NPTEL

NPTEL ONLINE CERTIFICATION COURSE

REINFORCEMENT LEARNING

Thompson Sampling

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

There is only one thing which I want to mention more in every, you know very, very high level I do not want to get into the proofs and other things the guys who did the course last year actually had a guest lecture from somebody who works in this field and who was able to talk about another technique which is gaining a lot of popularity recently, so he talked about that in a gory detail in fact it did some proofs from one of the papers that they wrote and so on so forth but unfortunately I do not have the guest lecturer here this time and was also I mean just I introduce you that ideas to you and then if you want you can read up on those later right but in some sense.

Whatever we have done so far gives you all the tools that you need you know it gives you the basic tools for it understand more complex algorithms and so on so forth so that is the whole idea for me to do this and if you actually look at the book right the textbook. They almost in passing mention all of these algorithms mean they'll have like one paragraph saying oh this is how you do you see be and this is how you do they don't talk about meeting limitation at all this is how you do something else which i will talk about in the next class called reinforce and they mention these things in passing and then they say we will build on some of these ideas as and when required for the full reinforcement learning problem right.

So in some sense the book treats bandits as something that you is a necessary evil that you need to know about on your way to learning about the full RL problem right. Given the focus of the book that's a valid way of treating bandit problems but in fact it bandits itself is a very, very rich large literature and it is very active area of research for a lot of rain for running people. And this

one of the reasons I actually covered bandits in a little more depth. Then I used to do maybe a couple of years ago, right. So couple of years ago. I did bandits in like one class, right all of bandits. Anyway so we one thing which I want to talk about was,

(Refer Slide Time: 02:30)

Thompson sampling which is gaining a lot of attraction now okay, so it is more descriptively you could call it posterior sampling depending on which literature reading some people call it posterior sampling some people call It Thompson sampling so the basic idea is the following right, So I am trying to solve the Bandit problem right and we know that if you know that the parameters of the Bandit if I knew all the Q stars right I can solve it very easily right so how what do i do, take the highest right so what I am going to do when I do this kind of Thomson sampling approach is to, take a somewhat Belgian way of looking at these things. That I am going to say that ok I am solving this unknown problem, I am solving this unknown problem.

But I am going to make some assumption about the parameters of this unknown problem. Right so essentially and the assumption I'm making here is about the, Q stars I am going to make an assumption that the true Q stars come from some distribution right. I will start with some guess for this the 2Q stars come from some distribution ,right so if my if my if I know that my rewards

are limited between0 and 1 like for like many of the cases we have been assuming the rewards are limited between 0 and 1 so what would be an appropriate assumption for this Q star distribution .Beta distribution so people remember beta distribution, the guys who are in ml, the guys who are in ml at least should remember the beta distribution so the beta distribution is is a distribution that is limited to zero one right and take all kinds of weird Falls right and ,so this is typically a prior that prior distribution that use you whenever you have.

We are trying to find parameters of a, distribution where the parameter is limited between zero and one ok so the beta distribution is a very popular prior that is used as like that I make some kind of an assumption let us say that i do not know anything about the underlying problem right. Then what will I assume, uniform right you can assume a uniform distribution saying that I don't know anything so my Q starts can come from anywhere so now what do I do and one way of solving this problem is to say that given my current knowledge about the parameters It given my current knowledge about the parameters which arm should I play okay.

So that I may reward is optimized that is one way of thinking about it so what should i do then i should take each bandit combination right I should I should say that ok, ok the probability of it so Q star 1 is some value say 0.3 Q star 2 is 0.5 Q star3 is something else 1.2 like that so I have to make one bandit problem and then figure out okay what is your best term to play here and then take another bandage problem figure out what is the best term to play if you take another banded problem figure out what is the best term to play and for each of these Bandit problems.

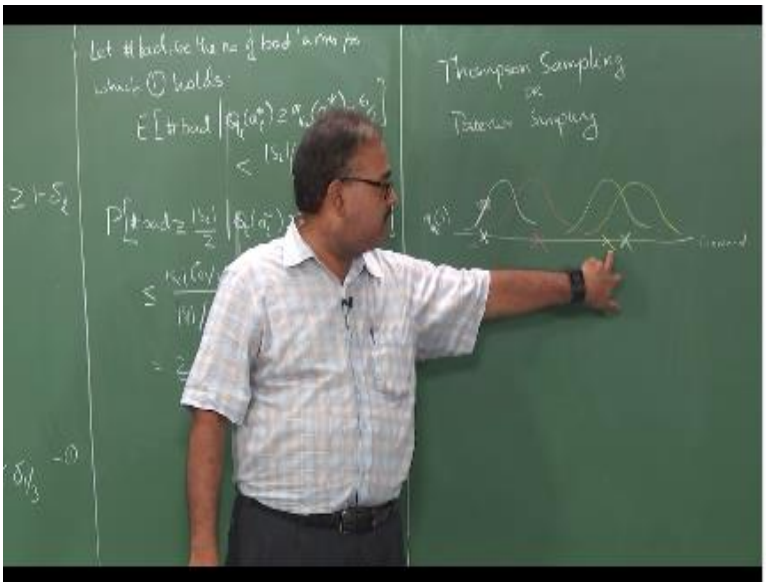
I should look at what is the probability that this is the true Q stars according to my current knowledge about the problem so if I start off with a uniform distributions and everything is equally likely right so all arms I mean can be optimal all arms can be non optimal and so on so forth so I basically have to play at random right but it turns out that doing this computation can be very, very cumbersome you can just think about it right I am asking you to figure out for each arm combination with what is the probability that the arm combination can occur right and depending on the nature of the probability distributions you are assuming this computation can become very, very complex right.

So the posterior sampling idea is a way of approximating this computation so what I do is the following is very, very simple okay. I have some distribution over what my Q star values will be have some distribution over by Q star values so what I do is I draw a sample from the distribution. Right so let us say something like this so my.

(Refer Slide Time: 7: 06)

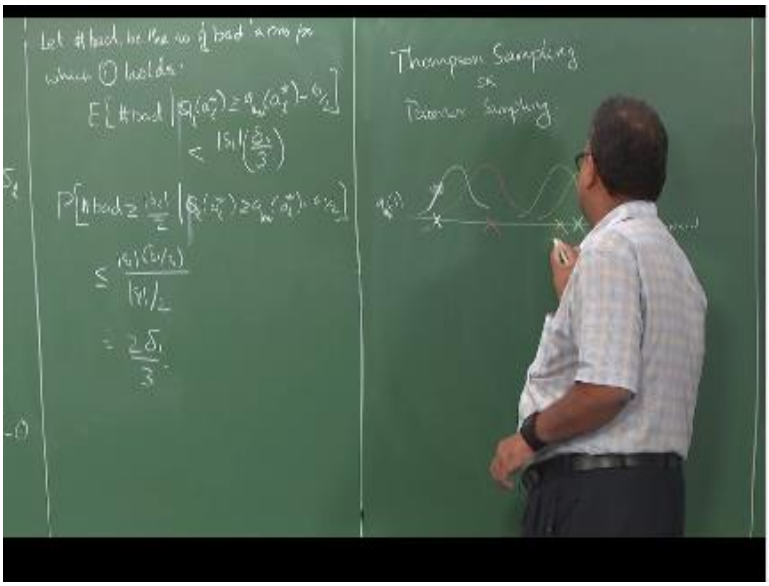
So Q star1 I am saying is something like this right and then, that Q star2. Tats Q star 3 .You have another color okay oh you're good four arms because running out of colors. Right so Q star 1 2 3 4 right this is the current distribution i have so i am assuming that the true value of Q star lies somewhere in this range, right and it is most probably this value but I do not know yet right there is some uncertainty about what the values are now what I do is, I draw samples you remembered we talked about drawing samples from Gaussians and other distributions I draw samples so I draw a so basically right so, this is basically my reward.

(Refer Slide Time: 8:38)



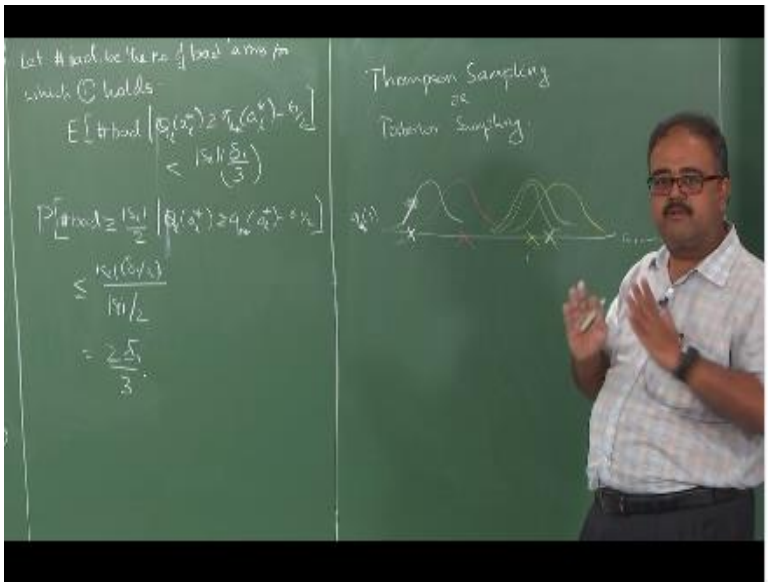
This my reward axis and these are the probability axis I draw a sample for that is one this is another that Is another Right. So now I have a specific bandit problem where my Q star 1 is this Q star 2 is this Q star 3 is this and Q star 4 is this so I have taken this distribution right and I have drawn one bandit problem from this distribution okay now I solve this problem the very basically pull on three because one thing gives me gave me the best value here so I pull on three now I look at the actual reward I get four on three.

(Refer Slide Time: 09:09)



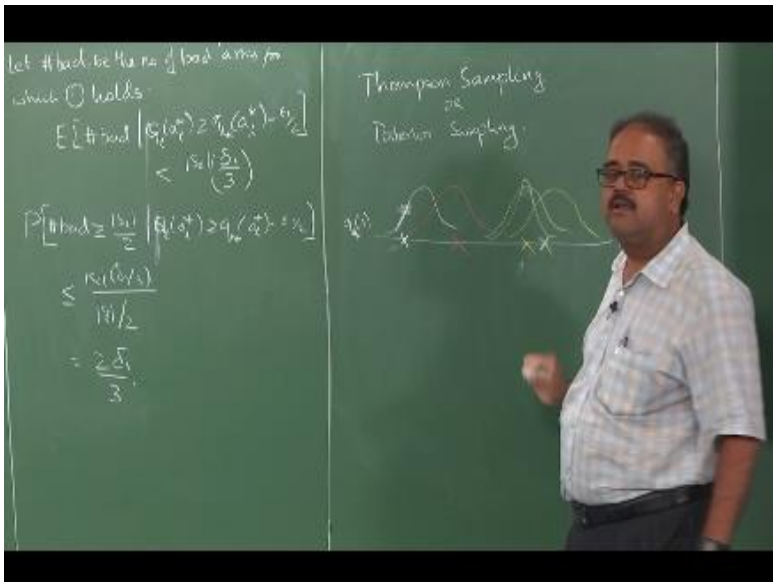
Let's I pull on three and let us say that my on three reward is that this is the actual reward that I get four on three and now I go back and update this fellow based on this reward right so what ,what do you think it should be.

(Refer Slide Time: 9:26)



Look um narrower because this is I kind of reinforcing my original belief that this is the most probable outcome.

(Refer Slide Time: 9:37)



So this will become my belief now this is my new distribution now again I try to sample so the probability that I will sample a high value for the green arm goes down now because I when I played the green arm I got a smaller vent so I keep doing this until my probability is kind of converged to the 2Q stars at the time in a sample a problem from there I will most probably be sampling that true problem right the actual problem underlying that which I am trying to solve this .This is called Thompson sampling or posterior sampling.

So this is getting a lot of attention recently mainly because people have been able to show that this can give you better regret bounce than you see be right so this can give you better regret bones than you see be and therefore there is a lot of attention that's being paid to Thompson sampling approaches. The analysis gets really hairy and in fact until recently there was no analysis of the Thompson sampling approach but more recently there have been several three four years there have been several papers that they have come up with new techniques for analyzing these kinds of algorithms different from what we have been talking about so far.

Kind around based right yeah. We are not eliminating any arms this what you are saying right well it turns out that you don't have to even without eliminating arms you are getting very good

complexity and typically what happens is after some time the posteriors become so focused right so the probability of you actually selecting the bad arm is gone I mean if you do not really need to do a round based computation there but yeah there might be a room for sorry after a few steps yeah I mean yeah this is possibility so the only thing is the only tricky part is once you do this kind of elimination glad to come back and analyze it.

So you saw how to keep this we came right I mean I gave you the right events to consider here right but to come up with these events is a tricky part right so once you come up with these events is easy to bound the probabilities and how do you know that this is the actual set of conditions that you should consider right for you to prove the whistle so that is the tricky part now once I say that okay here is a round based Thompson sampling approach then going and showing that it actually gives you the guarantees that you want becomes tricky I mean people just the plain Thompson sampling without any round basing itself resisted analysis for what a decade right and so, so it's not, not immediately clear.

Whether there is more interesting result that have come in yeah so people there is a very old style algorithm very world class of algorithms for solving bandit problems called learning automate and algorithm these are the people know about finite state machines right all of you know about finite state machines and have bunch of states and have some input signals symbols coming and then you jump around the states and so on so forthright.

So there is a class of finite state machines called variable structured finite state machines variable structure automata right and people have used variable structure automator to solve banded problems I am going to get into the details of it right so it's and it turns out that some of what the variable structure automate algorithms are doing is posterior sampling it turns out to be very similar to posterior sampling and very recently and earlier we knew results only for asymptotic convergence of this learning automaton algorithms.

So very recently like late last year October or November of last year people have shown that they also satisfy regret bonds they satisfy the block t regret bonds so that means a lot of interesting work that's happening in this phase and I mean of course If you're welcome to try a

round based a thousand sampling algorithm I don't know or we just check on arcade right the rate at which work comes out in this space and make me something already out there okay.

So stop here in the next class we will look at ways of solving the Bandit problem which do not depend on the do not depend on estimating the value function so everything you have looked on so far look finds the function Q right capital key so next class we look at one class of algorithms which do not look at maintaining the value function and directly try to learn the probabilities which you share to select the arms.

IIT Madras Production

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

www.nptel.ac.in