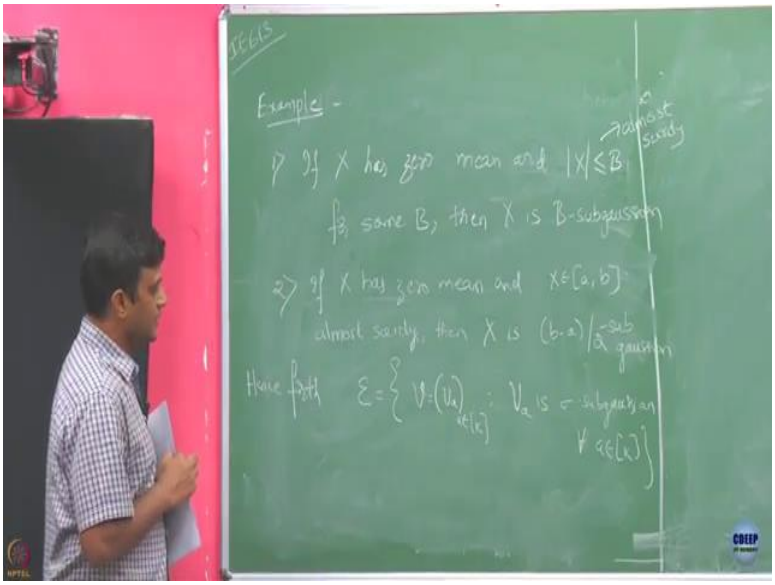
Bandit Algorithm (Online Machine Learning)
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture - 29
Regret Definition and Regret Decomposition

So, now at least one thing we are happy about is now we have a bound, which decays exponentially in n. Now, but this happens for sub Gaussian random variables that we need to have that property. Then the question is, what kind of random variable satisfies sub Gaussianity property?

One thing we already saw, if it is a Gaussian with mean zero that satisfies sub Gaussianity property.

(Refer Slide Time: 00:54)



So, here are the other ones, we need to verify all of them. If X has zero mean and then X is B sub gaussian and this X has zero mean and X is so, this is like almost surely.

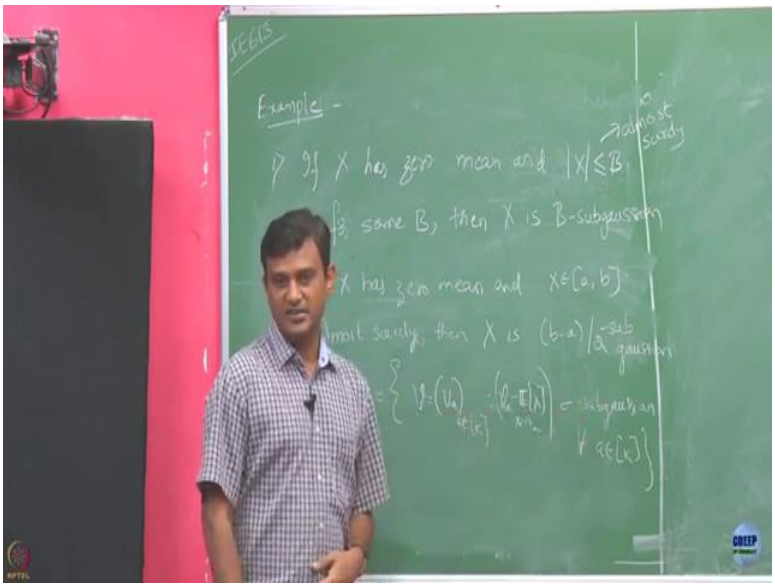
So, suppose you have a random variable X that has zero mean and its absolute value is bounded by B, if that happens, then that guy is going to be sub Gaussian with parameter B or other possibilities, if you have an X random variable that is zero mean, but it takes value in this bounded interval, some a and b if that is the case then also it is sub gaussian with a parameter b minus a by 2 ok. So, you see that whenever my random variable has

zero mean and bounded support, it is going to be sub gaussian with some appropriate parameter.

So, we will be henceforth, focusing on random variable, a distribution, that are just sub gaussians with some parameter. So, we really need not worry whether they are bounded like this or supported on some bounded interval anything like all this is fine we will just say that it is going to be subgaussian with some sigma.

So, we will say that henceforth; hence forth, according to our notation what we said? We are going to assume my distributions are going to come from a environment class in which, the distributions are sub gaussian with some parameter sigma square ok. We are going to say that my environment classes is υ which is υ_a such that; υ_a is sigma sub gaussian for all a.

(Refer Slide Time: 05:50)

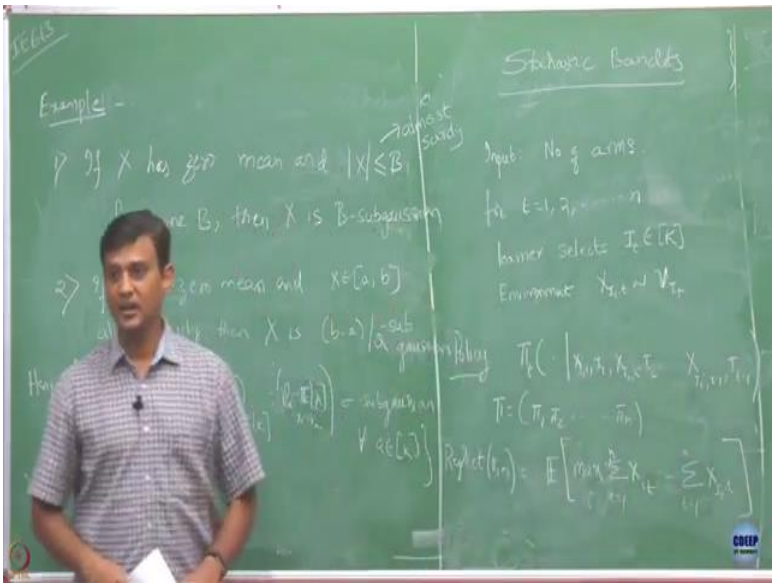


So, we are going to say that each of these arms are going to have a distribution that is, sub gaussian with some parameter sigma and that sigma we are we will start with saying that we know it, it is just that so, we are. So, if I am going to assume everything that is going to be zero right. So, if I am going to assume already all the distributions are sigma sub gaussian all of their means is going to be the same zero right. What we will assume is υ_a minus expected value of X which is drawn from υ_a this sigma sub gaussian.

So, what is that we are going to say that, these are the distributions assigned whatever the true means that is the mean, when I say expectation of X, where X is drawn according to distribution υ_a. So, this is nothing but, the mean value of this distribution υ_a. So, if you take subtract this value from the distributions right. So, distributions I mean so, if X is going to be drawn from that distribution and if you subtract mean from that. So, this going to be a zero mean random variable right and we are going to assume that is values are sigma sub Gaussian all of them.

I am just saying what is a good notation. Is this notation clear to you? We are just saying that, the distribution of the a term that random variable whatever associated if you subtract the mean value from that, this is going to be a zero mean quantity we are going to this is going to be sub gaussian with parameter sigma each of these arm will be. We do not know that what are the associated means with it, but if you just take this distribution with centered around that mean, then that is going to be sigma sub gaussian.

(Refer Slide Time: 07:36)



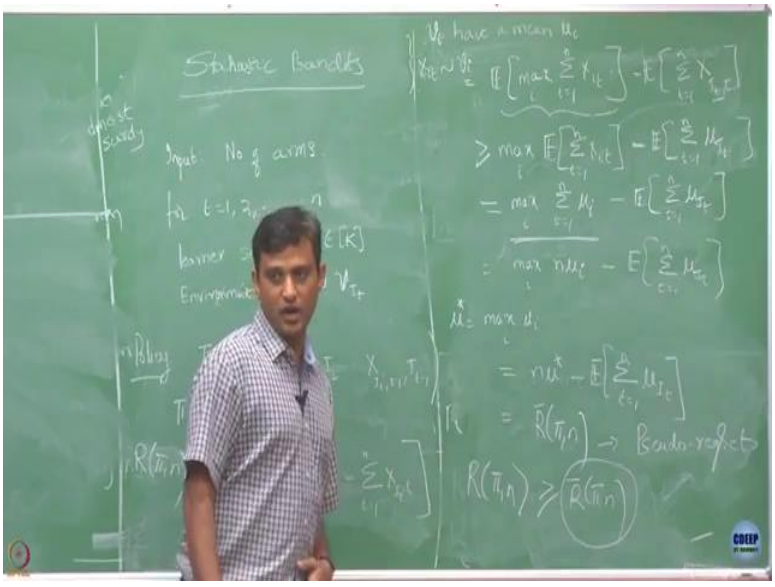
Now, let us return to our stochastic bandits. So, what would we say in our stochastic bandit? There is a learner, in each round he has to decide which arm to play and when he plays that arm, environment is going to generate a sample associated with distribution of that arm and learner get to observe that and learner's goal in this case is to quickly identify the arm which has the highest mean right ok.

So, let us rewrite the interactions we had. Our setup was like input is number of arms and for t equals to 1, 2 let us say n is known till that point, learner selects I_t from one of the arms and environments draws some random variable $X_{I\_t,t}$ this is drawn from the distribution υ_{I_t}; υ_{I_t} is the distribution associated with that arm I_t.

So, earlier we just wrote it X_{I_t}, first lets also append a t here, just to mention that this is the t-th round. So, this is how the interaction happens and based on these observations, every time the learner has to see what is the action that is going to see. So, we said that policy is in every round is based on their past observations in every round, this is based on the history the learner has to decide which is the action he is going to pull in round t and policy consists of all of this so, (Refer time: 10:59).

And we defined our regret to be so, we defined it bit so, this is our policy π over n rounds. So, last time we just defined this regret quantity to be let me define it more refined way. This is t equals to 1 to n rounds and this is like so, let us the so, what is this? If you play over t rounds and if you always happen to play the i-th arm, this is the cumulative reward you are going to get right and learner want to identify I mean you will be interested in taking your base to be the one, which would have given you the maximum reward and you want to compare it with the reward you get where you played I_t - th arm in round t ok.

(Refer Slide Time: 12:43)



So, this is same as saying expectation of max over i yeah I just flipped the expectation, but now this quantity here where I am trying to seek the maximum among the realization. So, these X_{I_t} 's are realizations right if you are done over t rounds this X_{I_t} for t ranging from 1 to n this is the realization you would have seen on the i-th arm.

We are looking at the maximum value of this, but in general as we argued in our adversarial case this point was maximization, sample wise maximization to deal with this is going to be hard. So, we will relax this and we will take the expectation inside. So, if you do the expectation inside, we are going to get a lower bound ok, but here so, ok.

Now, what is this expectation is over? This $X_{I\_t,t}$ are also random variables right and this I_t is also random variable because it depends on your past history. So, there are two randomness here, one is the randomness in pull pulling the arm even though the learner is not randomizing, he is not picking them according to some distribution, but even though he is doing deterministically, but that will be influenced by the past observation. So, because of that this I_t is has some randomness and this $X_{I\_t,t}$'s anyway they are drawn according to some distributions.

So, if you take into account the distribution of both, let us say first fix one sequence of this pulls if I fix this and if I pull only one expectation inside where that I will take the average of this samples. So, this is going to be t equals to 1 to n of $\mu_{I\_t,t}$ right because expected value of this sample is going to have this μ_{I_t} mean that is what we have said right each of these distributions have a mean and the ith arm. So, our thing is we said that, this distribution υ_i is going to have a mean of μ_i. So, this X_{I_t} is going to be drawn from υ_i. So, that is why this expected value of X_{I_t} is going to be μ_{I_t} here ok.

Now, what is this quantity? I can interchange the expectation here. So, this is going to be max over i and what is this? This is now nothing but, summation of t equals to 1 to n this is going to be what? μ_i right expectation of X_{I_t} simply μ_i here and this quantity here.

Now, what I am looking is, the maximum values right of the means over all things. So, suppose let us say and this is nothing but, max over i of n times μ_i and the other part remains the same. So, suppose define μ^* to be max value of μ_i. So, than this is going to be n times μ^* minus expected value of summation t equals to 1 to n of μ_{I_t}.

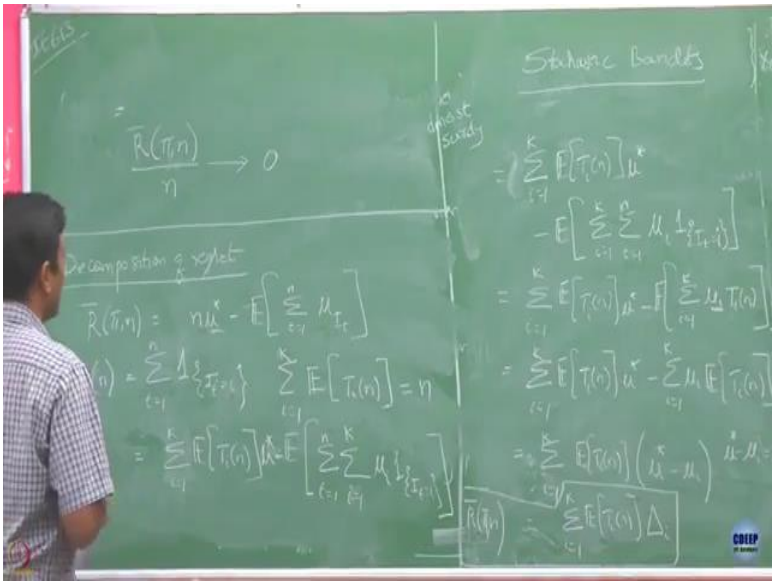
Now, what its expectation is about is expectation is on this I_t because I_t itself can be random depending on their past history, that past history is going to induce some stochasticity in I_t. So, this is what we are going to call it as. So, let me call this as simply $R(\pi,n)$ and I am going to basically I do not know how we call this one and this is the earlier did we denote the pseudo regret by R bar or the original regret by R bar?

Student: Pseudo by R.

Pseudo by R bar so, and now we are going to take this as pseudo regret and we will be interested in this regret bounding this regret. I do not know we if we discussed this during the adverse suppose so, we have a lower bound on this actual regret right, pseudo regret is a lower bound on that. So, if I give you an upper bound on this regret pseudo regret, will it be an upper bound on this actual regret? So, this quantity so, what we are finally, showed is $R(\pi,n)$ is up (Refer time: 19:20). So, this is the one we said henceforth we are going to focus on.

And when what we will usually do? We will develop algorithm which is going to give an upper bound on this. If I have an upper bound on this pseudo regret, will it be an upper bound on this actual regret no right because this is a lower bound itself, if we give an upper bound on this cannot be upper bound on this, but we will ignore that fact. And hence, we will only try to study this and see how we can upper bound is regret bound ok.

(Refer Slide Time: 20:11)



And as usual, our interest is to find policies π that makes this environment class learnable. What does that mean? We are interested in policies π such that so, we will interested in π such that. If you divide it by n it goes to 0 that is, we are interested in policies which gives us sub linear regret fine.

So, now, before started looking into what are the policies that gives me the sub linear regret. So, we will have one more property of this regret, which we call it as a decomposition yeah I mean decomposition of the regret let us say decomposition of the regret. So, henceforth like I will not keep calling pseudo regret, I will just say regret, but what we mean is actually pseudo regret we will be interested in pseudo regret ok. So, what is this we have or a policy pi we have nμ* minus expected value of t equals to 1 to n μ_{I_t}. So, here mu I t is the one that depends on your policy π.

So, now let us try to simplify this. So, this I will simply again rewrite this as t equals to 1 to ok. So, before this I am going to define now total number of pulls of an arm ok. So, you are playing this learner in each round he is going to play different different arms right. So, when after n rounds, he would have pulled each one of them different number of times. So, let us define number of pulls of arm i over n rounds to be equals to summation of t equals to 1 to n indicator right.

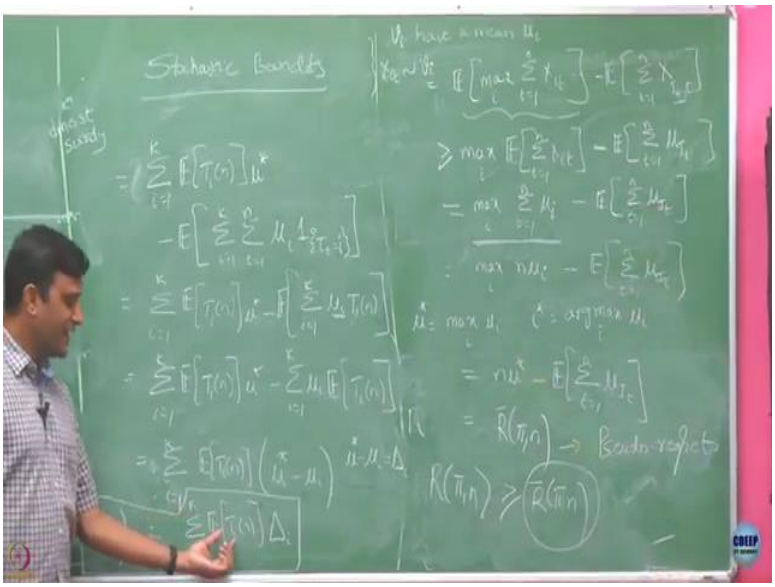
So, if in round t, if he happens to pull i then, it will get counted in the summation all we are saying is how many times are i is pulled over n number of rounds is this correct ok. So, now, using this, let us rewrite this quantity here this n mu star is simply t equals to n sorry yeah before this if I take summation of expected value of this T_i (n). So, tell me is this T_i (n) is a random variable? Yes right it depends on how many times you have pulled and it is going to influence how many times you are going to pulled.

So, let us take an expectation of this and sum it over all i equals to 1 to n, then is it equals to n right. So, total number of pulls either you would have pulled i or something else. So, if you would have pulled one of the k arms in each round right so, if you are going to take the sum of all the pulls, it has to be equals to n. So, this is the first thing I will do in this equation because this n I am going to sorry is it n or k here it should be k right, the sum of the pulls of a arm that should be equals to k.

So, this n I will replace by this quantity this one in this expectation I am going to add another summation. So, here this μ_{I_t}, I have written in this format. By doing so, did I

change anything from here to here. So, let us look what happens. So, this I_t has to take some value right. So, what I am doing is now looking at all the policies of i, i to 1 to n and I am saying what is that value I_t takes. If I_t takes that particular value, I will have an indicator here only that μ_i is returned everything will vanish right. So, because of that in this only that i corresponding to I_t will remain. So, that is the this term μ_{I_t}. So, these two terms are the same.

(Refer Slide Time: 26:18)



Now, let us do one thing. What I will do here is, these are both finite summations right, I will interchange the summation and also this is a finite sum so, I can also take expectation inside here, what I will do is, right there should be a μ^* here this μ^*. So, first thing I will do is, in this expectation I will interchange the first inner one and then, this is like μ_i I_t equals to 1.

So, now, this part here in the internal sum here what we are basically doing is, we are counting as I run from t equals to 1 to n basically it is going to be this indicator is going to be 1 only when I_t is equals to i right in this internal summation. So, then only when I_t equals to i, this sum this μ_i returns otherwise it is going to be 0 so, can I then say this, this is nothing but, number of times I have played arm i times μ_i. So, then this is like so, this is nothing but, μ_i times number of times I have played arm i and anyway this μ_i 's mean values are constants for me in this summation I am going to take it out and then, I will further simplify this.

And now, I know that μ^* is the maximum value of all this μ_i 's right that is how I have defined my μ^*. So, μ^* minus μ_i has to be positive. Now, what I will do is, I will define this difference as μ^* minus μ_i to be some delta i. So, what is this is telling you, what is the gap between the best arm and the ith arm? So, if I do that, now this is going to be and this is going to be called as your regret decomposition formula.

So, does this make sense like this formula, what are we are saying? This is saying that, if you have played your arm n times in that you have played arm i T_i (n) number of rounds suppose this i arm happens to be the optimal one then, if i is the optimal one this delta is going to be 0. So, it is not contributing anything to you, but if this i happens to be something other than the optimal, then in the regret it is always going to contribute an amount of delta i right.

So, what we are basically saying is this regret is nothing but, number of plays of the sub optimal arms and times the regret they are going to incur and this is taken sum over all sub optimal arms. So, we have even though then taken it some over all arms, but for the optimal arm this term is going to be 0.

So, now, ideally suppose let us say in this let me denote i* to be the arg max of μ_i that is i* is my optimal arm. So, what you expect whether T_{i^*} you want to be much higher than the other ones or is it they can be all equal if our algorithm your policy is good.

So, let us say you are playing it n number of rounds, for arm number 1 you have played T_1 number of rounds, arm number 2 you have played T_2 number of rounds and the optimal arm i* T_{i^*} you have played some number of rounds. If your algorithm is good, you expect the number of times this T_{i^*} has been played should be much much higher than the other ones right that is the number of play of the sub optimal arm should be much much lesser than compared to the number of plays of your optimal arm.

So, suppose if we can bound and expected number of plays of this arms depends on your policy π. Now finally, suppose if for a given policy, if you can bound how many times you are going to pull the suboptimal arms compared to the optimal arms, then you can come up with a bound on this.

So, all their regret analysis that we are going to do exactly going to exploit acquired this line of thought. Once you have given the policy, they will try to identify how many times

you have played an optimal arm and how many times you have played a suboptimal arm, they will find a come up with a bound on that and based on that you will end up giving a bound on this. So, that is the line of attack we will take to give the regret bound.

Now, so, before we start with any algorithms, you have any idea about how to go and minimize, how to get a small value of this regret anything new you can think of. So, this all boils down to how quickly you can identify correctly estimate the means right. So, if you somehow get a good estimate of the means, then you know already what to do. You have to just pick the arm with the highest mean right.

Now, how you are quickly going to narrow down on the arm which has the highest mean? So, here the thing is how quickly you can end up identifying the arm with the highest mean? It is not even necessary that every arms you estimate the value correctly, it is about just identifying who is the best among this, we do not know how good he is compared to the others, but all we need to know is this is best arm compared to the others. If we can do this, then we are done right. So, how you are going to do that? Ok fine. So, will discuss different algorithms in the coming classes.