NPTEL

NPTEL ONLINE CERTIFICATION COURSE

REINFORCEMENT LEARNING

Contextual Bandits

with
Prof. Balaraman Ravindran
Department of Computer Science and Engineering
Indian Institute of Technology, Madras

Another class of algorithms in the Bandit space which ideally I should have spent a lot more time but we are really running out of time in Samsung's I do not want to spend too much time on bandits I want to get on and talk about other topics in reinforcement learning so there is something called contextual bandits.

(Refer Slide Time: 00:44)

So what are contextual bandits one way to think about it is to say that I am relaxing my bandit assumption that there is no state. I am introducing a state, right so this is like I am taking one step from my immediate RL problem towards the full RL problem right. Now so what did they

mediate our real problem had it had all this Explorer exploit trade-off stochastic right it has the are the noise in the selections that it had the fact that I knew about the rewards only through sampling I do not know the true underlying reward distribution so all of these characteristic of the reinforcement learning problem it had.

Right so what is what is the third thing we are adding now and adding this state dependence right in different states I am going to have different kinds of reward functions so what is still missing, sequential dependence remember I was talking to you about when cycling you know when you might make a mistake initially but then that is going to influence how hard it is going to be to balance in the later stages right so I do not have that sequential dependence now right I so I give you a state you give me an action right independent of what action you selected will give you another state.

They keep going like this so, so essentially what you are going to do is look at an input and then you are going to take an action right and then you're going to move on right you get to the wall and then you will move on to another state so one other way of thinking about it is to say that hey! I am not solving one bandit problem I am solving a set of bandit problems right and I will tell you at every point which bandit problem you are going to solve okay.

Here is problem three tell me what is the option, option here is problem five tell me what is it option like the right so one may way to think about Saul solving this well it's not really nice if the way I have stated the problem ok it's pretty nice so one way of solving this is to essentially say that okay. I am going to maintain suppose I have 10different bandits i will maintain 10 different asset of action values or 10 different sets of policies right I am way is going to do this all independently and I solve it okay.

So that is one way of handling it but what if I tell you that these problems can be many not just like a handful but i am going to have a huge set of problems right . For example i told you one of the success stories of reinforcement learning was in like those new story selections and add selections and so on so forth right people remember that the very first class i showed you some

examples where I said Yahoo's chose add new stories to show on the web page by treating it as essentially they are treating it has banded problems right.

So I have like 20 new stories when somebody comes to the web page I will show him one out of the 20 stories essentially like pulling an arm and if the person clicks on it I get a reward of 1 if they do not click on the story I get a reward of 0 it's like a binary bandit problem but then if i solve it as a single banded problem it is not very interesting right because that means that everybody who comes to the page will get shown the same news story or if you go back again will get shown the same you story.

That is not going to be is not a good thing so what you really want to do is depending on who comes to your page right sure you want to be able to change your choices so one way of thinking about it is now to say that hey i am going to group all the people coming to my page right if somebody from Group one comes I will use this bandit, somebody from group 2 comes I'll use this bandit somebody from group 3 comes I will use this bandit right does it make sense .

And now somehow found a way of grouping them and how do I group them. Best well there are too many options here so you could look at their browsing history you could look at if they already have an account on say Yahoo or Google or something you can look at the data that they filled in right. A variety of things right so you can look at which country they are coming from what time of the day they are clicking from are they are browsing from a phone, or from a laptop so there's so much information that each one of these internet sites gathers about you right in fact they know more about you then you do right.

That is not is not as stupid as you think it is in fact there are lot of times when people are consciously not aware of what they are doing right but then by watching the person's behavior they are able to form a more informed model about you what you are trying to do then what you yourself are doing you quite often you are diseltarly truly doing things right so you do not really realize what you are what you are doing what you want anyway so fine so you can do all kinds of things you can group the users right.

Is that something you can do better than hmm speech variables which variables I use include them exactly so why, why would that help. It will reduce the number of parameters for each person you can define a unique set of parameters and that's not interesting right it's going back to solving the same problem for you do not want them to have unique parameters you want them to have some kind of grouping so essentially what this will allow you to do is it gives you greater flexibility to figure out what kind of generalization you want to happen over the users right

When I say fine groups right and then assign one bandit to each group in some sense you are making everybody in the group have the same recommendations right the same, same stories will be shown to them right by doing this kind of a parameterize function so you could be sharing some parameters with him on four accounts right so you should be sharing parameters with him on three other things right so essentially there will be like a smoother generalization across people right so the extreme case of this would be where everyone has a unique parameterization but.

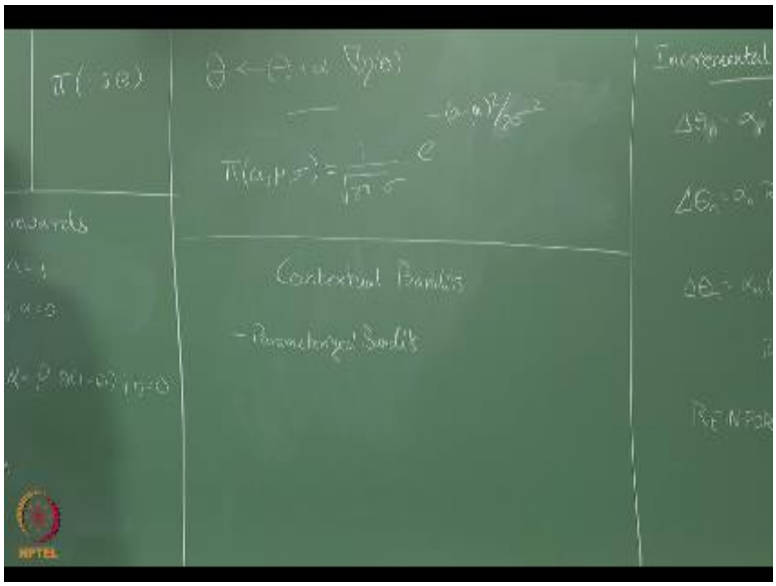
If you make everyone you force them to be unique then you are not doing much, much generalizations right so you essentially want some kind of generalization to happen so that will kind of fallout from doing this in a parameterization fashion right. So there are many ways so contractual bandits is just a problem setting there are many ways of solving it one of them is to do this kind of parameterization of the value depending on, the state as well as the action.

So what I mean by the action here the action can have parameters so what do I mean by that let go back to the new story case right, so I can think of it as new story one I can also think of it as political article, from India right about the central whatever so I mean I can I can think of some kind of attributes that describe the new story as well right so these are attributes that describe my arm they are not describing my states.

They are describing my arm so in this case state or the people who are coming to the page and arms are the new stories right so I could have attributes that describe both the new stories on the arms I am sorry and the users and i can use all of these satellites to define my value function so I can have a parameterization that depends on both the actions and the state it apart from that I

could have additional parameterizations like, like this for defining my value functions and my and my policy's.

(Refer Slide Time: 9:39)



Whatever alright so and people talk about parameterized bandits typically they talk about parameter zing the value functions okay not necessarily the policy gradient kind of things even though this is also a form of parameterized bandits. When typically when people talk about parameterized bandits in the contextual bandit can kiss they are talking about parameter zing the value functions okay.

And just, just point out these are very close and sometimes they are used interchangeably like parameterized bandits or linear bandits. Where the, the function value function is assumed to be linear in the set of parameters that you have right so that is linear band it's right and contextual bandits are often use interchangeably but there is a certain difference contextual bandits is a problem setting while linear bandits is a specific way of solving that problem.

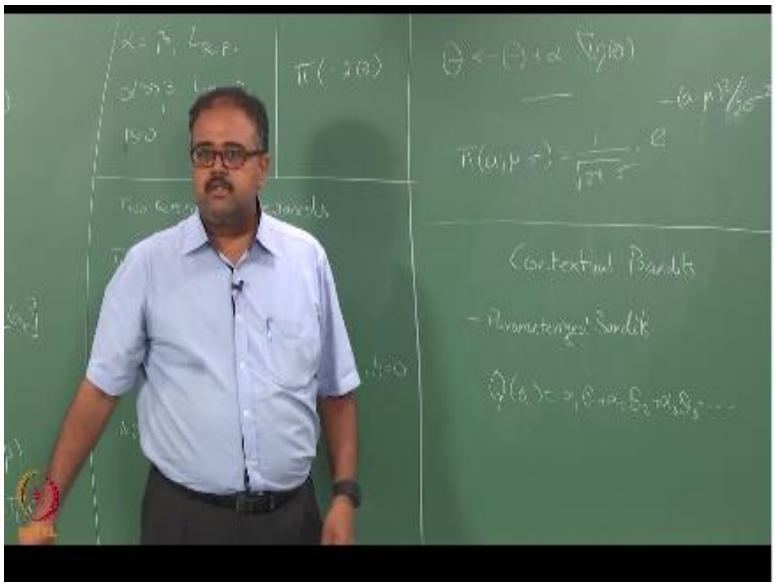
So contractual bandage is essentially like I said there are many ways in which you can solve it right you can just replicate the bandits you can do group and then solve we do not really have to assume some kind of parameterization but that's it so this again very, very popular now in fact

most deployments of bandits right use this kind of parameterized contextual bandits so for example the one of the most popular deployments of UCB.

Right only UCB has become the most popular bandit algorithm right but the most popular deployment of UCB we find is called Lin UCB which is UCB with a linear parameterization assume for the value functions so I will assume some set of parameters I will assume my Q of a right that is what I am estimating my cue of EA is some function of these parameters and that function is linear in the parameters right. So that's essentially what Lane you see we would do right and then you have to have some mechanism by which you determine what the upper confidence bounds are right.

So, it's a slightly different update rule then what we do for UCB but it is more or less the, the same principle right so I will be doing something like this should be something like this.

(Refer Slide Time: 11:48)



So θ1 θ2 θ3 will be my parameters I assume for the representation. So fine so I will stop here there will be the end of bandits. So next class I will start with chapter 3 in the book so if you want to read ahead please do so.

IIT Madras Production

Funded by
Department of Higher Education
Ministry of Human Resource Development
Government of India

www.nptel.ac.in