Bandit Algorithm (Online Machine Learning)

Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture – 56
Introduction to Pure Exploration and its Lower Bounds

We are going to start looking into another setup in multi arm bandits called pure exploration today. So, where we do not care about regret or what is the reward I am going to accumulate till some point or it may be the case that what I only worry about is whether finally, the arm you gave me is the optimal one or not ok. So, let us consider an example.

So, you all know that like before all these drugs are tested on in a laboratory in a lab on some animals, let us say mouse let us say or on mice. So, what they will do? They will have set of drugs they want to test and finally, identify which is the drug which is most efficient.

When they are doing these trials, they will not worry about how many mouse or mice they killed right like because some of the drugs what they want is at the end I whether I identify the best drug or not, so that when I apply on humans the risk are minimal or that is the most effective drug. But, while I am applying on testing it on the mouses or mice I do not really care.

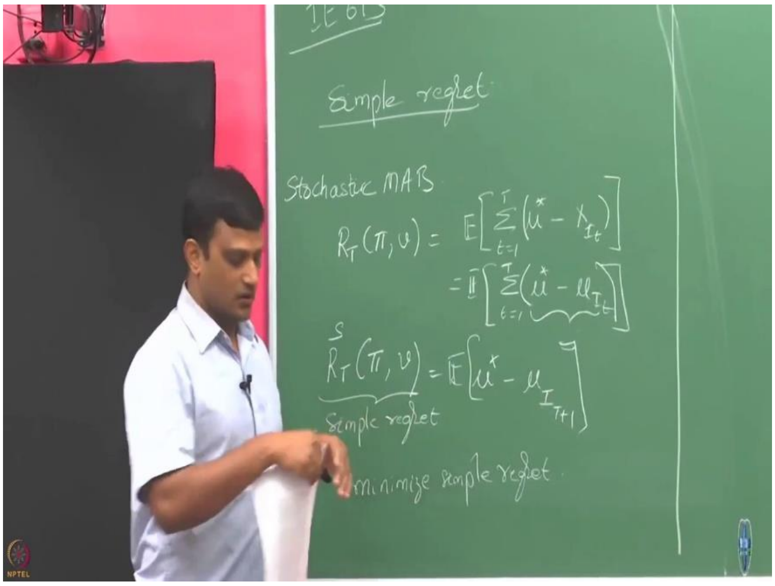
So, what for me there matters is whether did I identify the most efficient drug or not. So, during the exploration I am not worrying about like how many mouses I really saved. It is fine like as long as that that experiment is helping me in identifying which is the good drug I have full freedom.

And, similarly let us say you want to; you want to identify which is the best restaurant in surrounding area of Powai what you could just do is if you do not care about the money you are going to spend you just go and eat in every place every time and at the end you have to recommend to some one of your friend or let us say your parents have come and you have to take them, then you are going to say ok, so, let us go for this I have already explored and let us go to this restaurant.

So, when you are exploring you did not care whether you are to get the best or not, but what you are trying to do is just identify which is the best one. So, in this setup in a way you are not worried about the regret right what is happening, what is the cumulative reward or loss you are going to incur some time.

Now, such problems are usually called as pure exploration problems and in that also there are different-different versions ok. So, we are going to start with the case of simple regret ok.

(Refer Slide Time: 03:29)


IE 613
Simple regret
Stochastic MAB
$R_T(\pi, \nu) = \mathbb{E}\left[\sum_{t=1}^{T}(\mu^* - X_{I_t})\right]$
$= \mathbb{E}\left[\sum_{t=1}^{T}\underbrace{(\mu^* - \mu_{I_t})}\right]$
$\underbrace{R^S_T(\pi, \nu)}_{\text{Simple regret}} = \mathbb{E}\left[\mu^* - \mu_{I_{T+1}}\right]$
minimize simple regret

Let us recall our standard stochastic multi arm bandit case. So, in the standard stochastic multi arm bandit case, how did we put our criteria? We evaluated the performance based on the regret we are going to incur. How was the regret defined there?

In the stochastic we defined regret of an algorithm π over an environment ν as I could write it as $R_T(\pi,\nu) = E[\sum_{t=1}^{T}\left(\overset{*}{\mu} - X_{I_t}\right)]$ or let me just write it as $R_T(\pi,\nu) = E[\sum_{t=1}^{T}\left(\overset{*}{\mu} - \mu_{I_t}\right)]$ So, this is the reward sample you got when you played I_t arm, I am just took the expectation, but I_t still a random quantity that is why because that depends on all the past observations.

So, we are going to look at this. Let us take a particular time t, what is this quantity here? This quantity we can treat it as the loss of the regret instantaneous regret I incurred in round t and the regret is what? It is sum of the instantaneous regret you are going to incur in each of these rounds.

Now, instead of worrying about what is the instantaneous rounds and, how they accumulate over a period of time suppose let us say I am only worried about what is the; what is the regret I am going to incur of. So, I_{T+1}, what is I_{T+1}? This is the arm that is recommended by this algorithm at the.

Student: (Refer Time: 05:57).

After the T-th round that is in the (T+1)th round, you are worried about this. You whatever in T rounds you wanted to do let us say you did it, but now you wanted to worry about this quantity and let us worry about the expected value of this quantity. And, now this we are going to call it as a simple regret. Simple regret I am going to write it as simple of what our term π you are going to have.

So, till first T rounds whatever I got I do not care, but now my performance is defined in terms of what would I have got in the (T+1)th round and my goal now I am going to set it as minimize this quantity. This is (Refer Time: 06:58). So, anyone has simple algorithm to minimize this simple regret what could be simple way?

Student: Arms equal number of time.

Ok.

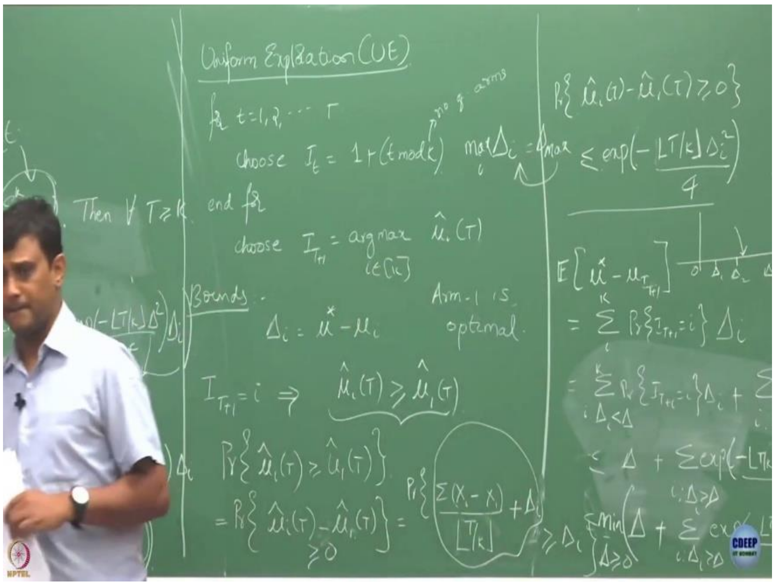
Student: And then choose the arm like mean will be maximum.

Which has the highest mean.

Student: Highest mean.

Empirical mean right ok. So, that is what is that is what similarly we actually did in explore and commit algorithm also right, but there we also worried about the entire regret, but let us say I apply the same algorithm here and evaluate my simple regret ok.

(Refer Slide Time: 07:50)


Uniform Exploration (UE)
for t = 1, 2, ... T
choose $I_t = 1 + (t \bmod k)$
end for
choose $I_{T+1} = \arg\max_{i \in [k]} \hat{\mu}_i(T)$
Bounds
$\Delta_i = \mu^* - \mu_i$
Arm-1 is optimal
$I_{T+1} = i \Rightarrow \hat{\mu}_i(T) \geq \hat{\mu}_1(T)$
$\Pr\{\hat{\mu}_i(T) \geq \hat{\mu}_1(T)\}$
$\leq \exp(-\lfloor T/k \rfloor \Delta_i^2 / 4)$
$E[\hat{\mu} - \mu_{T+1}]$
$= \sum_i^k \Pr\{I_{T+1} = i\} \Delta_i$

So, whatever the algorithm you set that is like let us call that as a simple uniform exploration. What this algorithm is going to do is simply the first T rounds it is going to spread along each the arm equally and then it is just going to play the best one at which is best in the sense that which is empirically best. So, for. So, here k is the number of arms and then.

So, what is $\hat{\mu}_i(T)$, is this is the empirical estimate of arm I you have gotten till round T fine. So, you are just doing this then the question is how what is what will be the simple regret for this simple algorithm. How to bound this? Ok. Let us try to get the bound for this we already have done it the bound we are going to get is very similar to what we did it in the explore and commit algorithm.

I am going to denote by $\Delta_i = \overset{*}{\mu} - \mu_i$; $\overset{*}{\mu}$ is the mean of the optimal arm and μ_i is the just like mean of the i-th arm. Now, let us say $I_{T+1} = i$. So, let us say I happen to choose I_{T+1} equals to i-th arm, what does this mean? It means that.

Student: maximum.

It should be ok. So, I am going to assume that arm one is the optimal arm ok. So, I am just without this is without loss of generality algorithm does not know, but for unless point of view just assume that arm one is the optimal arm. Now let us say.

So, in this case if I have chosen i-th arm then it must have happened that the mean of this would have been at least equal or greater than the mean of the optimal arm ok. Then we want to bound what is the probability of this. By the way how many times each arm is sampled in this?

Student: T by k.

T/k. Let us say T/k is a integer for time being, now how to bound this? Ok. We know this what is this quantity this is nothing but so, or I am going to just write it as $\Pr\{\hat{\mu}_i(T) - \hat{\mu}_1(T) \geq 0\}$. This I am going to simply write it as I want this to be greater than or equals to 0 ok.

I did some manipulation here. I know that $\hat{\mu}_i$ has been obtained by averaging T/k samples of i-th arm right. And, similarly $\hat{\mu}_1$ is nothing, but T/k. I have so, average of X_i averaged over T/`k number of rounds.

Student: sir shouldnt like the floor to keep the (Refer Time: 13:48).

Floor of what?

Student: Floor of T/k.

So, the number of samples I am going to get for both of them.

Student: Because total number of times we are going to place T.

T yeah.

Student: So, it should be less than t by (Refer Time: 14:07).

Let us keep it like that let us say the everything for us goes through and I am just not writing the indices here, but these indices are this sum is assumed to be over T/k number of samples. And, I know that X_i has mean μ_i and X_1 has mean μ_1 and the difference between their means is Δ_i that is what I added on both sides ok. But, the difference here is $-\Delta_i$.

Student: $\mu_1 - \mu_i$.

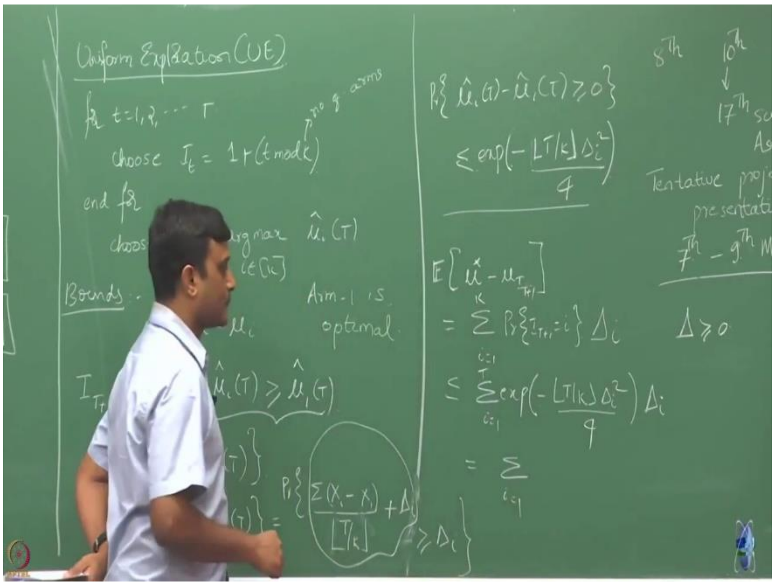
Yes, $\mu_1 - \mu_i$ is Δ_i, but that is fine this is the expectation of this is going to be.

Student: (Refer Time: 14:58).

$-\Delta_i$. But, if you look into this, is this a zero mean random variable? Yes, right it is a zero mean random variable. Let me make one more assumption that assume that the each of the arms have sub Gaussian distribution and specifically we are going to assume that they are all one sub Gaussian. So, that the sigma square parameter is 1 and they are independent across the arms. So, now, in that case what is the distribution of this going to look like? Yeah?

Student: Root 2 sub.

(Refer Slide Time: 15:55)


Uniform Explration (UE)
for t=1,2,··· T
choose $I_t = 1 + (t \bmod k)$
no of arms
end for
choose ... argmax $\hat{\mu}_i(T)$
$i \in [k]$
Bounds
Arm-1 is optimal
μ_i
I_{T+1}
$\hat{\mu}_i(T) \geq \hat{\mu}_1(T)$
$\Pr\left\{ \frac{\Sigma(X_i - X_1)}{\lfloor T/k \rfloor} + \Delta_i \geq \Delta_i \right\}$
$\Pr\{\hat{\mu}_i(T) - \hat{\mu}_1(T) \geq 0\}$
$\leq \exp\left(-\frac{\lfloor T/k \rfloor \Delta_i^2}{4}\right)$
$E[\hat{\mu}^* - \mu_{I_{T+1}}]$
$= \sum_{i=1}^{K} \Pr\{I_{T+1} = i\} \Delta_i$
$\leq \sum_{i=1}^{T} \exp\left(-\frac{\lfloor T/k \rfloor \Delta_i^2}{4}\right) \Delta_i$
$= \sum_{i=1}$
8th 10th
17th
Tentative proj presentati
7th – 9th
$\Delta \geq 0$

It is going to simply $\sqrt{2}$ sub Gaussian. Now, if I am going to apply. So, if I am going to apply $\Pr\{\hat{\mu}_i(T) - \hat{\mu}_1(T) \geq 0\}$. Now, if I am going to apply the property we have for a sub Gaussian noise this is going to be $\exp\left(\frac{\lfloor T/k \rfloor * \Delta_i^2}{4}\right)$. So, in the denominator we usually got $2\sigma^2$ right, but σ^2 is what here?

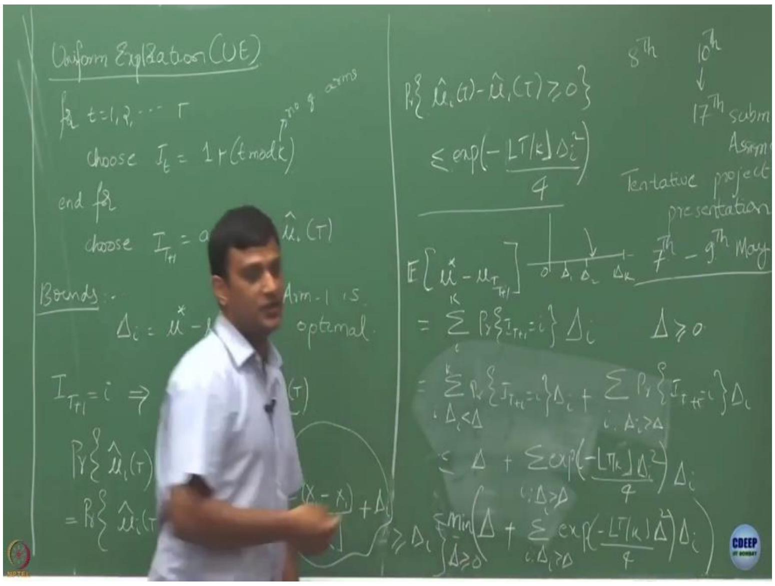
Student: 2.

It is going to be 2. So, it is going to be 4 now. Notice that the when you are going to do the averaging right the number of samples are fixed. The number of samples are no more random because I know that after T rounds I would have had each the number of samples is this floor of T by K. So, this quantity is a deterministic in this case. That is why I could apply my result that I have for the sub Gaussian case ok.

Now, once I have this I am now going to bound my regret sorry my simple regret. What is my simple regret? My simple regret is this is nothing, but probability that and what is this probability? Probability that I_{T+1} equals to i is upper bounded by what? By this quantity right because this probability that I_T equals to 1 is exactly this condition and we have bounded that probability.

So, this is going to be like. So, now, to tighten this bound what we are going to do is let Δ be some quantity whatever you choose delta be some quantity. Now, I am going to split this sum as maybe like before this I will do this not here.

(Refer Slide Time: 19:07)



So, I am going to split this sum as $\sum_{\Delta_i<\Delta}^{k} \Pr\{I_{T+1}=i\}\Delta_i + \sum_{\Delta_i>\Delta}^{k} \Pr\{I_{T+1}=i\}\Delta_i$. I have split it into these two format based on what is the delta I have chosen ok. So, I know that my now this is going to be d_0 this is somewhere Δ_1, this is going to be Δ_2 all the way up to this is going to be somewhere Δ_k .

I have chosen somewhere Δ here and that is why I could split it. I mean it may happen that delta is beyond this point also. In that case in this case it is going to be simply one sum, the second sum will not arise, but as of now I have chosen this Δ arbitrarily.

Now, this quantity here I know that this Δ is upper bounded by so, the Δ is upper bounded by this quantity Δ I find, then I will going to pull it out. Then this is nothing then it is going to be the sum of probabilities for which this condition holds, but I know that there is some probabilities can be at most 1. So, that is why I am going to upper bound it as Δ, but the other quantity is going to still written like this, but in the other quantity yeah, I am going to I am going to.

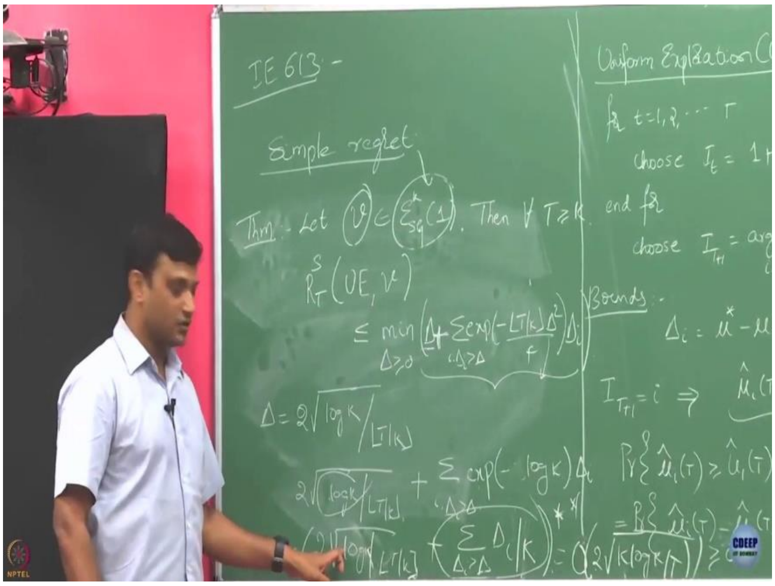
So, this one I cannot bound it in a way I have done here because Δ is a lower bound on Δ_i here.

Student: Probability.

Yeah? Yeah, but I will use the probability term here bound on the probability. But, here this Δ_i^2 I have this Δ I know that this is lower bounded this Δ is lower bounded by this Δ right. So, if I replace this Δ_i here by Δ will I still get an upper bound? So, I will do that.

Now, the way I have chosen this Δ here is arbitrary right. I have just said there is some $\Delta \geq 0$. So, if I am going to do instead of this min over all these quantities such that $\Delta \geq 0$ it is bounded still a valid bound ok. So, this is what we are going to state as a result now what we have just proved.

(Refer Slide Time: 23:16)



So, I am going to use this notation $ò_{sg}^{k}(1)$, this means this is the class of distributions which are sub Gaussian with sub Gaussian parameter 1. So, when I say this ν belongs to this means each of the arm is sub Gaussian distributed with parameter 1 and k denotes already the number of arms in that.

Then for all T greater than or equals to K what we have just showed is R_T simple regret of this guy uniform exploration is upper bounded by what? min over (Refer : 24:31). Now, what does this bound tell us? So, if you look into this how does the T coming into picture here the number of rounds? It is coming as negative exponential in this it is coming as a negative part in this exponent, that is if I increase T this term is going to vanish exponentially fast right.

So, this uniform exploration is going to be asymptotically optimal in that sense right because as I let T goes to infinity, this term is going to vanish and minimization I am going to take it toward equals to 0 and then this is going to say upper bounded by 0. I know their simple regret is already non-negative quantity. So, it is going to be 0. So, this simple algorithm is going to give us a simple regret that is going to be falling exponentially fast in the number of rounds.

Student: (Refer Time: 26:15).

Central limit theorem what do you want to say?

Student: T is approaching to 0 we will get the empirical mean close to the?

Yes. So, he is claiming like if I am letting T go to infinity there also I am getting the right means right. So, I should not be making any mistake. The same thing is also being said here yeah.

So, when I said this is going to go to 0, this simple regret exponentially fast that is for a you fix a environment and then you increase the number of rounds ok. But, now let us say if I fix the number of rounds to T and if I keep changing and am applying. So, let us say I will give you the I will give you fixed T and I am now asking you to apply it on different-different environment. So, how is that?

In that case if I fix T and if I am going to change this environment that is the Δ_i are changing what could be the bound on this? So, you see that this guy here if I increase Δ this guy is falling exponentially fast, but it is also there is an increasing parameter Δ_i here that because that is going to increase the regret. So, let us see and also there is a Δ here.

So, let us see if I am going to fix T and let us try to get a problem independent bound here that irrespective of whatever Δ is what is the regret I am going to get? One thing you can do is suppose I set environment and let us say if I am going to set my I will going to choose my Δ such that; so, Δ I said is arbitrary right it is just going to be greater than or equals to 0 in this I am going to choose a particular Δ here in this part here.

If I am going to choose a particular Δ in this fashion can you quickly plugin this Δ into this equation and see what is the value you are going to get? So, Δ is going to be in that case this quantity is going to be Δ is this plus what is this going to do? It is going to if I square it just let me plug in myself here. So, T is going to be this is going to be 4 right one 4 get canceled, this guy get canceled I am going to get.

Student: Minus minus one.

$\exp(-2\log k)$, I am just going to written it and this is going to be simply Δ_i.

Student: Sir we get 0.

Will not only this 2 get canceled with this 4 here?

Student: Sir (Refer Time: 29:55).

is it that? Let me just cross verify with calculation. So, if I do del square here

Student: (Refer Time: 30:01).

it is going to be 4 into 2 log k divided by this quantity. So, 4 and 4 get canceled. Is it not because of this 2 this 2 will complete?

Student: no its two root log k because Δ

2 2 log k.

Student: (Refer Time: 30:16) written something (Refer Time: 30:17). I think that delta.

I say did I fine. So, it is not like that this is (Refer Time: 30:22) right in that case it is simply going to be what? 2.

Student: Plus.

Log k one whatever this term T/k, this is going to be 1/k; and how many terms here at most? k times will be there, right and if we are and what is this Δ_i?

Student: Bounded sub Gaussian bounded.

Yeah, it is sub Gaussian right we do not know like whatever ah, but let us let us write it whatever it is a fixed content right. So, this is going to be k*(1/k) it is simply going to be in that case is Δ_i is this.

$$2\sqrt{logk/(T/k)} + \sum_{\Delta_i > \Delta} \Delta_i / k$$

Student: one by (Refer Time: 31:16).

1/k I multiplied there is at most this sum can be at most K, right?

Student: Yes.

I just I removed the summation ok that is summation is to be ok, if I am going to keep that ok. Now, let us see if I want to make it independent of this quantity fine. So, this we do not know like really what is the value of μ_i right, but we know that these are some finite quantities ok. I mean they are not at least infinity rights and so, that this Δ_i are what the difference between the means. So, they are further bounded ok.

So, if you are going to assume that all this means are all some finite then this is going to be just a constant right. So, this summation is going to be some constant in that way. So, then I have I will just write it as if I assume that the support is between [0,1].

Student: (Refer Time: 32:23).

They are all be bounded by 1 that is fine, but here if I if when I started saying that I have just said that this guy belongs to a class of sub Gaussian distributions. So, I have not specified about where the support is if the support is in the interval [0,1] it is still true and everything I can bound it by 1, this is going to be a constant.

If not I can always assume that this means are some finite value. For this entire class I can assume that for this whatever my entire class it is there right. I can assume that the means are some finite, once I know that is finite then this quantity is going to be bounded in terms of that constant value ok. So, suppose assume that you have you are dealing with the class of sub Gaussian noise right.

Student: Sir, you do not know how many terms have Δ_i right?

No, I know they are at most k terms in this.

Student: So, then it know obviously counted by k times the Δ max right.

Δ max yes, but what is that Δ max?

Student: So, you can (Refer Time: 33:26) if you are (Refer Time: 33:27) about the mean it is not then (Refer Time: 33:28).

So, that is indirectly right, if I bound me Δ max is already bounded.

Student: Yes, you do not need to (Refer Time: 33:34) indirectly, you can directly bound that (Refer Time: 33:35).

So, I am trying to make sure that instead of writing Δmax, delta looks like a problem dependent quantity right. I want to just write it down as a constant which depends on the entire class, not for that particular distribution. So, as I am like when I want to trying to write this I want to come up with a bound which is problem independent.

Δ max could be depends on the particular problem instance, but I do not want it to a problem instance dependent. I want it to be a parameter which depends on the entire class not on the particular instance.

Student: (Refer Time: 34:18).

See like one thing is you have this Δ_i you can choose max over Δ_i and call it as $\Delta_{\max}$

Student: (Refer Time: 34:26) sir even if you let the supposed to be 0 1 or (Refer Time: 34:31).

That I agree. So, there when I say support to be [0,1] right?

Student: Yes.

That is not the property of the instant that is the property of my entire class.

Student: Suppose it is 0 1, then how do you (Refer Time: 34:42).

So, if I do this in this case irrespective of what is my problem instance this Δ_i are always going to be in [0,1].

Student: Yes.

That is why $\Delta_{\max}$ is going to be at most 1.

Student: Yes, assuming just (Refer Time: 34:54) k times 1, right?

In that case yes.

Student: So, there is no need of the entire analysis you already knew we already knew it was less than 1.

It is less than 1, right. I am only trying to do deal this part separately.

Student: but, sir mu star minus mu i t plus one obviously, be less than 1 equal to (Refer Time: 35:12).

that is fine I know this if the support is in the [0,1], I know this is already between [0,1].

Student: Yes.

But, that I am taking I am not making that assumption right initially.

Student: (Refer Time: 35:25).

I am saying the support can be an anywhere.

Student: Problem dependent all I agree, but I am asking what is that (Refer Time: 35:30) problem independent.

That is what I am saying like when you want to go for problem independent bound.

Student: Yes, sir.

Let us say that this class is such that the means are bounded.

Student: Ok [0,1] ok.

Whatever like [0,1].

Student: (Refer Time: 35:47).

Which is that is something, but whatever that value is it is bounded. In that case I know that this Δ_i are bounded.

Student: Yes.

So, then it is going that is going to give me some universal constant $\Delta_{\max}$ that is only depends on the on this class not on a particular instance.

Student: Yes, sir what I am saying is (Refer Time: 36:04).

Yeah.

Student: So, some of if you bound $\sqrt{1/T}$ by some $\Delta_{\max}$.

Ok.

Student: And, if you can say $\Delta_{\max}$.

Ok. So, you are saying that R max is any way upper bounded going to be $\Delta_{\max}$ if you are going to. So, that is fine. What I want to write it here is in terms of.

Student: And, sir we are (Refer Time: 36:27).

In terms of number of rounds, yeah?

Student: Yes. Then the bound we are getting is weaker than what we have already got initially.

Fine let us figure out that. So, what I am going to get if I upper bounded by $\Delta_{\max}$ whatever right. So, that is Δ_i / k into some part which is not at k.

Student: Yes.

So if that that is all k agree then it becomes in $\Delta_{\max}$ that is a trivial bound I have here.

Student: Yes.

But here I right now do not know that is the case.

Student: (Refer Time: 36:57).

Because so, that the other part is also there ok.

Student: Yes.

I think this needs a bit more refined argument. So, let us see if we can argue that bit later, but fine anyway. So, if we are going to like how the sub Gaussian noise and let us say we have some parameter which defines how large this delta i's are in terms of that then I can treat this as a constant. And, what I have is here a simple regret bound which is of the order log what is this log K divided by square root of this is what is this sorry this is floor of T by K ok.

Then what I want to write this as so, this is like order if I just do this manipulation which is going to be $o\left(2\sqrt{klogk/T}\right)$. I have just taken this k in the numerator. So, this is K log K divided by T if I am going to read this as a constant here in terms of the problem class dependent constants.

So, what we have done here is we have an upper bound on this simple regret where that is growing that is falling exponentially fast, but this problem is this bound is like problem dependent like it depends on the particular one. So, now if I make it get rid of the problem dependency by bounding its all the parameter, now I have a constant here where regret is like falling like the simple regret is falling like inversely in T.

Student: Sir, can we actually say.

But, this is any way constant right like I mean I am not changing this is not changing with time. I am only talking with respect to time how it behaves.

Student: Yes sir.

Only time. So, this is going to if you are going to fix whatever the problem instance you and then you are going to increase T what is this is saying is your simply regret is going to 0 exponentially fast. But, if you are going to treat it across all parameters irrespective of a particular instance, it is saying that there could be a worst case if you are going to treat where it can make you go like only like order like $\sqrt{1/T}$.