NPTEL

NPTEL ONLINE CERTIFICATION COURSE

REINFORCEMENT LEARNING

Thompson Sampling Recap

Prof. Balaraman Ravindran

Department of Computer Science and Engineering

Indian Institute of Technology Madras

Right I will try to figure out what is the model parameter right and then I will try to solve the problem corresponding to the parameter so if you think about I do not know how many of you did the ml course and if you remember the maximum likelihood estimate and maximum posterior map estimates and things like that we said we will find one estimate for the parameter θ right and then we solve the problem corresponding to that parameter θ right.

So in a very rough sense right so I have some data that comes to me I am trying to find out what is the process they generated the data right so some data is coming to me right so I have the huge collection of data from that I will try to figure out what is the process that generated the data and then try to answer questions about the process, right so some kind of expected value equations or whatever.

So translating that to the RL setting or especially the bandage setting so what is the question is okay here I have given you some data what is the data lot of arm pulls and rewards later given you some data based on this data figure out something about the parameters right figure out something about the parameters and then solve the problem post by the parameters right so if you think of the value function based methods that we talked about it the ε really kind of an approach if you think about it what do I do there.

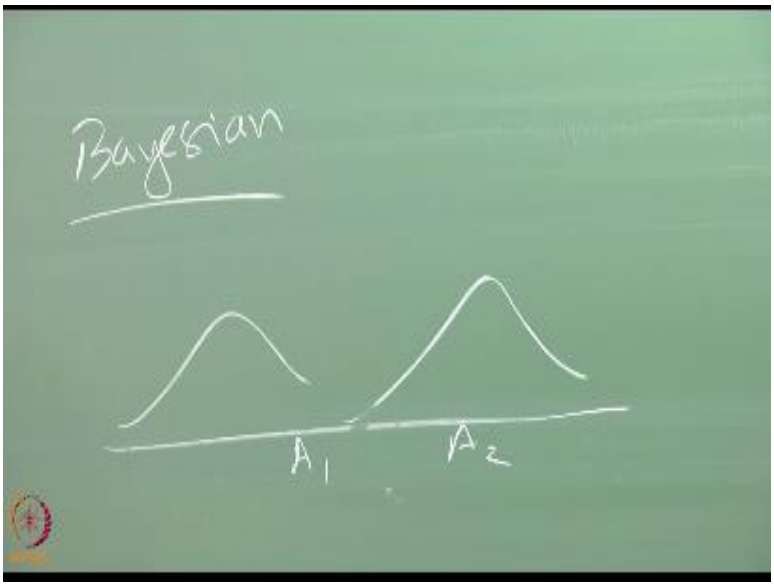
I take the problem right I take the data that is coming to me then I make a guess about what is the problem what kind of a guess do I make how do what is the guess it is a maximum likelihood gifts if you have no other assumptions about the arms so what I am doing there is essentially making a maximum likelihood estimate about the problem that I am solving I do not have no

other information right I ask you what is the possible reward for arm A basically I will take the average of all the rewards I got for arm A right that is the maximum like your estimate right alternatively you can think of doing a map version of this what will I do there I am going to make an assumption about the probability distribution or the prior means right.

I am going to assume that something about the priors of the means right and now I get some samples and based on those samples I will modify the prior I will get a posterior so and then based on the posterior I will find out what is the maximum which value of probability has the maximum value in the posterior maximum probability and the posterior right and then I will pick them there are two probabilities and talking about here.

The first probability is the probability of getting reverse and the second probability is the probability that parameter is the fight parameter okay this is the map estimate then I can do the full be in version of it so what is the full page in version of it, it start off with the prior the people understood what we are doing in map right so that is crucial if you did not understand what I did in map then you would not understand what I am doing here.

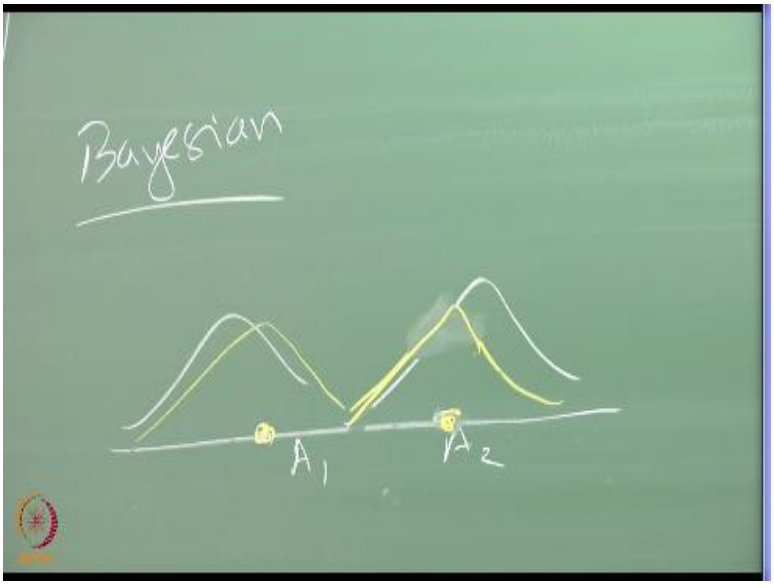
(Refer Slide Time: 03:18)



So in map let us say I am doing something like this right so I have let us say two arms I have two arm A1 and A2 and I am going to make some assumption about this is the probability of our arm A1 well this is the probability for the arm A2 about the mean μ, μ is what I am interested in right so I am going to assume this is the mean μ that is the mean μ and I am going to me this is μ1 this is a distribution for μ1 that is a distribution for μ2.

So what do I mean by distribution for μ1, I really do not know what μ1is, okay for a given problem μ1 and μ2 are numbers but if I do not know anything about the problem I do not know what that number is. I am going to assume that okay the real μ1 corresponding to a1 okay somewhere in this range right given my knowledge it is more likely to be here and less likely to be here and so on so forth so that uncertainty about my μ1 I am encoding it as a probability distribution okay.

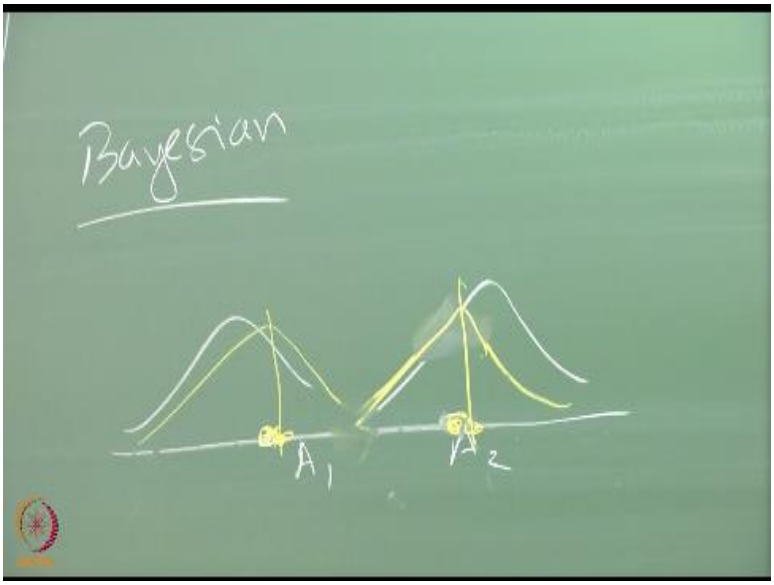
(Refer Slide Time: 04:16)



So now I pull arm A1 and I get a reward here right and let us say I pull arm A2 and a greater one here right so now what will I do?

I have to move my distributions right so I will maybe shifted slightly somewhat okay somewhat like that not necessarily the peak need not necessarily come that right but I move them slightly more to this side or that side now if I am doing the map estimate it what will I do it said I look at this distribution figure out which is the outcome that has the highest probability.

(Refer Slide Time: 05:11)



So this outcome has the highest probability this outcome as I as probability. I will take these two numbers to be my values for arm a1 and a2 that when I am doing maximum likelihood estimate I do not have any such distribution assumptions so I just keep pulling a one multiple times I will just take the average offset and put that as my estimate when I am doing map I have that assumption I take the one that has the highest thing I put that as my assumption and then I to this one, okay.

And bring a full machine estimate the trick here is to say that no I am NOT going to pick one value for my mean I am NOT going to pick one value for my mean I am going to assume that all these values are equal all these values are likely not equally likely all these values are likely and the likelihood is that distribution I given you right so one thing that you should remember is this is a distribution over the means right.

So even if I assume that the mean is this one okay when I pull arm a1 not that what I am Going to get there will be a distribution around that right so I can compute the expected value of the reward I am going to get assuming that this is the mean if I assume everything is Gaussian it is the same but you can assume different distribution so you will get different expected rewards right.

So if I pull if I say this is the this is the distribution over the means right and say the mean is here then I will get some expected reward with this as the mean right so there will be another distribution that is listening component let us make this thing as confusing as possible okay say suppose this is a mean I choose the true reward will probably be something like that right and that will have a different Sigma that will have a different variance.

It cannot be the same variance as this one that will be a different thing and that will not change because I am assuming the sigma are fixed for all the action say so it will be a thing like this and for this let me see if I assume this is the thing and it will be a similarly shaped curve centered around that now I can compute the expectations of that right and now I pull the arms based on whichever gives me the best expectation.

And then what is this expectation I am taking this is the expected reward that I will get assuming that may mean you could have come from a specific distribution starts becoming little complicated right I am assuming the mean could have come from a certain distribution and for each value of mean that I could have gotten there I am going to find what is expected reward and I am going to average it by the probability of that mean being character.

So that is what I mean by taking them the expected reward when the mean is coming from a certain distribution that Pick up certain reward value okay find out what is expected reward I will get 1 I am sorry I am going to pick a set a new value and find out what is expected reward will get for that μ value right and then what is the probability that the μ is correct so integrate overall μ right integrate.

So there will be two integrals SF 1 integral running over a fixed μ what is the expected reward there will be one integral then another integral running over all μ times the probability of that μ okay since little gets a little hairy right so instead of that what we do we come up with a way of approximating this expectation right so what I do instead of either taking the map estimate or trying to find the full expectation.

So I take the posterior I take the posterior I just draw a sample according to this and then I assume those two samples are the means of a1 and a2 I just taken one sample instead of computing the full expectation I just take one sample assume that a one in a to the sample I draw are the means right now I just go solve the problem according to that so bandit case is very easy just take whichever is maximum pull the door right.

Now I will get some data point I will go back and change my the posterior again right that distribution will again change now what I do I just go pick a random point according to that distribution keep that as a my μ 2 and again solve the problem I keep doing that so what you can show is right instead of computing this expectation in this very complicated double integral fashion we are doing this repeated sampling from the posterior.

I am in effect computing the expectation that is why this posterior sampling kind of this is called posterior sampling now this sampling is called posterior sampling so posterior sampling is effective because in effect you are trying to solve the Bayesian version of this wrong because there are more technical differences between the two but the intuition as to why we do this is this in pleases you can kind of get this intuition just by looking at maximum likelihood map.

And this so we do not have a really a map version of the Bandit problem as we have ml version right then we also have this kind of a Bayesian approach, see UCP already making some kind of an assumption about the distribution that is very able to derive the upper conference bound anyway right, like we are assuming the process that we can have a bad day for the system you can incorporate that in their learning and another hostility in you.

Because you can assume something about the distribution not if you want to start UCB with the prior is it well if you want to sure I am going to have a Bayesian UCB yeah so base UCB is yeah you could if you reduce a prior then essentially that is what you are going to end up, right more or less think about it, exp full version you incorporate expert axles in each time schedule outside Victor issue I think you are scaring the class.

We can discuss this afterwards yeah my god you are really giving my scared look be careful you might be incapacitated from writing the exam, so any other questions on this so I will try to figure out what is the model parameter right and then I will try to solve the problem corresponding to the parameter so if you think about I do not know how many of you did the ml course and if you remember the maximum likelihood estimate and maximum posterior map estimates and things like that.

We said we will find one estimate for the parameter θ right and then we solve the problem corresponding to that parameter θ right so in a very rough sense right so I have some data that comes to me I am trying to find out what is the process that generated the data right so some data is coming to me right so I have the huge collection of data from that I will try to figure out what is the process that generated the data and then try to answer questions about the process.

It is some kind of expected value equations or whatever so translating that to the RL setting or especially the bandage setting so what is the question is okay here I am giving you some data what is the data lot of arm pulls and rewards right I have given you some data based on this data figure out something about the parameters right figure out something about the parameters and then solve the problem post by the parameters right.

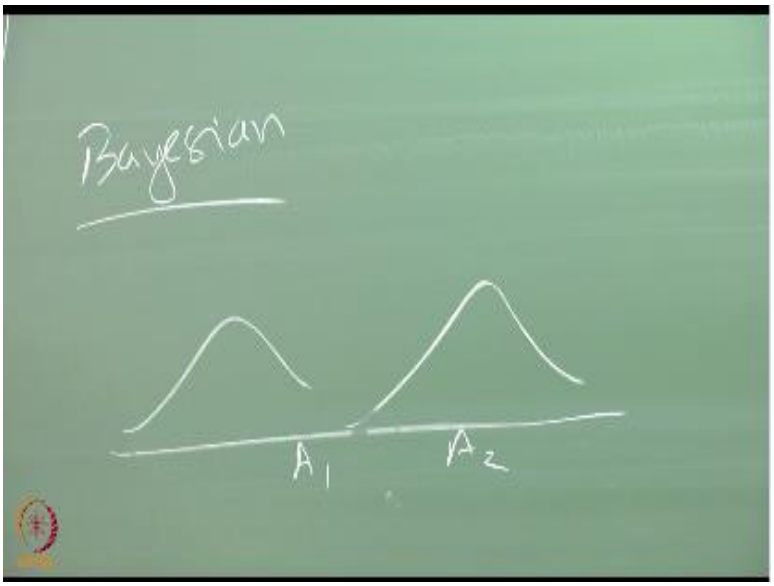
So if you think of the value function based methods that we talked about it the ε really kind of an approach if you think about it what do I do there, I take the problem may take the data that is coming to me then I make a guess about what is the problem what kind of a guess do I make how do what is the guess that is a maximum likelihood guess if I have no other assumptions about the arms.

So what I am doing there is essentially making a maximum likelihood estimate about the problem that I am solving I do not have no other information right I ask you what is the possible reward for arm A basically I will take the average of all the words I got for arm A that is the maximum likelihood estimate right alternatively you can think of doing a map version of this what will I do there I am going to make an assumption about the probability distribution or the prior means right.

I am going to assume that something about the priors of the means right and now I get some samples and based on those samples I will modify the prayer I will get a posterior so and then based on the posterior I will find out what is the maximum which value of probability has the maximum value in the posterior maximum probability and the posterior right and then I will pick them there are two probabilities I am talking about here.

The first probability is the probability of getting rewards and the second probability is the probability that that parameter is the right parameter okay this is the map estimate then I can do the full Bayesian version of it so what is the full Bayesian version of it, I start off with the prior the people understood what we are doing in map right so that is crucial if you did not understand what I did in map then you want to understand what I am doing here. So in map let us say I am doing something like this so I have let us say two arms.

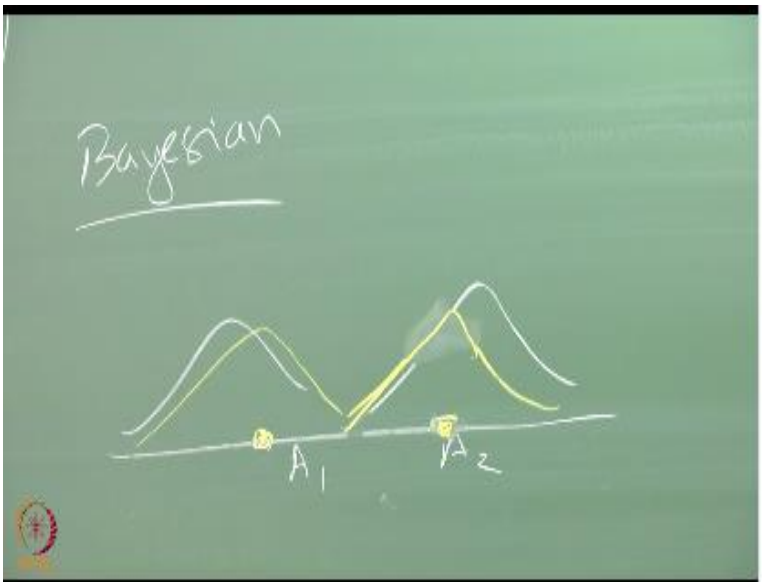
(Refer Slide Time: 15:16)



I have to UM say a1 a 2 and I am going to make some assumption about this is the probability of our arm a1 this is the probability for arm a2 about the mean μ2 so μ is what I am interested in right so I am going to assume this is the mean μ that is the mean μ and I am going to me this is μ1 this is a distribution form you one that is a distribution form you too so what do I mean by distribution for μ1 I really do not know what μ 1 is okay, for a given problem μ1 and μ2 are numbers.

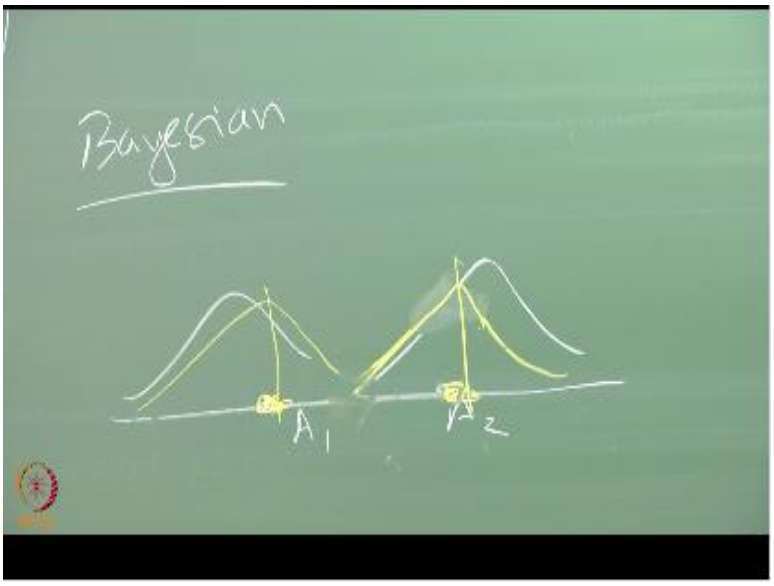
But if I do not know anything about the problem I do not know what that number is I am going to assume that okay that real μ1 corresponding to a1 okay somewhere in this range right given my knowledge it is more likely to be here and less likely to be here and soon so forth so that uncertainty about my μ1 I am encoding it as a probability distribution so now I pull arm a1 and I get a reward here right. And let us say I pull arm a2 and I get a reward here right, so now what will I do.

(Refer Slide Time: 16:21)



I have to move my distributions right so I will may be shifted slightly somewhat like that not necessarily the peak need not necessarily come that right when I move them slightly more to this side or that side.

(Refer Slide Time: 17:06)



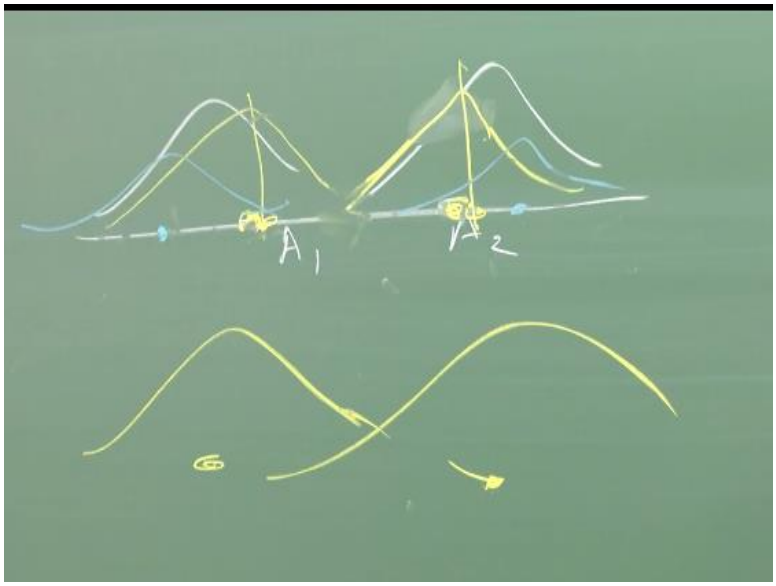
Now if I am doing the map estimate what will I do he said I look at this distribution figure out which is the outcome that has the highest probability, so it is this outcome has the highest probability this outcome has highest probability I will take these two numbers to be my values for arm a1 and a2 that when I am doing maximum likelihood estimate I do not have any such distribution assumptions.

So I just keep pulling a one multiple times I will just take the average offset and put that as my estimate when I am doing math I have that assumption I take the one that has the highest thing I put that as my assumption and then I to this pulley okay I am doing a full Bayesian estimate the trick here is to say that no I am not going to pick one value for my mean I am not going to pick one value for my mean.

I am going to assume that all these values are equal all these values are likely not equally likely all these values are likely and the likelihood is that distribution I given you right so one thing that you should remember is this is a distribution over the means right, so even if I assume that the mean is this one okay when I pull the arm a1 is not the word I am going to get there will be a distribution around that right.

So I can compute the expected value of the reward I am going to get assuming that this is the mean if I assume everything is Gaussian it is the same but you can assume different distribution so you will get different expected rewards right so if I pull if I say this is the this is the distribution over the means right and say the mean is here then I will get some expected reward with this custom mean right so there will be another distribution that is listening let us make this thing as confusing as possible okay.

(Refer Slide Time: 18:55)



Say suppose this is a mean I choose the two reward will probably be something like that right and that will have a different Sigma that will have a different variants it cannot be the same variance as this one that will be a different thing and that will not change because I am assuming the sigma are fixed for all the action say so it will be a thing like this and for this let me see if I assume this is the thing and that will be a similarly shaped curve centered around that.

Now I can compute the expectations of them right and now I pull the arms based on whichever gives me the best expectation and what is this expectation I am taking this is the expected reward that I will get assuming that my mean μ could have come from a specific distribution starts

becoming little complicated right I am assuming the mean could have come from a certain distribution and then for each value of mean that I could have gotten there I am going to find what is expected reward and I am going to average it by the probability of that mean being character so that is what I mean by taking them the expected reward when the mean is coming from a certain distribution that pick a certain reward value okay find out what is expected what I will get I am sorry.

I am going to pick a set a new value and find out what is expected reward will get for that μ value right and then what is the probability that the muse correct so integrate over all μ right integrate so there will be two integral size of one integral running over a fixed μ what is the expected reward that will be one integral then another integral running over all news times the probability of that μ okay.

Since little gets little hairy right so instead of that what we do we come up with a way of approximating this expectation, so what I do in stuff either taking the map estimate or trying to find the full expectation so I take the posterior I take the posterior I just draw a sample according to this and then I assume those two samples or the means of a1 anda2 I just taken one sample instead of computing the full expectation I just take one sample assume that the a1 and a2 the sample I drew are the means right.

Now just go solve the problem according to that so bandit case is very easy just take my server is maximum pull the door right now I will get some data point okay I will go back and change me the proposed to you again write that distribution will again change now what do I do I just go pick a random point according to that distribution keep that as my μ 1 and keeping another point keep that as μ2 and again solve the problem.

I keep doing that so what you can show is right instead of computing this expectation in this very complicated double integral fashion but doing this repeated sampling from the posterior I am in effect computing the expectation that is why this posterior sampling kind of this is called posterior sampling thumbs and sampling is called posterior sampling, so posterior sampling is effective.

Because in effect you are trying to solve the Bayesian version of this wrong because there are more technical differences between the two but the intuition as to why we do this is this in pleases we can kind of get this intuition just be looking at maximum likelihood map and this so we do not have a really a map version of the Bandit problem that we have ml version right 10 we also have this kind of a Bayesian approach.

IIT Madras Production

Funded by

Department of Higher Education

Ministry of Human Resource Development

Government of India

www.nptel.ac.in