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

Lecture – 42
Stochastic Contextual Bandits

So, during the first half we saw all bandit problems right, like we started with the full information setting in the adversarial case, then we graduated to the case where we have only bandit information and then we consider a case where it is not adversarial, but your environment is stochastic.

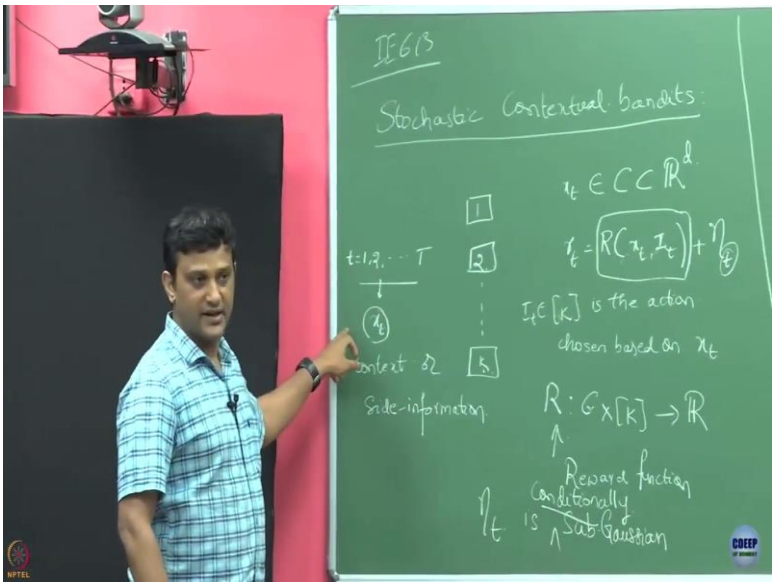
And, we looked into different algorithm and at the end we showed that what is the minimax regret one can get; on all these instances and what our algorithms actually got we showed that all our algorithms that we discussed in the class they were almost minimax optimal and one algorithm was exactly minimax optimal. So, others who were half by a factor square root log t sometimes.

So, yeah before I start this the next part, so, in the last we ended the first half by giving the minimax regret, but there is another notion of called problem dependent lower bounds assume today which we did not do, but there is a separate chapter on the book you can look into that.

So, remember like we gave 2 kinds of bounds on each algorithms right, one is problem dependent and problem independent bounds. And, for a problem independent case we looked at the minimax lower bound and showed that whatever the problem independent bounds we are getting they are matching the minimax lower bound up to the log factors.

But, we gave problem dependent bounds upper bounds, but we do not know whether they are optimal right, what is the lower bounds, problem dependent lower bounds. So, there is again what are we have they are almost optimal, we will not go and prove it a lower bounds problem dependent lower bounds.

(Refer Slide Time: 02:27)



So, today we are going to start this Stochastic Contextual Bandits. So, if you recall the set of we studied so far, what was my interest? I have a set of actions, each action gave me some rewards stochastic reward, but it has it is own mean value, all my interest was to find out the action which has the best mean, right.

And, we posed that problem as regret minimization problem that, if I how best I could what is the total reward I am going to accumulate in expectation over a period of time compared to the case where I knew exactly which is the best arm, that is which is the arm which has the highest mean.

So, the goal there was to find the action with the highest mean or the best action there. So, often it so happens that the action, the best action for you in a particular round depends on that round itself, something that as related to that round itself. It is not that you are interested in one best action over all rounds, but it may happen that the best action could be different at each rounds.

So, what could be the example? For example, suppose nowadays you might be doing lot of shopping over e-commerce sites and all right, like you know when you just log in you already see that the things which you like, but you have never told to the website they already started popping up on your screen.

So, what these guys are doing? They are basically trying to recommending items which you are going to like. So, they what is the action set for them? For them, the set of actions is let us say, either advertisements or the products that they want to sell.

Everybody, will be users will be log in to that website, let us say whenever I new user logs into that website that is like a new time that a will be looking at ok. And now, naturally that whenever a new users logs in that time that websites like to show a product or an advertisement that is most likely clicked by that user right.

And, for each user the likings could be different. So, if a new user logging into the system is a different events or the times, then optimal action at every time is potentially different right that depends on that user itself who is logging in at that time. And, the website where you have logged in may have some information about you. For example, you when you registered onto the website you might have told about your date of birth, your sex and maybe some the region where you reside and all these things.

So, these are the information that the website has and maybe this information is useful for the website to tell what is the best advertisement or the best product to recommend for you ok. So, now the question is fine this is a problem; I have let us (Refer Time: 06:29) stick to the case that, I have a set of advertisements, I want to show it to users and my goal is to gets maximum clicks on the advertisements.

So, whenever a advertisement you have shown if you click on this, that websites make money out of it right. So, it wants more clicks to happen on the advertisements. So, it is going to get more clicks, if the advertisement it shows you, you likely to click on that. Now, it over a period of time it want to maximize the number of clicks you are it want to get.

Now, do you think I can pose this problem as a bandit problem which in a setting which I already know from the what we did in the first half, like either in a stochastic bandit setting or in the adversarial bandit setting? If I want to do this what how can I do it?

Student: (Refer Time: 07:29).

What?

Student: Mean could vary (Refer Time: 07:32).

The mean, what is the mean here?

Student: I mean the reward, the.

So, reward here is the probability of click. If I use a clicks; that means, with that probability he got some money ok.

Student: (Refer Time: 07:46).

Now, can it, if I have to use my earlier setting right, what I will eventually end up finding up? I will end up finding up single best advertisement that I want to show to everybody.

Because, that earlier band earlier set up only cares about one best action and here my actions are set of advertisements. It cared about showing one single best action among all that would have got more number of clicks. But, here if you make and it will try to always, it try to find that single best advertisement which gets maximum number of clicks.

But, now if you customized your advertisement, based on the user information, the user who has logged with them do you expect to get a better clicks than always showing a single best advertisement across all users or you want to show an advertisement which is most likely to clicked by each individual users? You are going to make it personalized or you want to choose a one which is globally I mean optimal across all?

Student: Personalized.

You want to make it personalized right. The question is did whether the earlier setting allowed for that, earlier setting we had did it allowed for it? It did not allow for it right. Now, the question is how to make use that how to incorporate in this possibility in that setting. Yeah.

Student: You can (Refer Time: 09:31) algorithm for every user separately.

Ok fine. You want to run every user separately, but user 1 comes, he does click on or not click on something then he may vanish right he may not again come.

Student: Ha sir.

But from that did you learn anything about and there are so many such users right and you cannot design one algorithm for every user.

So, you see that right you could do that potentially like you could treat that every user as some bandit instance and I on which I want to learn, what is the best for him. But, that guy's short time like he just comes and leaves. So, in that way like I cannot directly use that standard bandit setting right. What could be the other issues, if you want to use that world setting?

So, one case is already we discussed like this number of users could be many many and I cannot run separate algorithm for each and other things whatever I have learned from one user, let us say whatever time he was in the system whatever little I learned from it, can I transfer that knowledge to other user? The earlier setting also did not have that facility right.

So, now you are going to study this version of stochastic bandits called stochastic contextual bandits where, depending on the context the best action can change and what is the context? The context is whatever the information I have about an instance at that time.

So, let us say I am going to just denote my set of actions. So, my let us say the actions are this and I am let us say our time period. So, think of these times as instance happening for example, somebody logs into your system. Now, what the earlier setup did is? It try to see these are actions.

It tries to find which is the single best advertisement I should be showing to each one of them, but, now let us say now I have some contextual information for every user at time t, let us say I have some contextual x_t that I could observe. For example, as I said when you log in the website already knows through your profile what is your age category possibly what is the shopping you have done earlier all this information.

And on observing this, now you have to find which is the best action that need to be showed for this instance x_t. So, the x_t here is what we are going to call it as context or side information and this context x_t could be drawn from set some C, which could be some subset of R^d.

So, this x_t could be a vector of dimension d and it could be coming from some subset of R^d, do you understand this. For example, if I am only going to consider your age, height and your location only 3, then this x_t is a 3 dimensional vector for each person ok.

Now, the reward you are going to get, if you are going to play an action i after observing a context x that we are going to denote it as some R, R may be yeah this is some context x and what is I_t? I_t is such action chosen this I_t and this I_t is.

So, the learner in round t first observe this contextual information and based on that he plays an action I_t which is one of them. And, then he observes a reward in that round t, which is depends on both the context as well as the arm that he played. But, what he observes is not just this plus some noise t.

So, he is going to observe a reward which is noisy for example, like when you enter something right; you may like something and you may click it, but it may happen that you like it for some reason you did not click it just because, you just happened that I mean you just saw a new mobile which had all the features, but it so happened that just a week you had just bought an another mobile and because of that I cannot just buy this mobile.

So, because of that, whatever in the clicks or no clicks right, I have to take them somewhat assume that they are nosier right. For that reason, we are going to assume that the reward that learner is going to is observe is noisy version of the actual reward which depends on both the context as well as the arm plate ok.

So, what is this R? This R we are going to assume that this is a map from whatever the set of context into the set of arms to my R. So, this reward function tell you for a particular cost and a particular action that is the context action pair what is the reward I am got.

Now, yeah.

Student: First I_t it is closer and (Refer Time: 17:42) it also depends on the x_t.

It could depend on the x_t.

Student: (Refer Time: 17:50).

So, what I am saying is yes, it could depend on the x_t. But, I_t you selected after observing x_t, user may use it or ignore it, but he has observed it and after that he is choosing this.

And now, but for the learner this reward is unknown, he do not know what is this reward? If he knows this reward, so suppose let us say if for every context and action learner knows this reward function. Now, what does what will you do? What is the best thing for him to do?

Student: (Refer Time: 18:36).

For him the best thing is to do is.

Student: (Refer Time: 18:40).

Just, for find out for each context and action what is the best you can get? So, whenever you observe a particular context just go and see, which is that action that is going to maximize this and you are just going to play that right if you know that.

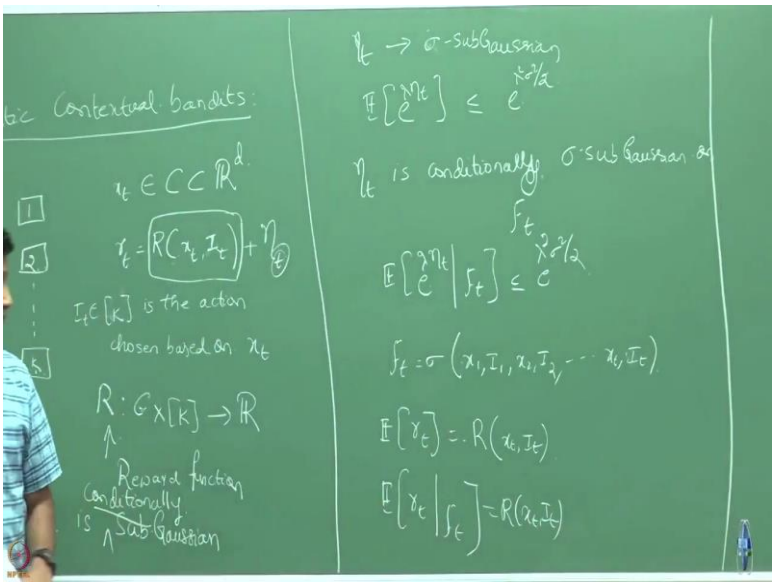
But the user learner does not know it and he is one of the goal is if you want to select a best action in each round, he has to figure out what is this function is ok. For in this current set he has to figure out what is this function for every possible pair of context and action or at least he has to figure it out for whatever context he is going to observe.

And, for the and all possible action pairs he has to find out what is this function is. Now.

Student: Sir.

We are going to call this as our reward function and we are going to assume that, this eta is sub-Gaussian. So, we are going to say that is conditionally sub-Gaussian I will. So, what I mean by this? We will just going to assume that, so we know already what I mean by a sub Gaussian noise right.

(Refer Slide Time: 20:39)



If I am going to say that, theta is what, we say that R let us say sigma sub Gaussian.

Student: x_t. Sir, you wrote.

What we know? We know that e to the power lambda eta is upper bounded by what?

Student: (Refer Time: 21:06).

e to the power.

Student: x square.

Yeah.

Student: Sir, (Refer Time: 21:17) function of the contents also for.

It could be that is why we are making it time dependent it could be. And, now what we are saying is that is why that is why I have bought this conditionally; now I am going to say what I mean by this. So, this if see; sigma eta t is sub Gaussian, but if I say that eta is conditionally (Refer Time: 21:56) sigma sub Gaussian or F_t; I will tell what you meant F_t; that means, we are going to say that e to the power lambda eta given F of t is upper bounded by sigma square by 2.

And what is this F of t? F of t is actually going to be my sigma algebra that depends on context x_1, I_1, x_2, I_2 all the way up to x_t and I_t, so, ok. So, let us understand what I mean. So, I am saying that, this eta t is conditionally sub Gaussian and this conditioning is on the sigma algebra generated by your observation so far.

Right. What is this? x_1 is your first context after that you have played action I_1, x_2 is second context you have played action I_2 till round t you have observe x_t and played I_t. So, conditioned on this; this noise in that round is going to be sub Gaussian ok. So, you are just saying that, if I know what has happened till now, the noise that I am going to observe in the reward, that is going to be simply a sub Gaussian conditioning on my observation so far.

So, you see that I have already assumed that, I have known this seen the context and which action selected conditioned on that this is sub Gaussian. So, that is it could potentially depend both on the context as well as the arm we are going to select in that round. This is about noise part.

Now, since we know that if it is a sigma sub Gaussian noise, what is its mean?

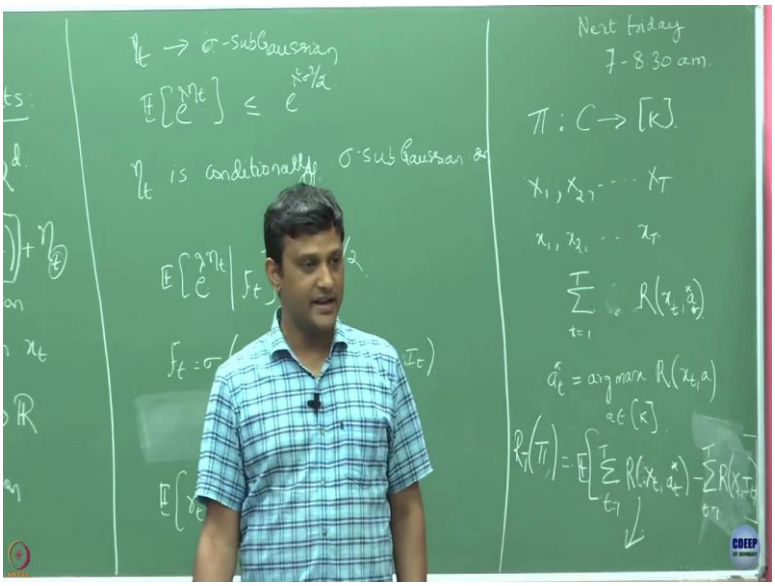
Student: 0.

It is 0, right. So, what is the reward in any round we are going to get? What is this value is? It is going to be simply R of x_t and whatever the arm you are going to play in that round I_t. The way you are going to write this is given F of t. So, then I am going to observe my reward in round t this is based on F_t includes all the observation that I have so far; that means, here if I condition F_t; I already know which arm I have played.

So, now this is exactly this, so I should be writing this actually not because I have to know what is the reward I am going to get in I_t, I am I should be knowing what is that context I have observed and what is the arm that I have played and this is the one. So, now what I have to do? If what is how should we select, so this is a problem set up right.

What is happening? The environment is generating the contexts; you observe that context, the learner the and the learner has to decide which action to or which arm to play or which action to apply. So, then what learner has to do? Learner has to come up with a mapping that maps a given context to an arm.

(Refer Slide Time: 27:01)



Any mapping that maps a context to arm is going to be a policy here. So, what is basically learner will be doing? He observes a context and then he has to decide which arm I have to play. So, that is like a mapping he has to find out to given a context what. Now, our goal is he want to come up with a policy pi, that gives him the maximum reward right.

What is the maximum reward? The maximum reward, what is the maximum reward that a learner can get? Suppose, let us say he knows the reward function, what is the maximum that he can get over a period time t. So, suppose let us say

How we have denoted the context ok, I have denoted context by x_t right in round t x_t. But, this context I have not yet specified how they are generated right.

So, they could be generated stochastically in every round and we are going to assume that, these are generated in an IID fashion. This context are generated according to some common distribution and that are revealed to the learner. So, I am going to denote that whatever the contexts I am going to observe over a period, let us say X_1, X_2 up to X_T these are random; so, I am just writing it by capital notations.

Then, suppose, let us say you have observed a particular realization x_1, x_2 and x_T over a period of time t, over this time t what do you think how what is the best what is the best amount total reward that the learner could get. So, the best thing the learner could get is,

if you have observed x_n, let us say in round t in let us say this is from over t equals to 1 to T right.

If what is the a you should play in round t?

Student: Maximum value.

Let me write it a t star let me write this a_t star, can you tell me what is this a_t star should be?

Student: argmax.

a_t star is?

Student: argmax (Refer Time: 30:21). R of x t comma (Refer Time: 30:29).

So, if you know this function R and if you observes this in round x_t, you should be playing an action a, which maximizes this ok. But the learner right a priori do not know this, R functions right the reward function that is hidden from him. So, he will play whatever actions he feels or whatever the actions that is derived from the policy he is using.

And, now we are going to define the regret of that policy with respect to this; this is the best he could get right. Now, we are going to define regret of a policy pi say let me as the expected value of. So, this is the best he could get, if he applies the best action in round t; a_t star is the best action you could apply if you have know the R function that is how we have defined a_t star.

And, now we are comparing these against if we place I_t in round t whatever that action he is going to play, this is the total reward he is going to get. And, now we are going to compare this total reward against, the best total reward he can incur over a period of time t.

Student: (Refer Time: 33:39).

Yeah.

Student: This we can only.

This we can compute only. So, this is what I am saying this is the best you can get if you know what are these reward functions and this is you are getting through your policy with by playing I_t's. Here, you do not know this reward function. And, now we are going to define and what is I_t? I_t is whatever the policy pi tells you to play in round t after observing the context here ok.

Now, what is this expectation about, what all it averages over here?

Student: (Refer Time: 34:25).

So, one randomness is already over, there are different contexts right. And, then what is the other randomness?

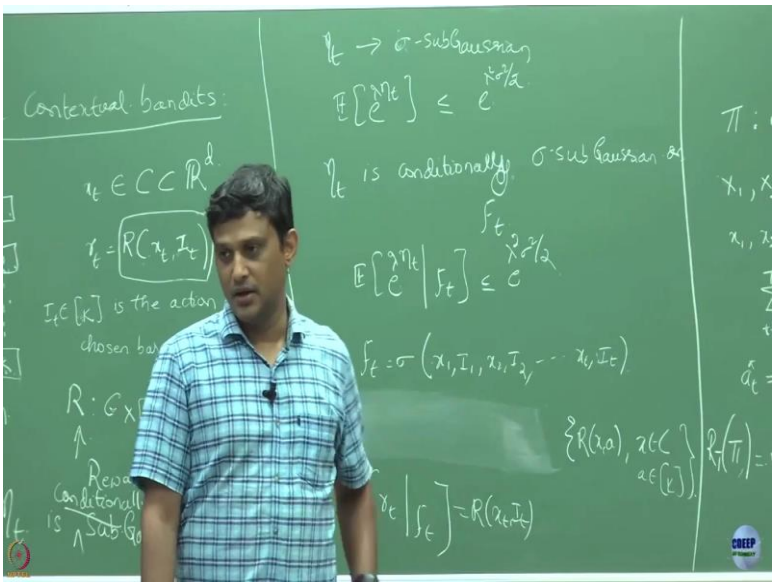
Student: (Refer Time: 34:37).

I_t could also be random right, the arm I am going to play in every round maybe deterministic, but it could be random also. So, I have averaging over both of this randomness here.

Now, suppose for every possible context, an arm a I know the mean reward is given by this reward function right that, if on a particular context if I am going to play a particular arm, I am going to observe this noisy reward, but it is mean value is given by this reward function right.

So, I know that for every context and arm that is context arm pair I know let us say, I can treat this as an arm, one arm instead of treating these as arms; I am going to treat context and paired with this arm as all possible arms. Can I do that? So, in that case, how many pairs I am going to get? Cardinality of C into cardinality of K right, for every context I am going to pair one arm and I am going to get.

(Refer Slide Time: 36:09)



So, let us say I am going to collect this R (x, a) where, x belongs to C and a belongs to k. Now, if I do not know R function, if I knew R function I can compute all these values and I am going to treat this R (x, a) as the mean value, mean reward I am going to get when I going to play action a when I observe context x. So, now suppose if now if I am going to think this as a some bandit problem, where my arms are this pairs x,a pairs.

Now, I am figuring out, if I can find out estimates for each of this pair what is this reward value then, also I can find out if I can estimate all these values right, then I can figure out whenever I have a context what should be the best arm I what is the arm I should play on that.

Or alternatively you can just think of these are different arms, I am going to play them and the goal is whenever I observe a particular context, I want to play arm which has the highest mean whatever.

Student: Sir but (Refer Time: 37:50).

Yeah.

Student: (Refer Time: 37:53).

Right. But I am this is what like you enumerate all of them, you that is why I enumerate for all context and all arms. So, in each, yeah.

Student: This C could be completely filled with a finite value.

Yes. It could be but for a time being assume it could be it is finite. It is finite number of arms is also finite. So, we have finite pairs here ok. Now, I am going to treat them as a different arm, now my goal is just find for every context which is the best one. So, if I am going to treat them as in separate arms, can I use my standard bandit algorithm to figure out what is the best action eventually for each context?

Student: (Refer Time: 38:46).

Yes right, like basically for every context you just observe play an action and estimates it is reward and similarly you do it for every context. So, now you have this value and based on that you just whenever you observe a new context just correspondingly play whichever has the highest mean from whatever your estimates.

But, what is the problem with that?

Student: Set of arms.

The set of arms is could be potentially large right for especially, like as you are seeing if this context could be very large even if it is a finite, but very large; the number of arms is already large and we already know that how does the regret scale with the number of arms?

Student: Root k.

Root k right like in the minimax setting it is scales like square root k. And, if I am going to treat these many pairs as arm instead of just k arms, the regret can be very bad ok.

Student: (Refer Time: 39:56) not be using the information that will result on all the (Refer Time: 40:00) to evaluate.

No, but right now how you are gathering? Is there any right now we have just said for every context and pair this is the reward. Did I say anything about how one pair says anything about the other pair?

Student: (Refer Time: 40:17).

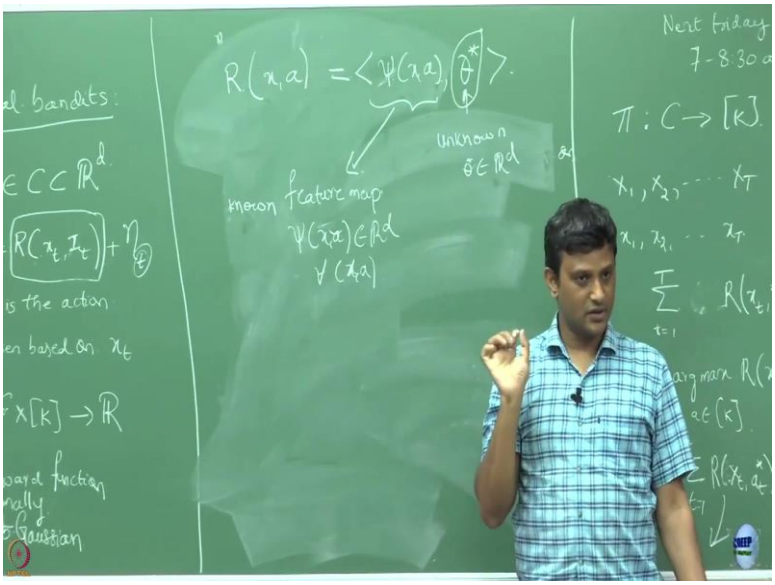
We do not have any such as of now we have not bought in any of those angles right. Now the question is, yes if you are going to do like this; if you are going to treat blindly this problem as in earlier bandits setting with considering these many pairs as arm, your regret bounds can be very bad right. Now, the question is when I can actually do better?

As you are seeing, maybe way if there is an information that I can extract from one context about the other context maybe I should be able to do better in this case and it is often the case. When you have this contextual information, it is true that right like those visitors who are coming to let us say on a web page usually, that kind of things the youngsters do could be potentially very different from the kind of things the old people do right.

So, if you know that some people are in a certain age category you kind of already kind of narrow down what there could be interest and, infer from one guys whose interest with the another guys who are in a similar age group. All these things you could potentially do that right.

So, what we are going to next assume is, this rewards here they can be parameterized such that, in a parameterized such way I there is some correlation of rewards across the contexts ok. So, what we are going to do is, we are going to assume that now we are going to make a special; so far we did not say anything about how this function R look like.

(Refer Slide Time: 42:36)



Now, we are going to assume that this function R, for a given context x and arm a is going to look like; in this function, what is phi? We are going to now assume that, this is some theta star, this is unknown and this theta star belongs to some R^d and this is we are going to say that this is a feature map, and known feature map.

And, this reward is given as the inner product of these two quantities and of course, this feature map is also giving me vectors which is in R^d or has a dimension d.

So, known feature map belong and now phi of x of a, belongs to R^d, for all x a.

Student: Sir, (Refer Time: 44:17).

Yeah.

Student: Is this psi known?

Yeah psi is known. That is what I have written right this known feature map. So, let us understand this, what we are saying here? We have a fixed theta star here which does not depend on what context or action you are talking about. This set of parameter is independent of this.

And this is an unknown. Now, we are making an unknown quantity theta star which does not depend on context or action and we have a known map, which depends on context and actions and we are now saying that the reward is in a product of these two.