Bandit Algorithm (Online Machine Learning)

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

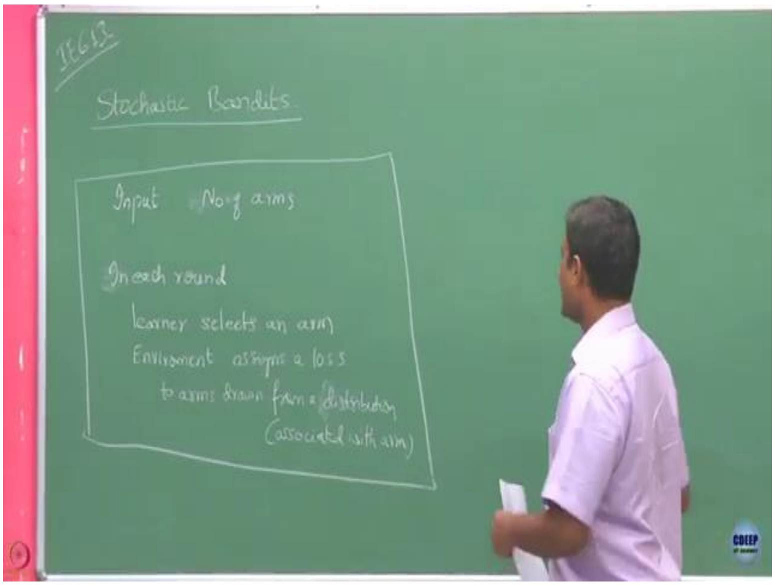
Lecture - 26
Introduction to Stochastic Bandits

So, far we have been dealing with learning problems in which we did not make any kind of specific assumption on how the environment is generating your loses right. It could have it could be even adversarial like the environment could be just trying to it make you incur as much as loss as possible. So, in that we define the notion of regret and try to see develop algorithm which minimizes this regret.

Now, we are going to switch to the stochastic setting where we are going to a priori assume that the loss that the environment assigns to an arm or an action is drawn from a distribution. So, it is always going to draw the losses from a fixed distribution. And now the goal is to identify which of the arms gives with the smallest mean in expectation ok.

So, what we are doing is we are now kind of restricting ourselves in the sense the loses are now coming from a fixed distribution, that distribution is not going to change once the game has started ok. So, because of that we are saying that the environment is following a fixed rule that is a fixed distribution according to which it is generating the loses.

(Refer Slide Time: 02:11)



IE613

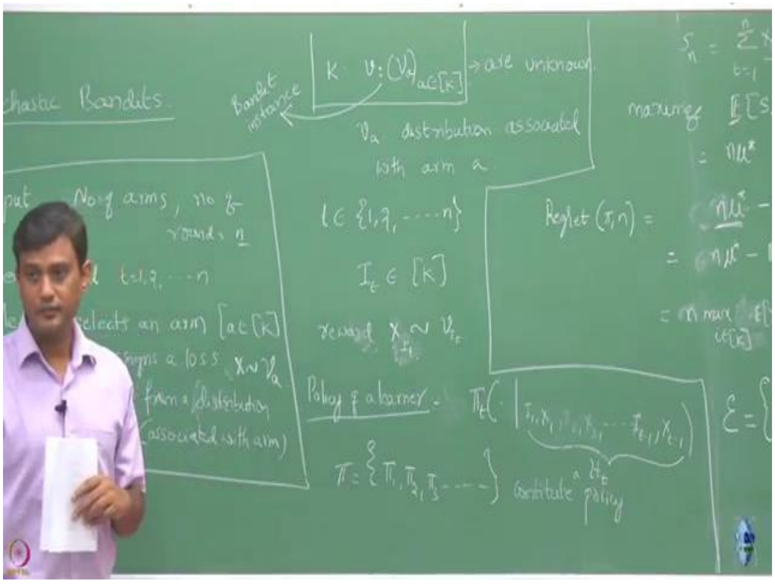
Stochastic Bandits

Input No. of arms

In each round
learner selects an arm
Environment assigns a loss
to arms drawn from a distribution
(associated with arm)

Now, how the interaction happens in the stochastic bandit? Here we are going to say that as usual there is the template is like input. The input you are going to say how many arms are there. And in each round, learner selects an arm and environment distribution, this is going to be associated with arm ok.

(Refer Slide Time: 04:16)



hastic Bandits
Bandit instance
$K \quad \upsilon : (\upsilon_a)_{a \in [K]}$ → are unknown
υ_a distribution associated with arm a
put No of arms, no of rounds n
t=1,2,...n
selects an arm $[a \in [K]]$
signs a loss $X \sim \upsilon_a$
from a distribution (associated with arm)
$t \in \{1, 2, \ldots n\}$
$I_t \in [K]$
reward $X_{I_t} \sim \upsilon_{I_t}$
Policy of a learner
$\pi_t(\cdot \mid I_1, X_1, I_2, X_2, \ldots I_{t-1}, X_{t-1})$
H_t
$\pi = \{\pi_1, \pi_2, \pi_3 \ldots\}$ constitute a policy
Regret (π, n) =
$S_n = \sum_{t=1}^{n} X$
maximize $\mathbb{E}[S$
$= n\mu^*$
$n\mu^* -$
$= n\mu^* -$
$= n \max_{i \in [K]} \mathbb{E}[$
$\mathcal{E} = \{$

So, let say there are k arms. And I am going to denote υ_a. What is υ_a here? υ_a is the distribution associated with arm a ok. Now, what we are going to say, learner select an arm let say from a belonging to k, and then environment assigns a loss which is going to be drawn from which is going to be an loss x which is drawn from your υ_a the associated distribution with arm a ok.

Now, this game is going to be played in every round right. So, let say for any round, and here n is the number of rounds, this may be a priori specified here. If it is a priori specified, the number of rounds we are going to play it that many rounds; or it may not be specified a priori how many rounds you are playing it will be you keep on continuing it ok.

Now, again we are going to denote this I_t to be the action selected by arm a learner in round t, this is going to be one of this t. And then X_t let say denote this X_t is the reward, so or loss. So, again when I am talking about now this bandit setting, I am now going to switch from loss setting to the reward setting ok.

So, now, I am going to assume that when a player plays an action, what he is getting is a reward, and now his goal will be to accumulate maximum number of rewards ok. And he is going to reward, he is going to get X_t which is now drawn for υ_{I_t}. So, I_t is the action you selected in round t. And this is the you get an reward which is the drawn from this distribution v_{I_t} here fine.

Now, in this stochastic setting, the setting for the environment is simple right. He has simply the environment already decided; the distribution for each of this arm that is fixed. And every time if you select an arm, it is going to draw one sample from that distribution, and it is going to reveal to you ok.

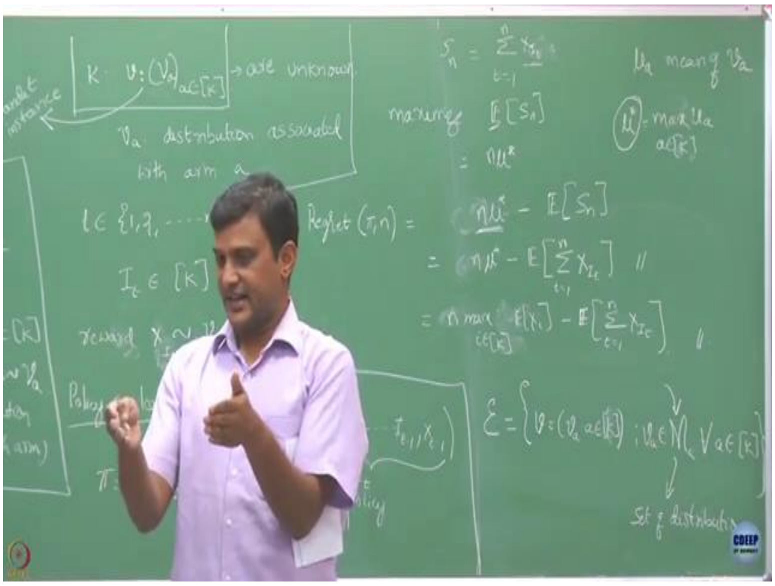
Now, when though I said here this is the action selected by the learner in round t, now depending on what is the action that is going to select his performance is going to change right. So, this selection of action in round each round, we are going to call it as the policy of the learner. And is policy of the learner will depend on what is the actions and the corresponded rewards he has seen so far ok.

And now this I am going to denote it as π_t, which is so in round t you would have observed what he had all these history, and I am going to denote it has its history time till time t. What the history includes, what I played in round 1, what is the reward I observed; what are played in round 2 what is the reward observed, all the way till t minus 1 round. I have this information. Based on that, he has to decide what is the action he is going to play in the next round ok. So, this is like I_t will be coming from this. And this is in the round t.

And now the sequence of π_1, $\pi_{2,}$ $\pi_{3,..}$ whatever you are doing to do, this pi this will constitute policy. So, basically these are rules. You are going to say now do not need to. So, these are the rules like if you tell me what I have observed so far, then it is going to tell you what is the action you are going to play in every round. And this collection of all these rules in every round is going to be your policy ok.

Now, what is the objective here? As I said your objective is to accumulate as many as much reward as possible, but you do not know these distributions, you do not know a priori what is the distribution that each arm is associated with ok.

(Refer Slide Time: 10:58)



Regret (π, n) =
$S_n = \sum_{t=1}^{n} X_{I_t}$
maximize $\mathbb{E}[S_n]$
μ_a mean of ν_a
$\mu^* = \max_{a \in [K]} \mu_a$
$\mathcal{E} = \{\nu = (\nu_a\ a \in [K]) : \nu_a \in \mathcal{M}_a\ \forall a \in [K]\}$
set of distributions

So, let say if you run it for n rounds, your total accumulated reward is ok. And my goal is to maximize expected value of S_n , n p r ok. And you have to do it without knowing this distribution. So, can anybody say what could be the maximum value of this expected value of S_n? Suppose, I tell you a priori these are the distributions associated with each of these actions ok.

And let me say also let me denote as μ_a to be mean of a distribution a ok. So, there are this distribution, the distribution associated υ_a is μ_a. If I have to maximize this, what is the maximum value I can get? Max of?

Student: (Refer Time: 12:31).

Max of mean let say I have this right like I have this μ_a, and there are, and let me call this as some μ^* ok. So, in that case, what will be this quantity will be, it is going to be n times μ^* right, because you are running it n times, and in each round what will be going to play a do is you play the arm which has the highest mean, and you are going to get it as $n\mu^*$. So, this is what I am going to get in expectation when I know the distributions of all the arms.

But we do not know them a priori that has been selected by the environment, but that has been not told to me. And now I want to, goal is to achieve this much of expectation in reward. So, I am going to set this as my benchmark when I am going to define regret of a my policy. So, here I am going to define regret of a policy pi over n round as what S_n minus, so this is this $n\mu*$ minus expected value of S_n. In the policy have said here I am just saying that this could be deterministic, there is no randomization by the learner here.

So, based on this history here is he should tell him this history, he will find out based on this quantities tell what is the arm he have to play in the round t ok. So, he, we are not allowing him to do any randomization here ok. In this, it is just like claimed deterministic given in the past. What is X_t is the reward you are going to observe in round t if you happen to select action I_t will be drawn by υ_{I_t} is distribution is υ_{I_t} in round t.

So, in round t, we are playing some action right let say you have played action I_t, then its distribution of this reward will be like this. So, because of this, this X_t is a random quantity here the reward you are observing in each round is the random quantity. And now when you define this S n here like this, this is the random quantity because this X_t's are random here.

Now, what you are interested in this is the expected reward, total reward in this case ok. And now what I am saying is this is the best you could get if you happen to know all the distributions. How you achieved this? We achieved this by playing the arm with the highest mean in each round, and this is what you are going to get based on whatever you are.

Now, I am going to slightly denote this read X_t, I am going to append to this X_t here, this I_t also like this is to just say that the reward I am going to observe in round t this is associated with arm I_t which is random quantity ok.

So, now, when I did this expectation here, this expectation is over what, there are two random quantities here right. Even though you are if I give you this history, what I am going to select in round t could be deterministic, but this sequence itself is random right because this x 1, x 2 these are random quantities here the sample you have observed. Because of that this will induce as a randomness in the choice of my arm.

So, because of this, this I_t's are random variables. The they depend on what is the samples you have observed in the past ok. So, because of that, this expectation here will consist of will involve two randomness; one is the randomness in the samples itself, other one in the choice of your actions ok, so even though the samples have induced the randomness in the choice of action I_t ok.

So, I am just me look what is the right notation for this once I have this. So, just let me continue to use this notation may be I will just write this in round t the observed this is the reward you are going to observed may be I will just write it like this t I_t is going to be this or t I_t means looks confusion let us just use this. The you have played action in I_t in round t, and you are associating a random variable with this is going to be drawn like this.

So, when if I do this then this is like. So, this is not eta, this is just nμ* minus this is not expectation of what is this X_{I_t} equals to 1 to a n number of arms. And I am going to call this as regret. And how is the policy influencing this regret? The effect of the policy pi comes on the choice of I_t's here. If you are going to change this pi the way you are going to choose your I_t's is going to define will different and that is going to affect this regret part ok.

So, what we are doing here? In this case we are our regret base is still the same as we did it in the adversarial case. What is we are saying? This is the regret incurred by playing your policy; this is the regret maximum regret you got. So, this maximum regret is basically seen that this is nothing but what like I am if I just this is like max this is n times max i summation of this expectation of X_i's right this is like i equals to 1 to, so i this is i coming from k.

So, I am still looking at what is the mean reward I can get if I have to play a single arm throughout n rounds. So, if I am going to play the arm which gives me the highest mean for n rounds is the total reward I am going to get right. And I am going to compare it with whatever I am going to get by playing my policy.

So, in that way this what we are basically doing is I am comparing whatever I got by comparing it with an arm which I if I play it throughout all the n round that would have give me the best possible rewards. So, here my benchmark is still playing a single arm, but that arm is now the one which gives me the highest reward fine.

Now, we say that these are unknown, these distributions. Now, question is can this be any distributions that will allow, or we allow only some specific set of distributions on the arms, or is that is there some special structure we are going to assume on this distributions?

And that is going to kind of define the environment class we are going to look at. Right now we have said that environment is just a priori choosing these distributions and fixing it, that is not revealed to me and my goal is to identify the one with highest mean ok.

Now, we are going to say that this environment class itself is going to define like this. I have going to denote this to be an environment. So, let say an arm this number of k s are fixed the number of arms are fixed. Now, this is the set of distributions on this arms right. We are going to say that the set of distributions once you fix it that is going to define your bandit instance ok. Now, what I just said is we can assume that this bandit instance are coming from some special class ok, or we can say that these bandit instances are drawn from some environment.

So, let us define this. Let say so we said that one of this new defines a bandit instance right. Is a notion of bandit instance clear to all of you? So, we are saying that there are k arms right, you decide, let say environment decided one particular set of distributions arm 1 is going to take this distribution, arm 2 is going to take another distribution, arm 3 is going to take another this distribution.

And it has for each arm it has given a distribution that will define me one bandit instance ok. So, may be later tomorrow the environment assigns a different set of distribution to each of this arms, now that will make an another bandit instance ok.

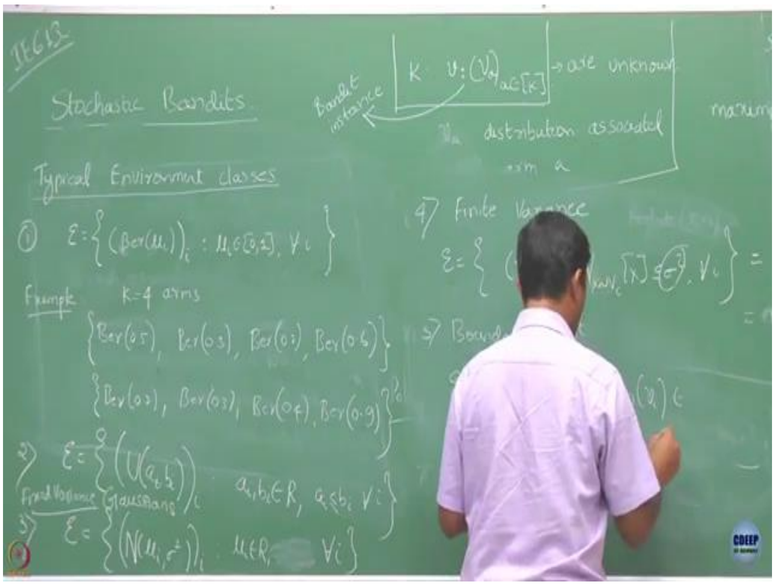
Now, what this we have defined is irrespective of what is the bandit instance the learner is facing we are going to define this to be his regret ok. And notice that this maximum value here the maximum mean it depends on a bandit instance. Once you fix this distributions their associated means are fixed now for that there an associated maximum value, and we are defining that regret in terms of this. If you are going to change your bandit instance, your maximum value could be change different ok. What?

Student: (Refer Time: 26:08).

So, I am saying that this is some class, and we are going to this is like some set of distributions. So, the environment can say that I am going to assign a distribution to this arm, but the set of the distributions and can assign to this arm there will be coming from some set. So, this distribution itself, this M$_a$ is basically the set of distributions here. Let say environment has some bunch of distributions for one particular arm, and it has another bunch of distribution for another arm like this.

So, each time it can pick one distribution from this set and assign it to some arm. And from the second set it can take one distribution and assign it to second arm like that and depend. So, that is why I am saying this is basically all possible collection of bandit instances, and that will define my environment class. At any time the environment can pick one bandit instance from this and I will be faced to learn against it ok.

(Refer Slide Time: 27:52)


IE613
Stochastic Bandits
Bandit instance
K
$\nu : (\nu_a)_{a \in [K]}$
→ are unknown
ν_a distribution associated arm a
Typical Environment classes
① $\mathcal{E} = \{ (Ber(\mu_i))_i : \mu_i \in [0,1], \forall i \}$
Example K=4 arms
{Ber(0.5), Ber(0.3), Ber(0.2), Ber(0.6)}
{Ber(0.2), Ber(0.3), Ber(0.4), Ber(0.9)}
2) $\mathcal{E} = \{ (U(a_i, b_i))_i : a_i, b_i \in R, a_i \leq b_i \forall i \}$
Fixed Variance Gaussians
3) $\mathcal{E} = \{ (N(\mu_i, \sigma^2))_i : \mu_i \in R, \forall i \}$
4) Finite Variance
5) Bound
maxim
CDEEP

Now, what are the typical environment classes? So, now, I am going to define you what could be this different environment classes ok. So, for example, one possibility could be, just an example the environment assigns Bernoulli distribution to each of this arms with different parameters. For example, how you write Bernoulli? So, this is one what environment can do is for each of the arm for each arm it can pull a value which is from between 0, 1, and then associate Bernoulli with parameter μ_i to the ith arm ok.

So, for example, let say one, one example could be, if you let say you have let say k equals to 4 arms. The first arm could take Bernoulli with parameter 0.3, 0.2, and 0.2, and Bernoulli let say 0.6. So, this in this case this will make one bandit instance. So, in this bandit instance, what is the maximum value or what is the best arm the one and this one because this has the highest mean right, the mean here is 0.6.

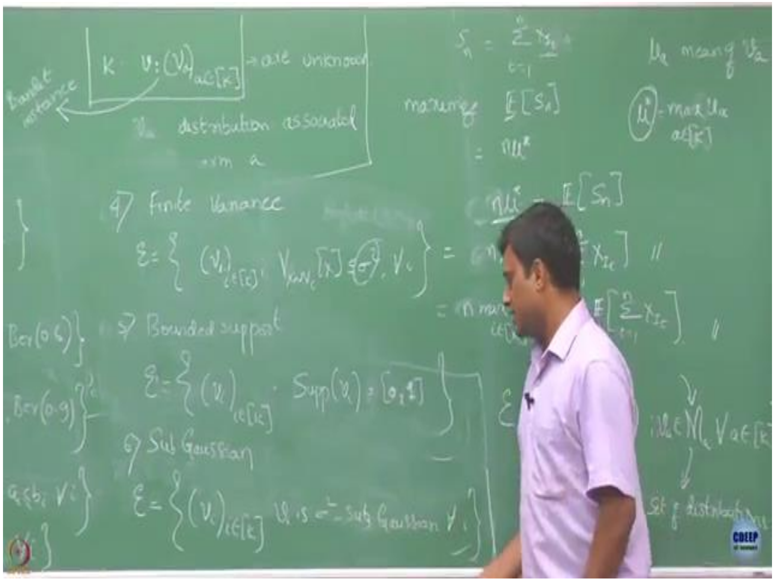
And another bandit instance could be like just like now instead of taking 0.5, it could take let say any value in the, so it could be just 0, 2, and it could be like it can be same something like 0.4, and Bernoulli 0.9 something like that. So, that is what I am saying this here the environment class is like this ok. Other environment class you can think of is where to each arm, the environment selects a distribution which is uniform ok. And now uniform distribution is defined by what parameters?

Student: a.

a where lower limit and the upper limit the range. And depending on the range, it can come up with different, different uniform distributions. So, that I can write it as and here this a_i and b_i could be some real numbers for, and we will assume a_i is less than b_i for all i. So, for each i, it is a uniform distribution and this parameters a_i , b_i can be drawn from real numbers.

Similarly, other examples could be all this distributions could be Gaussian with mean μ_i and variance sigma square. So, here again μ_i can be real, and sigma square can real, and this is for all i. You can think of like different environment classes like this.

(Refer Slide Time: 32:55)



Bandit instance
K $\nu : (\nu_a)_{a \in [K]}$ → are unknown
ν_a distribution associated arm a
$S_n = \sum_{t=1}^{n} X_{I_t}$
maximize $E[S_n]$
μ_a mean of ν_a
μ^* - max μ_a $a \in [K]$
4) finite variance
$\mathcal{E} = \{ (\nu_i)_{i \in [K]} : V_{X \sim \nu_i}[X] \leq \sigma^2, \forall i \}$
Ber(0.6)
Ber(0.9)
$a_i \leq b_i \; \forall i$
5) Bounded support
$\mathcal{E} = \{ (\nu_i)_{i \in [K]} : Supp(\nu_i) = [0, 1] \}$
6) Sub Gaussian
$\mathcal{E} = \{ (\nu_i)_{i \in [K]} \; \nu_i \text{ is } \sigma^2\text{-Sub Gaussian } \forall i \}$
$\mu_a \in \mathcal{M}_a \; \forall a \in [K]$
Set of distributions

So, another to be another finite another environment classes like something like finite variance. What here environment looks like is it is like all μ_i's. So, here now I am writing it has i such that the variance of a random variable which is drawn from v_i in this less than sigma square for all i.

So, what does this says, it says that the distribution assigned to each arm could be anything, the only restriction I am putting is its variance, if you draw the variance associated with the distribution should be finite sigma square, and this we assume that will be a priori given to you.

So, this could define another class of another class of environments right. Now, is this environment class contains this environment class here which is the Gaussian distributions? Then what could be sigma? So, here we are saying this sigma i's are just real sigma i squares are some positive numbers right.

But here we are saying that this is some bound we are giving, so if you are going to take variance of the distribution of the ith arm it is going to be sigma i square. How do you know that is going to be less than this given sigma square?

Student: (Refer Time: 35:06).

How many take like this could be any positive number, this sigma i square can come from any positive number right. Suppose, I relax this environment and I say that all of them have the same variance, so then I do not need to write this. So, this is the one with fixed variance. So, what I done is I have just taken a Gaussian distributions for all the arms, and I am saying that all of them can have different means, but all of their variance is going to be the same sigma square.

Sigma square is fixed that is know only thing is their mean values I do not know. Does this fixed variance Gaussian distribution will be a subset of this class right? I do not say anything about the mean, mean can be anything here right. And only thing I am allowing is all these variances are just request to sigma square. So, here this environment. So, here this is finite variance ok, finite variance this is Gaussians.

Another thing could be simply the bounded support ones. So, here the environment could be simply set of all distributions such that the support of v_i belongs to this some subset of a,b. So, all of you understand what I mean by support of a distribution, what is support of a distribution? Probability is not 0 ok. So, this bounded support one will it contain this Bernoulli environment here. How?

So, suppose, if I define this a, b to be let say 0, 1, then will the support of this Bernoulli distribution will lie in the interval 0, 1 right, because the Bernoulli what are the values either 0 and 1 right that will be always, but it is only thing is now in this case I am when I look at this guy I am looking at all possible values of this ok, or support of this for all of them let us take this instead of subset. In that case, this is going to can be containing this environment class already.

So, another could be like my all my distributions could be such that υ_i is sigma square sub Gaussian. So, how many of you know what is the sub Gaussian random variable? How many of you know sub Gaussian random variable? We will define it a bit later, but this will like more generalization of a Gaussian random variable.

So, if you are going to say it is a sigma square sub Gaussian, we will see that its variance is already going to be less than sigma square. So, because of that this guy will already incorporate this, this guy already going to include this family of distributions with finite variance, then I take this sigma square to be the same as this sigma square.

So, as you see the environment class can be anything but we may a priori restrict ok, I know environment can choose this distributions to the on the arm, but those distribution we are going to be drawn from this environment class like that. So, most of the cases what we will be focusing in our discussion is something like this the one with the sub Gaussian.

By the way this as I said is sub here though here I am saying the support is bounded here is the support bounded for the Gaussian random variable, it is not right. Even here for a sub Gaussian it is not bounded, but we will see that later as we go on even though its support is not bounded for the sub Gaussian, but its analysis is not much different from when we are going to deal with the family where support is bounded ok.