**NPTEL**

**NPTEL ONLINE COURSE**

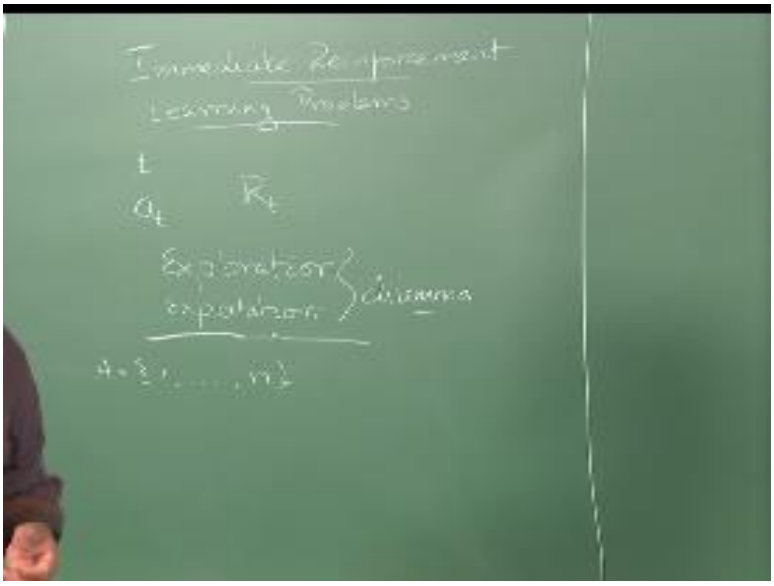
# REINFORCEMENT LEARNING

## Introduction to Immediate RL

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

So what we will start looking at today is what.

(Refer Slide Time: 00:24)




Right so what we call immediate RL immediate reinforcement having problems so if you remember I said there are many, many characteristics of a reinforcement learning problem and one of those was that you get this evaluation from the environment and could potentially be delayed right so I said you might actually lose a piece very important piece in the middle of a

chess game but you get a reward only at the end it is not like the final move you made it someone that causes you to lose it could be somewhere along the way right.

So what we are going to do now is to actually consider a much simpler version of the problem where we do not have this the temporal nature to it right, so we are not we are not going to assume that the rewards are the scalar evaluation that comes from the environment is going to come at a further time than when the action is actually performed right so I am going to assume that as soon as you perform an action like you are going to get a evaluation right and that is it basically right.

So we are going to consider a much more simplified version of the problem that is at every time instance I am required to pick an action right, so every time instant T right I pick an action $a_t$ right, and then I receive a reward $R_T$ and that is it my interaction is ended right. So I can learn something from this interaction and go back and pick a another action at time t + 1 right so that is why we call this immediate reinforcement learning problems because the outcome is immediate so as soon as if take an action I get a reward right.

So what else is missing in my description here there is no real input to the system right there is no state signal that the system is seeing so all the input it is getting are the rewards right there is no state input to the system right, so it is like it's just solving the same problem again and again right so instead of thinking of it as not having a state you could think of it as having like every time I give it the same input and I am asking you to give me a action in response to that this is seem like a reasonable thing to have you think of cases where this might actually be an appropriate model.

So one example which people give us in drug testing right so I have I have a certain symptom that is presented to me right and I can choose on one of four different drugs to treat that symptom right so I first I give Doug one it may or may not cure the person if it cures the person then I get a reward if it does not cure a person I do not get a reward right but it could very well be that for a variety of reasons the person did not get cured it is not because that the drug one was not the right give for this person right.

For this symptoms but for a variety of other reasons the patient might not have been cured so I cannot conclude by just administering the drug once that the drug is ineffective for a set of symptoms and I have to keep trying the drug multiple times right before I can come to a conclusion that the drug is ineffective or a person okay alright so for a certain set of symptoms so it is like saying that okay there are multiple you know samples that I have in a lab and each of them is exhibiting the same disease.

Let us not talk about human beings administering drugs and killing them or something can think of what samples in the lab right I'm going to add some drugs to that and see if that particular organism dies right in that in the when I add the drug and so far it is essentially the same presentation as far as the drug testing is goes right it is the same organism that I am trying to kill say microorganism that I am trying to kill across all of these different instances it is like the same state that is being presented to me over and over again but then I have multiple choices that I could make and as soon as I do the choice I get a feedback saying okay.

Yeah the organism died or it did not die right, so I have to keep trying this multiple before I figure out which is the right drug to administer okay good it clear so that gives you an example right so there are a couple of things which I kind of rolled into their example as well so one of the is this that the outcome can be stochastic right, so every time I take an action right so I did not be given the same evaluation back right so I take a tissue sample added up to it right so i might not succeed every time right so sometimes I will succeed sometimes I might not succeed so that is some kind of a stochastic nature to the outcome right.

So as a consequence of the stochastic nature what did you have to do I have to repeat things multiple times right so this is these are things which you have to keep in mind when we are talking about this kind of immediate reinforcement learning problems that we are going to have a stick outcomes and then we have to think about repeating choices multiple times before we can be sure of the outcome right is there anything else that was that could have been critical to the successful outcome of this let us say that I tried drug one okay.

And the organism survived okay, so then I just can I keep using drug one for all the other organisms as well let us say let us keep the problem simplified I just will say heavily to drugs drug or drug too okay I take the first organism give it drug one its advice I take the second organism give it to does not survive okay, now can I conclude that which one is more effective during go in order to win do not know me assuming is oneself the survival what we are looking for not survey well what we are looking for not survival talking about microorganisms he is right I actually flipped it around I do not know but nobody actually noticed it.

So yeah so I am looking for non survival of a microorganism right so in this case drug to appear to be more effective than drug one right so that does not mean I can keep giving drug to all the other samples and I need to actually go back and try drag one a few times right so this is the exploration and I was telling you about exploration in the previous class size this is the exploration here that I have to do exploration so the critical thing here that even the simple case of this immediate reinforcement learning introduces to is the need for exploration right but can I just keep exploring can I just keep giving random drugs to each of the samples and at some point I have to stop right.

I am interested actually in killing this microorganism today at some point I have to stop and I said okay here this is the actual drug that is going to kill the microorganisms and I have to start giving that drug right this did not be an explicit point where I stopped we will see as we go along but I am interested in eventually exploiting the knowledge that I am acquiring about this the whole system right, so either I explore or I have to at some point I have to stop and start exploiting so there is his inherent tension between exploration and exploitation okay.

They are in some sense at loggerheads with each other right if I stop at some point and say okay I have know enough about the system and then I start exploiting then I necessarily have to stop my exploration right but then if I do not explore enough I will be exploiting a poor solution right so this exploration exploitation so sometimes people call it a conflict sometimes it is called the called the Explorer exploit dilemma right so too should I explore Philly explain so that is essentially what this immediate reinforce will any problems captures for okay.

Since this is a very crucial question in reinforcement learning so we will look at this immediate reinforcement learning problems in a little bit of detail so that it will motivate some of the other things that we will do later when we move to the full RL problem right and so what I will do which is slightly different from what is there in the textbook, so how many of you have actually looked at the textbook good see I did not ask how many of you have read the book right. So what I will do the slightly different from the book and I am still not decided whether it will be for both versions of the course let us talk a little bit more about some of the critical questions that people are interested in addressing in the immediate RL setting right.

It is a book just gives you a very algorithmic very prescriptive viewpoint of this right this is or here is how you solve the immediate reinforcement learning problem right it does not go beyond that it does not give you anything about right so why do people solve it the way they do it I will go a little bit beyond that and try to give you some way intuition in that mainly because that seems to be a very active area of research nowadays in reinforcements and many of the ideas that people are using there are also being transferred to see the full reinforcement learning setting.

The one with all the delayed rewards the states and things like that right so lot of ideas from the research in this immediate RL setting is also being used to the full URL setting so I want to introduce you to those topics right, yeah so despite the makeup of the class right just predominantly undergrad this is a graduate level course rate so I want you to want to introduce people to research questions and other things Sarah mean demands for I to actually teach an advanced trouble codes at some point I know at least one person who'll take it I am going to try this.

Now every class I am going to make people turn and look at you at least once I manage to do that last class anyway so go back so we will look at a little bit of formalism on this right so what I am going to assume is that I have a set of actions A which will index by 1to n so action one could be anything right and this is just giving the arbitrary numbers 1 to n it could be drug one drug to drug three it or it could be add one add two add three whatever it is just one point is what we will denote it as right.

And each action when I pick that action it gets a payoff it is an evaluation so I am going to be using these terms interchangeably right reward payoff evaluation cost so all of this denote the same thing that dotted line that came in my box diagram right so depending on what literature you come from you use different by different names for it right so the reward typically comes from guess psychology it cost comes from control theory optimization yeah payoff comes from econ economics and what is the last thing I say evaluation comes from optimization okay fine.

So depending on which, which background you come from you give it different names right and so I will use them interchangeably for example for most of this literature I will be using payoffs because a lot of the work has been done in the in the o, r and economics community in when they are trying to analyze these kinds of problems right so when you pull each when I actually select an arm I select an action right what's going to happen is that I will get a reward or a payoff sample from some unknown distribution so what do I mean by that stochastic but what do I mean more I mean it can right this is something which I found people have difficulty so some people at least have difficulty when I say it is sample from a distribution so what I mean by mean is the following.

Let us  keep it simple so that people can conceptualize it so associated with each action let us say that I have a coil associated with each action there is a coin okay that has some probability of coming up heads and some probability of coming up tails so let us some probability P of coming up as they say pick action one then there is a probability P one of coming up heads when it toss the coin right and I will say that if the coin comes up heads ok I will give you a payoff of one if the coin comes up tails I will give you a payoff of 0 right.

So every time I select an action I am going to toss this coin so depending on whether it comes up heads or tails I will give you either a one or a zero right so the payoff that I will get every time I select that action is not the same right but what is same the coin right if I select actual action one I will always toss coin one if is elect action 5 and always toss coin five right I am not changing these coins okay, is it clear does it make sense to people and I will not change the coins but the coins will be the same right.

But every time I pick action I will toss that coin and I will give you the payoff or reward corresponding to the outcome of the coin toss so if I say that the probability of reward for picking action 1 is 0.8 what does it mean a probability of heads of the associated coin is point eight right in this case I can go further I can also say something else I can say that the expected payoff of picking action one is point eight as well right because the probability pointed will get a reward of 1 with the probability point2 will get a reward of 0 so if I take the expectation of this it will get will be point eight right.

We will understand this now so when they say select an arm select an action I will tell you why I keep calling it off in a second and then I will keep calling it up so when I select an action right then I am going to toss a coil right depending on the outcome of the coin I am going to give you one this is something which you should remember when you are actually going and implementing these things right so I had people do all kinds of weird things right and make sure that you remember what I mean when I say you sample from a distribution okay great.

So now I gave you the simple example of tossing a coin I could actually have any complex distribution right so if a toss a coin what is the distribution I am sampling from enrolling right I could have other kinds of distributions I could have a Gaussian distribution let us say so now what does it mean for every action I choose okay.

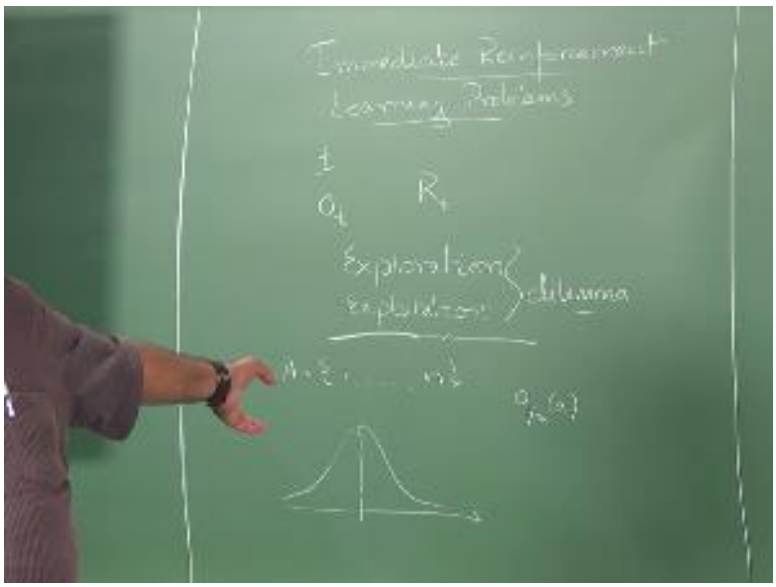
So for action when that is going to be a like some kind of a Gaussian distribution right and so what is the axis here x axis here the payoff the payoff for the reward x axis the payoff for the river and the y axis is the probability that you will see such a payoff right so that is the most probable pay off as it also happens to be the forgotten this also happens to be the expected payoff right so how will I sample from a Gaussian they can add a number and 0 and anyone else have a sample from a Gaussian okay.

How do you do that people who are in my minor I told you how to do this in the class okay little bit more efficient we have to get yeah and then what you do? Do not tell me you smite lab no man it is your role you know too many terminology just keep throwing them around go to CMC stand for in then there are two emcees then you just told me a Monte Carlo I said give me

something more efficient and then he keeps saying another Monte Carlo method what is it no one oh come on, fine okay Paulo you tell me, so the guys who were in the guys were there in my minor rights I actually told you how to sample me remember I said so I did this with a discrete distribution and then I told you how to do this with continuous distributions as well can you give the me write-up on model okay.

So basically the idea is you generate a random number right between 0 and 1 generate a uniform random number between 0 and 1 right and then keep integrating the area under the curve what is it called the CDF right and then when the CDF reaches that random number you generate so that is the number you sample does it make sense, so this is a complete aside.

(Refer Slide Time: 19:24)



Let us say that I have four possible outcomes right I want to generate one of those four outcomes at random right so the probabilities are $p_1$ $p_2$ $p_3$ $p_4$ how will I do that I generate a uniform random number between 0 to 1 right, no so I will see if the number is less than $P_1$ if the number is less than $P_1$ I will generate outcome 1 if the number is between $p_1$ and $p_2$ and generate the outcome of two this number is between $p_2$ and $p_3$ l generate the outcome of three generate number is between $p_3$ and $p_4$ is it correct or not, it should be between $p_1$ and $p_1 + p_2$ okay it should be between $p_1$

plus $p_{2a}$nd $p_1$ plus $p_2$ plus $p_3$ so it should be the cumulative sum right it should be the cumulative some previous cumulative sum and the next cumulative sum.

So if it lies between that then that is the outcome I will tend right so is less than $P_1$ I will generate one if it is between $p_1$ and $p_1$ plus $p_2$ I will generate two if it is between $p_1$ plus $p_2$and $p_1 + p_2$ plus $p_3$ I will generate three if it is between $p_1 + p_2 + p_3$ and one I will generate four right do not tell me if $p_1 + p_2 + p_3 + p_4$, so you can simulate the CDF for the question right yeah so there are ways of doing it right so there we have doing this so you can do this for any probability distribution for which you can define the CDF.

Not just Gaussian it's not just discreet so the discrete case it is very clear right so what you are doing is essentially cumulate adding up the probability so whenever you hit that number you say that is the Nazi sample that you have drawn right in this case you just do this oh my gosh she is okay so this is what I mean when I say sample from a probability distribution again it is not the most probable outcome like a nice a sample from this Gaussian I have actually seen people implement okay the reason I keep reiterating this people actually form the Gaussian and then keep returning the value that has the highest probability right.

So please do not do that of course if you are using MATLAB or something like that you don't even have to think about it we can just say Rand N and then it will give you a sample from a normal distribution which is what she was going to tell you what is the way to sample from a Gaussian but this is what it does internally wait and to give you the sample having said this now you can go back to using random yeah we had a question okay fine so going back associated with each of these actions right I am going to have some distribution from which the payoff Is being sampled okay is it clear.

And these distributions I am going to assume at least initially will be fixed for the duration of the experiment for duration of the learning this distributions will be fixed like I said you do not change the coin you don't change the Gaussian also okay now if I knew the Gaussian then or if I knew the coin send you the probability with which the coin will turn up tails heads right so how will I solve the problem there is no problem right I just picked the action that has the highest

mean right highest mean and for the Gaussian or the highest p 1 for the coins right whatever it is the highest expected payoff I just picked a topic that action right.

So the challenge comes when I do not know the distributions all I know is that is some distribution which I am assuming is fixed according to misty payoffs are generated I do not know the distribution so I basically have to figure out the distribution and then pick the pick the action I picked action corresponding to that right so is it clear so for each of these actions right for each of these actions I am going to assume that that is a true expected payoff which will denoted by Q* of A.

Let us say it's action so Q* is a function that will give me the true expected payoff of action a right so in the case of the coins Q star of one will give me p 1 this is a probability of heads that Q star of 2 will give me $p_2$and so on so forth in the case of Gaussians Q* of one will give me μ1 let you star of 2 will give me Bo 2and so on so forth and see what the Q* function is this is again unknown to me and if I know Q*I am done right I am a problem is solved right but this the assumption we are going to make so we have a set of actions A and we have a set when corresponding Q* that we do not know about right.

**IIT Madras Production**

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

**www.nptel.ac.in**