is 2 at most $2 f^2$ squared by t , we get this expression over here and we have some cancellations is the f^2 squared terms cancels out, we are left with 2 over t epsilon square and we want that 2 over t epsilon squared to be δ that is the probability the of the bad event where the random variable the y value exceeds. It is epsilon approximation range and so that has to be at most δ which means that we will get that requirements satisfied, if we set t to be at least 2 over epsilon squared δ which is all that we really need.


(Refer Slide Time: 35:11)



Conclusion

The final algorithm with $t = \frac{2}{\epsilon^2 \delta}$ is guaranteed to output an estimate Y of F_2 s.t.

$$\Pr(|Y - F_2| \geq \epsilon F_2) \leq \delta$$

 41

So, really with that we can conclude that the final algorithm that we stated a little while ago with t value set to 2 over epsilon square δ is guaranteed to output an estimate y of f^2 that ensures that the probability y deviates from f^2 by more than epsilon times f^2 is at most δ . So, this is exactly what we wanted and we have that. So, this is a very nice clean algorithm for what which is quite surprising because at first sight at least this problem does not come across like as having such a clean algorithms quite surprising, but we do have such an algorithm and very clean simple analysis which is well.

So, that brings us to the conclusion of this lecture.