Bandit Algorithm (Online Machine Learning)

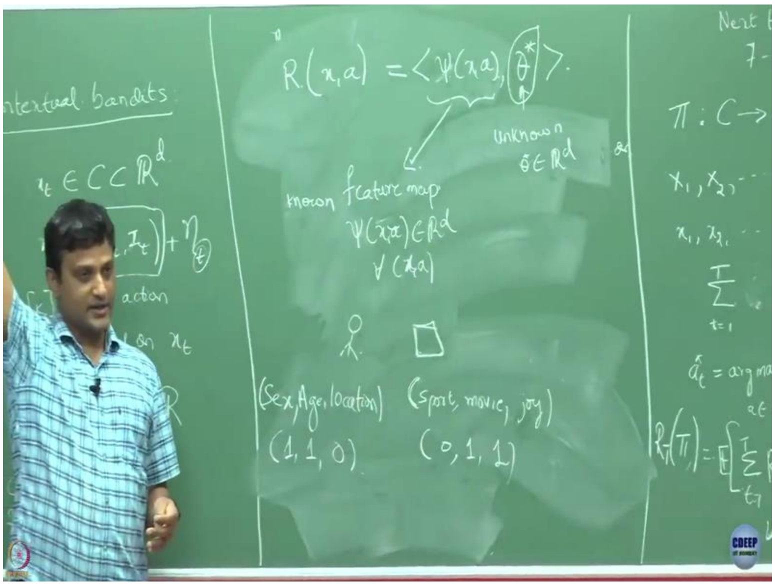
Prof. Manjesh Hanawal
Industrial Engineering and Operations Research
Indian Institute of Technology, Bombay

Lecture – 43
Introduction to Stochastic Linear Bandits

The feature map function here, now we are saying it is depends on both x and a right? So, now let us take the example of; again your recommendation system. Now, what we are going to do is users that are come into a system, you are going to categorize them. Maybe, you can say through different different category; you can say like ok, it you; it could be male, female, a old, young or from this geographical location, this geographical location you can just itemize.

And the movies or whatever the product you want to recommend, you can also categorize them. For example, if you are going to show a product, you can say that this is related to sports, this is related to some academics or this is related to some daily use material whatever you can put them into different different category.

(Refer Slide Time: 01:37)



ntextual bandits
$x_t \in C \subset \mathbb{R}^d$
action
on x_t
$R(x,a) = \langle \psi(x,a), \theta^* \rangle$.
known feature map
$\psi(x,a) \in \mathbb{R}^d$
$\forall (x,a)$
unknown
$\theta \in \mathbb{R}^d$
(Sex, Age, location) (sport, movie, joy)
(1, 1, 0) (0, 1, 1)
Next
$\pi : C \rightarrow$
$x_1, x_2, \dots$
$\sum_{t=1}^{T}$
$\hat{a}_t = \arg\max$
$R_T(\pi) = E\left[\sum_{t=1}^{T}\right.$
CDEEP

Now, depending on what is the context you have observed and now what is the corresponding arm, the phi of x comma a is going to tell what is the common feature for this pair. For example, let us say; you let us take this example ok, now let us say you have a some product and there is one guy here.

This product, you have come up with a categorization of this product; let us say this is sports or it could be movie related or it could be let us say sports whatever we are some joy some something. This product can be anything I mean, but you are going to put it in one of these category and this users, you can put him in a category like you can say its sex, its age group, maybe its location different ok.

Now, how you are going to make a feature map out of it. So, suppose let us say this is for your orbit; so for every guy you are going to set this features and for every product you are going to set this feature. For example, you will have a collection of products and different different people are logging into you.

For example, let us say you happen to show a DVD which correspond to some movie. Now, for that how you are going to assign? Definitely, it is not a sports related item that component you are going to set it as 0, it is related to movie maybe you can set it as 0. Joy, I do not know; you can set if it is thinking it is a good movie enjoyable; let us say it could be 1.

Now, for this person; let us say a girl happened to enter this website; for her it could be let us say sex can take 1 or 0; it could be like let us say 1 for female, we have this. And also you have class of age, you have two things very old and very young and let us say it is a very young, you can get this and location also maybe you would have categorized and let us say you got it.

You see that depending on that user and depending on the product; now you have this feature map like this which is given by this guy phi x comma a. So, this; the whatever the feature map I have written here right, it depend on what was the user and what was the product right. So, that is why it depends on both the user and the product and I can generate such features.

And it is all like all the recommendation systems are exactly working based on this features extracted right. When you are going to log in, the way it stores your informations maybe through such categorization and generating features out of your profile information yeah.

Student: But, we are agree with that (Refer Time: 05:27) known a priory.

Yeah, known a priory; this is what like you have already built in, like you have already by some experts; you have already built in, how to map a given pair of context and a product we like featuring like this. And I have just given them like; this feature extractor can run in have dimension of millions in it right because you can have so much of categorization. If you want a very more precise narrowed on information, you will have lot of features.

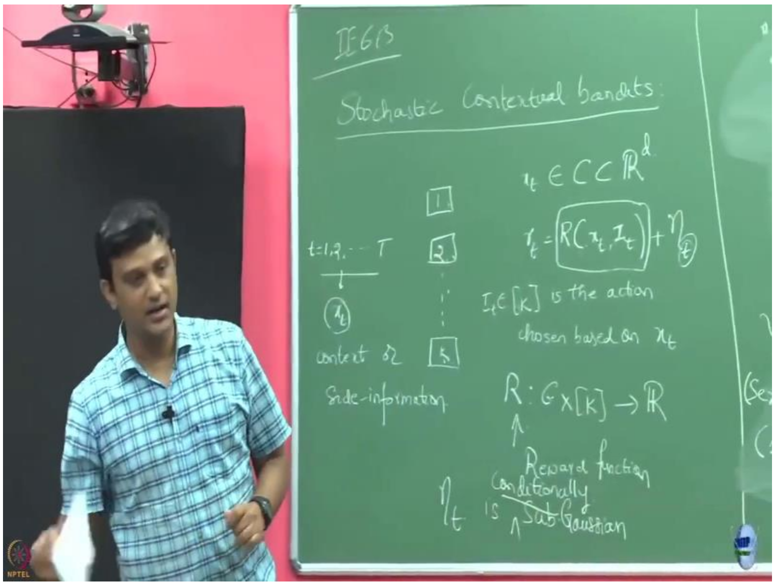
Student: Yes.

You could be logging in morning, what is the time of logging in, what could be the season; so, many information right because more information you have maybe better decisions you can make. But that is a known map; that you have come up with offline somehow with lot of discussion with your product team, how you should be generating this feature map yeah.

Student: Sir, these are the inputs to the feature map right; so.

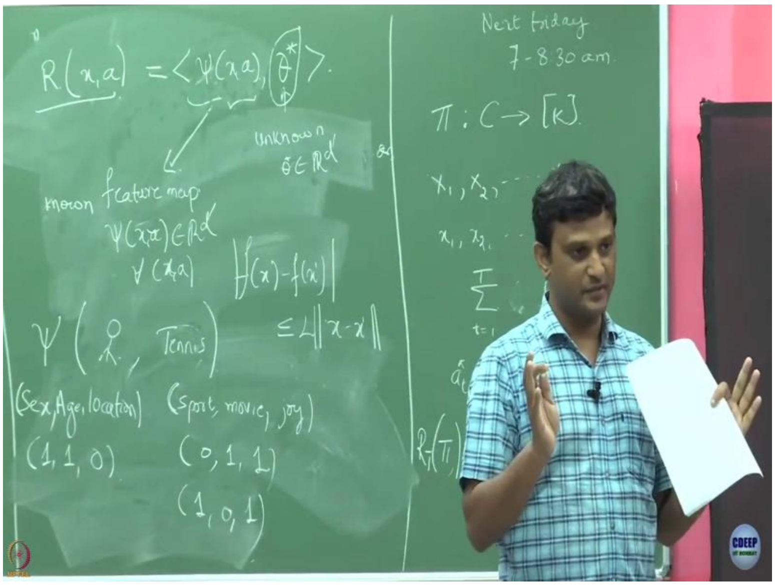
No, this is the output of the feature map; for feature map, this is the input you have; that is what I said right like. So, in this case here product is the one you want to sell.

(Refer Slide Time: 06:51)


IE613
Stochastic Contextual bandits:
$x_t \in C \subset \mathbb{R}^d$
$t = 1, 2, \cdots T$
x_t
context or
Side-information
1
2
k
$r_t = \boxed{R(x_t, I_t)} + \eta_t$
$I_t \in [k]$ is the action chosen based on x_t
$R : C \times [k] \to \mathbb{R}$
Reward function
η_t is Conditionally Sub-Gaussian

So, product is your kind of arms here; when a user comes you want to see which product you want to sell him ok. Let us say you have thousands of product; now your goal is to show him the right product. Now, users log in right like there could be so many users. Now, given a user and given a product; you should be able to come with the features like this.

(Refer Slide Time: 07:26)



If instead of this, so here I said it was a DVD right instead of DVD; let us say it is a tennis ball. Now, in that case this feature will change to 1, 0, 1 maybe right. So, depending on the product and the user, this features can change.

Student: So, the finally, vector will be a (Refer Time: 07:48) the vector of length 6 or like the combination of the two.

Yeah, it is together; so you can decide like how you want to come up with this vector, but only thing is I need to tell you, what is the context and what is the product for which you are want to generate this ok. And now this theta star is unknown and this part is known; now what it boils down; the problem boils down to?

Student: (Refer Time: 08:30).

Determining this theta star; if I know this theta star, I already know my reward functions right. Now, the question is; how I am going to determine this theta star from the interaction with the environment which is generating this context? Now, what is that?

Student: Sir, theta star is constant or not?

It is a constant, it does not change over time and it is independent of context and actions it is, but for a given environment; it is fixed. If your environment is going to change, maybe it can potentially change right.

Student: (Refer Time: 09:14).

So, you remember in the bandit instance; how we identify a bandit instance? Through the mean values right of the arms and if the mean values of the arms change, it is a different bandit instance; here if I have one theta star that is capturing one instance here and if theta star changes that it may be dealing with a different bandit, different instance.

Student: (Refer Time: 09:38) feature maps.

Feature maps we will freeze it like.

Student: Yes.

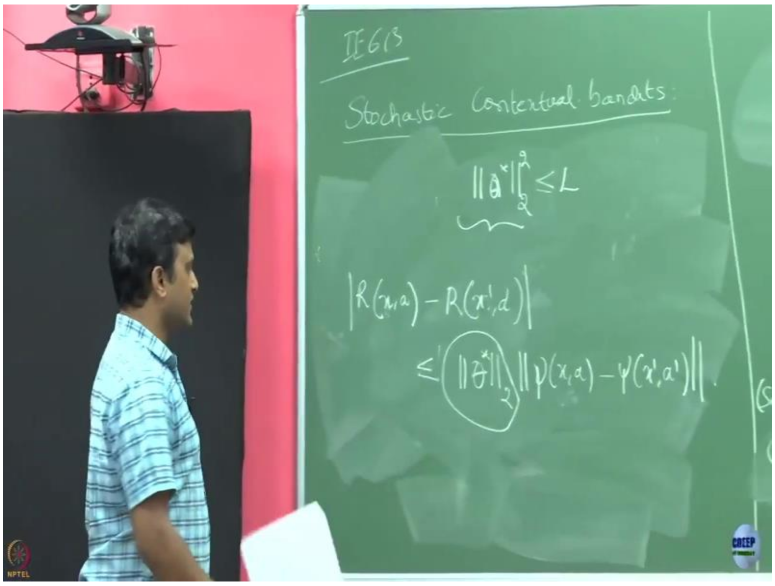
Because that is known to us right; there is nothing you this is only unknown this is known to you whatever you are going to use its fine; we do not care about it ok. What matters is the one; the quantity which I, it is not known to me, it should not change while the game is going on; we are going to assume that that is remains fixed throughout the game. And in that point; in that sense, we will also assume that once we fix the feature maps, we are going to stick to them throughout the game.

This reward map, I have written in such a way that this guy does not depend on x and a and this guy depends on x and a, but it is known to us ok. If you, for this setup here it is making it linear, but you could assume anything, if you do not assume anything about this values; as I said I mean it is as good as like you have to treat each pair as a different arm and you have to learn all possible pairs that could make the regret too large right.

So, we want to that is what and we also said that in such when you have such contextual information; it is natural to assume that one context to reveal some information about the other. So, that is why we have having this common parameter which will make whatever information I am going to get about one arm, I mean one context and arm pair to extract some information about other context arm pairs. Now, as I said; now in this, the problem boils down to like if I know this theta star; I know my reward function already and then I have I can know how to play optimally, but I do not know this theta star. Now, how to find this theta star?

Now, in general whatever dimension this theta star it is right; we assume that that dimension is known and we are going to assume that further, this theta star; the norm of this theta star is bounded.

(Refer Slide Time: 12:15)



IE613

Stochastic Contextual bandits:

$$\underbrace{\|\theta^*\|_2^2} \leq L$$

$$|R(x,a) - R(x',d)|$$
$$\leq \boxed{\|\theta^*\|_2} \|\psi(x,a) - \psi(x',a')\|.$$

So, we will further assume that like if you; what does this mean? These are basically saying that this additive norm of this is going to be some bounded quantity.

Student: We know the bound.

We know the bound; we assume that this bound is known to us ok. Now, what is this basically saying that; suppose; now if I am going to look this; suppose let us take one context pair and take another context pair. So, let us say x, a and one pair and x prime a prime is another pair.

If I have this assumption, this I can write it as; so this theta star is common right, it does not depend on what is x and a. I can just apply my standard order equality on this to get and then I can just; now treat as, if I know putting this assumption like this theta star is upper bounded by like this; is like some assuming that this function has this kind of Lipchitz property on it, you understand what is the Lipchitz function?

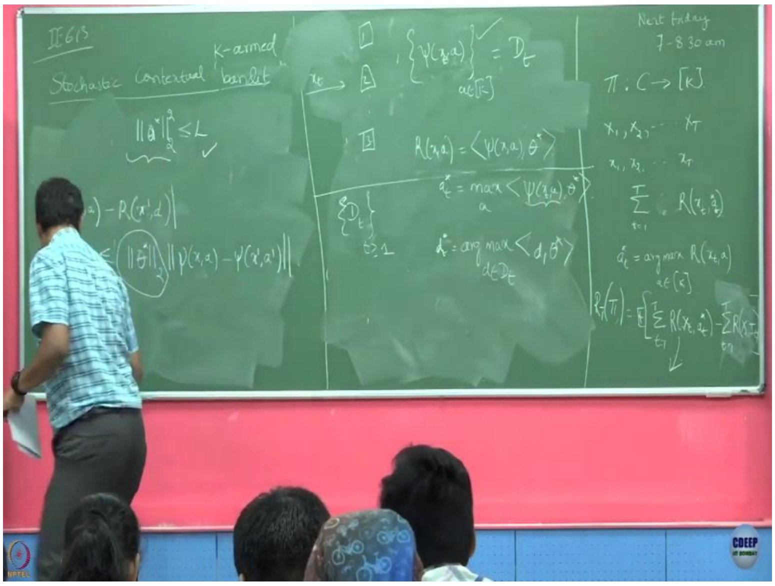
So, what is this basically? This function here is nothing, but as somebody said this is a linear function in theta star right ok. So, I am just; so what is the Lipchitz function; if we have f of x minus f of x star, what is this?

Student: (Refer Time: 14:41).

So, in the appropriate dimensions whatever right; so this is. Now, similarly what we are basically doing is for a for a given theta right; now if I am going to look this as a quantity as a variable, now what I am going to get is; just this quantity where now I am just putting a constraint on this theta star, known to be bounded is basically saying that this function is Lipchitz with that constant L ok. So, we are just going to assume this and that L is known to us ok.

So, basically limits that my theta star lies in some bounded space ok. So, now I want to look at slightly; so now we are going to stick to this assumption that my rewards are linear or they are parameterized by some theta star ok, that is unknown to us. Now, we are going to look it in a slightly different angle and that is going to that is another well studied set up. What we are going to now do is, what is basically happening to us?

(Refer Slide Time: 16:27)



When a context came to us; when a context came to us, this was my arm 1, arm 2, arm 3, right. What I am basically doing? What I am basically doing is; first I am looking for, I have already pre computed them; let us say, my feature maps give me for every action pair, what is that?

So, what I can do? Whenever a context comes to me, I can think of for every action a; I can look for this feature maps, for that context. And then what I has to basically do is look for a action or look for a feature; so I have these many features now which I have extracted from my feature map, for a that particular context. Now, I have to basically; what I have to do is, now because I know my reward is inner product right. So, all I have to now do is; find a feature, now x is fixed because x have observed. Now, I how to do this find that feature map that maximizes this product ok.

Now, the problem has boiled down to; looking for a feature vector that maximizes the product ok. Now, because now I have kind of abstracted out those, abstracted out this arms through the features now; now it is about looking for which is the best feature that optimizes my function in that round t. So, now, it is about like; now I can think the problem forget, throw away the arms; now it is about which is the best feature I have in that round ok.

So, in that way I can now think of in every round I have a bunch of features right and I have to decide which is that feature which maximizes this function right? Because I assume that already I have told you, I know this feature map and every whenever a contest come; I have this feature vectors. Now, the problem boils down to maximizing this linear function, over those feature vectors ok. So, now we; then we can just forget this and now we can think of as a linear optimization problem, over a feature set is that make sense.

So, what we are now basically saying is; now forget about my arms and the context coming. Now, I am thinking that in every round; I have a feature set, every round I have a feature set like this. So, x will give me this feature set right because I knew already a, I knew this phi function; so I have this feature set.

Now, I can think of in every round; I have basically a feature set and let me call this as D_t. So, now I am going to think this as in every round, in round t I have this feature set and I have to decide; which is the feature that maximizes my product ok? So, what I am; I am saying is, I what I was doing earlier in this setup? I am trying to find, in round t; I was trying to find an a; t star, that is maximizing this linear function right.

But, now this feature vectors; this set I am denoting it by say D_t, then this problem boils down to; see this D_t is or a collection of features now; this entire set as I said I am calling is D_t right. So, this set is collection of features that has come to me in round t. Now, I am looking for a feature; now I can equivalently write this as look for whatever; it is a feature or let me call it as yeah d_t; d_t star which is arc max of d belong to D_t times d; theta star.

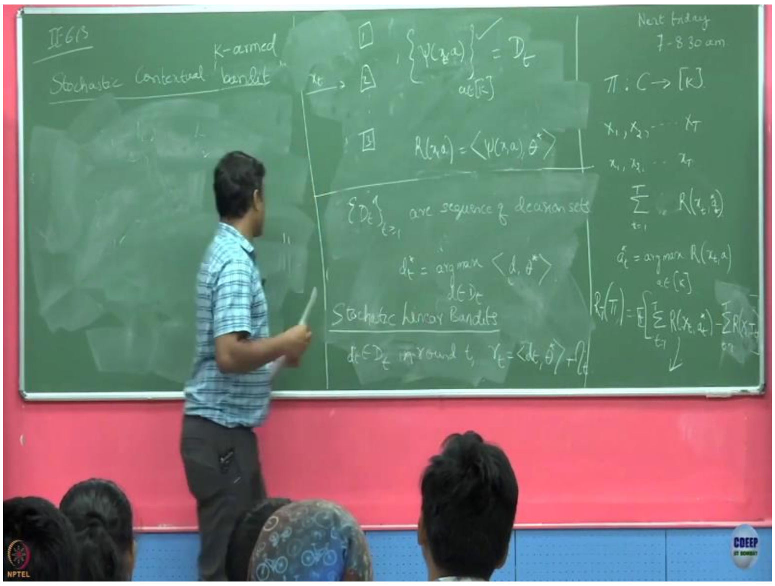
And this, can I do this? Can I map my problem of; so whatever the initial setup I give, they are going to refer to that as; so this set up where we have k arms and the contexts are coming, we are going to refer to this as; stochastic k arm bandit ok; stochastic contextual k armed bandit right. Because, this was like exactly the same as my stochastic k armed bandit, but it was just like we have the context there.

Now, after going through; once I made my assumption that, my reward function is linearly parameterize through this theta star. Now, arms has no I can, I do not worry about what the arms, but I have to worry about the features. So, if I know which is the feature vector that maximizes this function, I already know; what is that arm that is good for the that particular context x.

So, that is why; what we are now saying is I can think this problem ok; so now I am going to say that, I am going to now going to abstract this; now we are going to abstract this. Now, in every round; I have a decision set d_t, that I am going to call it as simply decision set which involves set of vectors which are feature vectors and now that is revealed.

And now, what I have to do is in every round, I have to pick one feature vector from them which maximizes my inner product.

(Refer Slide Time: 24:31)

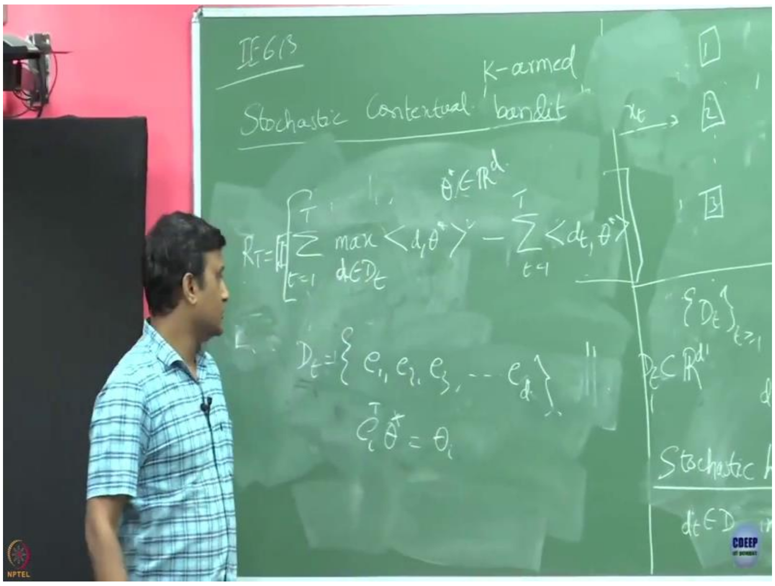

IE 613
Stochastic Contextual K-armed bandit
$x_t \rightarrow$ 1 2 3
$\{\psi(x_t,a)\}_{a\in[K]} = D_t$
$R(x,a) = \langle \psi(x,a), \theta^* \rangle$
$\{D_t\}_{t\geq 1}$ are sequence of decision sets
$d_t^* = \arg\max_{d \in D_t} \langle d, \theta^* \rangle$
Stochastic Linear Bandits
$d_t \in D_t$ in round t, $r_t = \langle d_t, \theta^* \rangle + \eta_t$
Next friday 7-8.30 am
$\pi : C \rightarrow [K]$
$x_1, x_2, \ldots x_T$
$\sum_{t=1}^{T} R(x_t, a_t^*)$
$a_t^* = \arg\max_{a\in[K]} R(x_t,a)$
$R_T(\pi) = E\left[\sum_{t=1}^{T} R(x_t, a_t^*) - \sum_{t=1}^{T} R(x_t, \ldots\right]$

So, now I have to choose a d_t star belonging to; so this is, I want to do this. And now, I want to identify this feature; if I can identify the feature and now through this mapping I also know, which is the action that is good for that feature.

Now, we I am going to call this setting; we are going to simply call this setting as stochastic linear bandits. So, I mean just in every round; I have this decision sets, this theta star is unknown; now my goal is to identify what is the best feature from that ok. So, now what I am going to observe? I am going to observe like in every round; if I am going to play, if I am going to; so this is I have this set D_t which involves set of features.

Let us say if I am going to play x_t belonging to D_t, in round t; the reward I am going to observe is now simply x_t times ok. So, let me not call this x_t because that is confuses that with that. So, now it is going to be d_t times; theta star plus the noise, this noise term is same as earlier. So, now I am going to call this setup as simply stochastic linear bandits , but now you see that, this is nothing, but again this is just an abstraction of our stochastic contextual k arm bandit ok. Now, what is the regret in this setting? The regret setting for this is simply what I have to do? I want to play select the best feature from my decision set in every round and if you are going to do this in every round, you are going to get the best actions, but you do not know always what is this theta star. So, you cannot play this in every round.

(Refer Slide Time: 27:56)


IE 613
Stochastic Contextual K-armed bandit
$\theta^* \in \mathbb{R}^d$
$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \max_{d \in D_t} \langle d, \theta^* \rangle - \sum_{t=1}^{T} \langle d_t, \theta^* \rangle\right]$
$D_t = \{e_1, e_2, e_3, \ldots e_d\}$
$e_i^T \theta^* = \theta_i$
x_t
1
2
3
$\{D_t\}_{t \geq 1}$
$D_t \subset \mathbb{R}^d$
Stochastic
$d_t \in D$
NPTEL
CDEEP

So, the best you are getting here is max over a minus this is what; t equals 1 to T and this is the regret I am going to incur. And we are and this is nothing, but this is the R_T, you are going to incur or which is simply d_t and theta star; yeah. So, if I am going to play d_t in round t, so the environment your; this is the value you are going to observe right?

Student: d i.

D?

Student: d i t; the arm.

No, that that arm has concept has gone right know?

Student: (Refer Time: 28:58).

We have abstracted out and now we are saying, it is every time thing in terms of the features now because now for any arm pair, we have this set of features now; in every round. So, when I have a feature x_t in round t, I had got this set of features using this feature map. So, this is what matters to me right now; I do not care what are the arms now.

So, now arms identity do not matter to me, only the features I have this set and then from what matters to me is; can I identify a feature that map optimizes this theta star? If I know which is that feature that optimizes this I already know what is the arm that is giving me the best reward.

So, that is why now I am saying that instead of worrying about what are these arm; I am only worrying about in every round instead of saying x_t has come, I am going to say that a feature set d_t has revealed to me and now I have to identify what is the best feature that maximizes this ok.

But, so if I knew theta star; this is what I have not done in every round, but I do not know; so I would have whatever my policy would have played d_t in that round; so this is the reward I would have obtained. Now, I am going to compare my performance; this is a total reward I would have got against, what? (Refer Time: 30:33) I would have got, who knew theta star. And I am going to call this regret and expected regret; you are going to and this is maybe I will just going to denote the expected regret of this quantity yeah.

Student: Expected.

So, let us look at this is quickly a couple of cases. So, now suppose let us say this set D_t, I said these are vectors right. Suppose, let us say feature vectors right; let us say these are my unit vectors, you understand what I mean by this? What is e_1?

Student: (Refer Time: 31:36).

1 all 0; e_2 is? 0, 1; all 0; so each e, e_1 is a vector of dimension k, vector of dimension d, where only the ith component is nonzero one. Now, suppose let us say all my decision set in every rounds are like; this my decision sets are like this ok. Now, what is this? Now, what is the R max? This is going to give me?

Student: (Refer Time: 32:22).

So, whichever the component the largest in theta star, it is going to give that right. Now, suppose this theta stars; assume that the theta star is a vector where each component corresponds to mean of an arm right.

Now, this operation is just telling that I am pulling that arm which has the highest mean right and what is this telling? Whatever the component you pulled, so now if I have this kind of decision set, is it not the same as my stochastic k arm bandit problem?

So, here what my decision sets is such that they have unit vectors in this and now I know that; if I am going to take e_i of theta star, this is nothing going to be theta i right. So, now you can just think that; this is always going to be just if my decision sets are like; this is just going to give me the corresponding component in my theta star.

So, if I am going to just treat in my earlier stochastic k arm bandit problem; if I am going to just think that the mean arms, means of all the arms I am going to concatenate and write it as a vector theta star. Now, it is about; now your decision is about which unit vector to pick from this decision right.

But this unit vector is; it is just telling that which arm to pick in this case. So, you see that this stochastic linear bandit, if am restricting this D_t to my unit vectors, this is already captures my stochastic k arm bandit setting ok; where this theta star, this unknown parameter here is nothing, but the vector of the mean values of the arms which were; which were characterizing the environment there.

Now, this theta star here is characterizing the environment ok. So, in that way for this special set of D_t; D_t which unit vectors the stochastic linear bandit is same as a stochastic k arm bandits. So, k means d; d arm bandits because it is dealing with d arms here.

If theta star has dimension d; that means, it is dealing with d arms and but the thing is the D_t need not be always unit vectors like this, they could be anything each D_t could be any subset of my $R^{d'}$; whatever is my feature space that I am looking into ok. So, if you are going to take this theta star to be belongs to R^d.

Student: Yes.

Ok. So, naturally I am; this D_t.

Student: Yes.

Each element, I am doing a new product. So, each of them has to be of dimension D.

Student: Yes (Refer Time: 36:03), but total half k; k arms right (Refer Time: 36:06) k and when we will require that k is small less equal (Refer Time: 36:09) in this (Refer Time: 36:12).

Now, what is k?

Student: (Refer Time: 36:20).

k equals to d?

Student: No.

k equals to d for us in this case?

Student: We are getting the; by that functions (Refer Time: 36:32).

Yes.

Student: Yeah, so e a belongs to now k.

a belongs to k, yes.

Student: So, now we have k (Refer Time: 36:46) d dimensional.

Student: k e dimensional; v dimensional vector right.

Right.

Student: (Refer Time: 36:52) theta k; d dimensional.

That is right, but; so that is when I did this. Now, I am saying that forget that like for each arm, you have a feature. Now, only look at a feature vectors now; the D_t is nothing, but the collection of feature vectors.

Student: (Refer Time: 37:11); we asked you k number of feature vector right.

If there are k number of features vectors.

Student: Yes.

Yes, fine.

Student: e_1 to e_k; e_1 again d dimensional vector (Refer Time: 37:28); when it is a d dimensional vectors and they were like k; k of k k of such (Refer Time: 37:35).

Exactly, so each arm. Whatever, this theta star here right; I want each of this.

Student: (Refer Time: 37:49).

Unit vector corresponding to one component in this; so, I want as many unit vectors, as there are; as the dimension of this theta star. So, that is why I like in this.

Student: (Refer Time: 38:05).

Case, I want k to be same as d; yeah there are that many unit vectors. So, is this clear why this is giving me back if my d t's are like this why this is giving me back my stochastic k arm bandits?

Student: (Refer Time: 38:23).

So, yeah clear; it has t has to be k for this mapping to happen ok. So, let us stop here; we will continue this in the next class.

Student: Ok.