

**NPTEL**  
**NPTEL ONLINE COURSE**  
**REINFORCEMENT LEARNING**

**Introduction to RL**

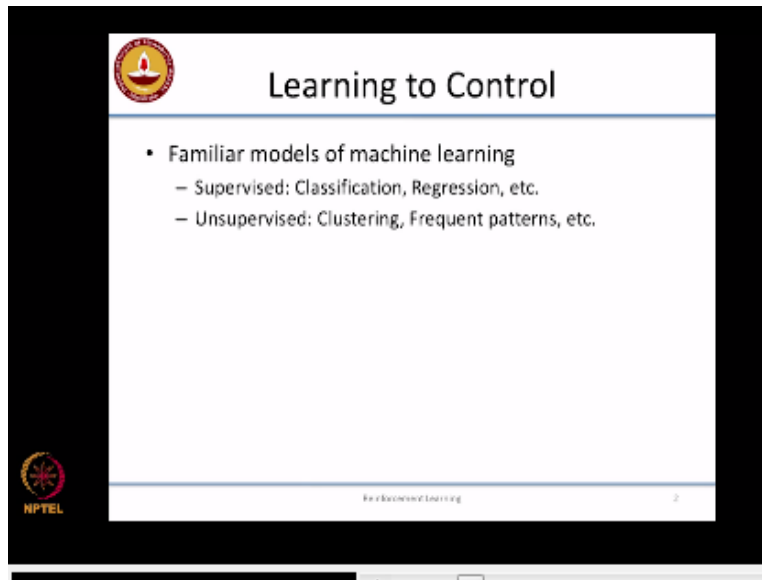
**Prof. Balaraman Ravindran**

**Department of Computer Science and Engineering**

**Indian Institute of Technology Madras**

So good, so we can finally get underway, so this is CS 6700 reinforcement learning, if anyone is here by mistake looking for the planning class still well you should leave okay, and yeah so how many of you were in the machine learning course just for me to get a sense large fraction of the people were enable and so, this is a very the different kind of learning than what we looked at it am I right? So in machine learning we looked at familiar modes of machine learning where the idea was to learn from data, you know so you had given lot of data as training instances for you and essentially we're trying to learn from those training instances as to what to do right.

(Refer Slide Time: 01:19)



And there were different kinds of problems that we are looking at, so one was supervised learning problem in which you are looking at classification and regression, oh yeah so in the machine learning class we looked at learning from data right primarily so one of the models we looked at was supervised learning it where we learnt about classification and regression the goal there was to learning from an input space to a output which could be a categorical output in which case in classification could be a continuous output in which case it was called regression right.

So if you haven't been in the ml class don't worry about it right because this is just to tell you that our RL is not what we learnt in the ML class okay. So she hasn't learned anything in the ML class then you don't have anything to unlearn. So don't worry so the second per second kind of learning thing we looked at was unsupervised learning where there was really no output that was expected of you right since therefore there was no supervision the goal was to find patterns in the input data. I will give you lot of data points you can you find out if there are groupings of similar kinds of data points can i divided them into segments right.

So that kind of a thing was called clustering, right or you are asked to figure out if there were frequently repeating patterns in the data right and so this is called frequent pattern mining or derived problem there was Association rule mining and so on so forth right so people have heard

me give this analogy multiple, multiple times before but this is the most apt one how did you learn to cycle, right. so was it supervised learning so how did you learn to cycle, somebody who hasn't heard me or then who hasn't been in ML, you haven't been in ML right okay, how did you learn to cycle, was did somebody tell you how to cycle and then you just followed their instruction?

Okay first of all tell you know how to cycle? Yes. Do you know how to cycle, yeah you? Yes. Okay how did you learn to cycle fell down a couple of times and that automatic you made cycle, you have to actually figure out how to not default down, right. So falling down alone is not enough but you have to try different things right it's not supervised learning is really not supervised learning how much everything because, now that I have given this talk multiple times people are getting wise to it, right earlier when I used to ask these people uses of course in supervised learning mantle was there holding me or my father was telling me what to do and so on so forth.

At best what did they tell you, hey look out look out don't fall down, right. So that doesn't count our supervision, so or keep your body right keep your body up or something some kind of very vague instructions was what they're giving you right, supervised learning would mean that so we get on the cycle somebody tells you okay, now push down with your left foot with three pounds of pressure right and move your center of gravity three degrees to the right, right. So this is I mean somebody has to give you exactly what is the control signals that you have to give to your body in order for you to cycle right, then that would be supervised learning.


If somebody actually gives you supervision in that scale you would probably have never learned to cycle we think about it because it's such a complex, complex dynamical system if somebody gives you control at that level gives you input at that level you will never learn to cycle. And so immediately people flip and say that it was unsupervised learning right, because here of course nobody told me how to cycle therefore it is unsupervised learning. So if it is truly unsupervised learning what should have happened is you should have watched hundreds of videos of people cycling figure out what is the pattern of cycling that they do okay, and get on a cycle and reproducing right.

So that is essentially what unsupervised learning would be you just have lot of data right and based on the data I figure out what the patterns are and then you try to execute those patterns that does not work right, you can watch hours and hours of somebody playing fly simulator you cannot go and fly a plane wait, so you have to get on the cycle yourself and you have to try things yourself right so that that's the crux here right. So what it's its how do you learn to cycle is neither of the above right is neither supervised nor and supervised it's an it's a different paradigm.

So the reason I always start out my talks not just in the class but in gender when I talk about reinforcement learning is because people will always talk about reinforcement learning and unsupervised learning and it's always really irks me because it's not an unsupervised learning just because you don't have a classification error or a class label doesn't make it unsupervised learning it is a completely different form of learning and so reinforcement learning is essentially this mathematical formulas for this trial and error kind of learning right. So how do you learn from this kind of minimal feedback you know, falling down hurts or somebody your mom or somebody stands there and claps, when you finally managed to get on the cycle you know that's kind of a positive reinforcement.


Where did he fall down you get hurt right, that's kind of a negative feedback how do you just use this kinds of minimal feedback and you learn to cycle then, so this is essentially the crux of what reinforcement learning is about trial and error right. So the goal here is to learn about a system through interacting with the system right it's not something that is done completely offline okay you have some notion of interaction with the system okay and you learn about the system through that interaction. Reinforcement learning originally was inspired by behavioral psychology right.

(Refer Slide Time: 07:05)



## Reinforcement Learning

- A trial-and-error learning paradigm
- Learn about a system through interaction
- Inspired by behavioural psychology!
  - Pavlov's dog



Reinforcement Learning 3

So one of the earliest reinforcement systems that were said he was a Pavlov's dog, how many of you know of the Pavlov's dog experiment? what is the Pavlov's dog experiment in association with whenever he's going to do okay, so whenever like he just utterly need above the dog started expecting the food and started salivating yeah so that is called a condition perfect, so when the dog looks at the food and start salivating it is a primary response because that is a reason for it to salivate on the site of food in any idea why? Exactly so it is preparing to digest the food you know and then show the food is preparing to digest the food so it starts salivating right.

So then now if you think about it right hearing the bell and it celebrates what is it doing preparing to digest a bell? So when you ring the bell and then serve the food the dog forms an association between the bell and the food right and later on when you just ring the bell without even giving the food the dog starts salivating in response to digesting the food that it expects to be delivered right so essentially the food is the payoff you know the food is like the reward for it and it has learned to form associations between signals in this case which was a bell like an input signal which was the bell and the reward that is going to get right.

So this was called behavioral conditioning right and so inspired by these kinds of experiments on the more complex behavioral experiments on animals and people started to come up with

different theories to explain how learning proceeds right. In fact some of the earlier reinforcement learning papers appeared in behavioral psychology journals right, the earliest paper by Sutton and Border appeared in brain and behavioral sciences Journal, just go back I needed to need to say something about Sutton- Border know that there's a larger audience you can tell that what then so the we're going to follow a textbook written by Richardson and Andy border right but more importantly they are also kind of the co-founders of the modern field of reinforcement learning right.

So in 1983 they wrote a paper adaptive neuron like element that learn control behavior from something to that effect right and that essentially kick started this whole modern field of reinforcement learning. So the concept of reinforcement learning like I said goes back to Pavlov and earlier right, people have been talking about those kind of behavioral conditioning and learning and sure, but the whole modern computational techniques that people use in reinforcement learning are started by Sutton and border. So what is reinforcement learning right, so it's learning about stimuli right the inputs that are coming to you and the actions that you can take in response to it right, learning about the stimuli only from rewards and punishments so you're not going to get anything else food is a reward right falling down and scraping your hand there's a punishment, right.

(Refer Slide Time: 10:25)



## What is Reinforcement Learning?

- Learning about stimuli and actions based on rewards and punishments alone.
- No detailed supervision available
- Trial-and-error learning
- Delayed rewards
- Sequence of actions required to obtain reward
- Associative learning required
  - Need to associate actions to states
- Learn about policies not just actions
- Typically in a stochastic world



Reinforcement Learning 6

So only from these kinds of rewards and punishments alone right, there is no detailed supervision available nobody tells you what is the response that you should give to a specific input right, suppose you are playing a game there are multiple ways in which you can learn to play a game right so we can learn to play chess by looking at a board position right and then looking at a table that tells you for this board position this is the move you have to make right and then you go and make the move alright. So that is a kind of supervision that you could get you know that gives you a kind of a mapping from the input to the output right, it gives you a mapping from the input to the output and essentially you learn to generalize from that.

So this is what we mean by detailed supervision so, another way of learning to play chess is just you have an open-ended sit in front of him and you just make a sequence of moves at the end of the move you win okay, you get a reward make somebody pays you say 10 rupees, okay if you lose you how to pay the opponent and ten rupees that's all, that's all that happens is all the feedback you're going to get right, whether you are going to get the 10 rupees or going to lose the 10 rupees at the end of the game so nobody tells you given this position this is the move you should have made that is what we mean by saying learning from rewards and punishments in the absence of detailed supervision is that clear okay?

And the crucial component to this is trial and error learning because since I don't know what is the right thing to do given an input right I need to try multiple things to see what the outcome will be right, I need to try different things to see if I am going to get the reward or not right if I don't try different things, I'm not going to be able to learn anything at all right. So we will I can give you more formal mathematical reasons for why we need all of this as we let go on but this is intuitively you can understand this as a requiring exploration so, that you know what the right outcome is right.

And there are a bunch of things which are also characteristic of reinforcement learning problems one of those is that the outcomes right they paid the rewards and punishments based on which you are learning can be fairly delayed in time they need not be temporarily close to the thing that cost it, I mean while you are playing a game let us say right, so you might you know drop a batsman right and then he goes on to score like 150 or something like that right so then you lose the match at the end of the day right but even that cost you to lose the match is the dropped catch that probably happened around the 12th over right, or it could be much more convoluted causal cause and effect right so and how many of you follow cricket?

My god really losing popularity, put your hands down, I'm not going to give a cricket example then forget it okay so a bunch of other things right so, so we talked about delayed rewards the rewards could come much later in time from the action that cost the reward will happen right for example, let's go back to our cycling case that I might have done something stupid honoring all I might have gone over a stone somewhere right, while I am cycling at a very high speed and might have been a small stone in this the road and that that will cause me to lose my balance right.

And then I will try my level best to get the balance back right I might not and I will finally fall down and get her that doesn't mean what caused the falling down is the last action I tried right, I might have desperately tried to jump off the cycle or something like that but that is not what caused the punishment right, the what caused the punishment happened a few seconds ago when I ran over the stone, right.



So there could be this kind of a temporal disconnect between what causes the punishment from the actual reward and punishment so it becomes a little tricky how do you are going to learn those things like learn the association's right. So quite often write your mind need a sequence of actions to obtain a reward and it's not going to be like a one-shot thing they're going to need a sequence of actions, so again going back to the chess example right, you are not going to get a reward every time we move a piece on the board right, you have to finish playing the game at the end of the game if you actually manage to when you get a reward so, it is a sequence of actions right.

And therefore you need to learn some kind of an association between the inputs that you are seeing in this case it will be both positions right or how fast the cycle is moving and how unbalanced do you feel and so on so forth right to actions so the inputs that you are getting which sometimes which we will call states right and the actions that you take in response to this input that you are seeing right. So this is essentially what you are going to be doing when you are solving a reinforcement learning problem. So this kind of associations is essentially called as policies right, so what you are essentially learning is a policy to behave in a world, right.

So learning a policy to play chess or you are learning a policy to cycle right so this is essentially what you are learning that you are not just learning about individual actions right, and all of this happen typically in a noisy stochastic world that the mix things more challenging, so these are all the different characteristics of reinforcement learning problems right and I will be looking at all of this as we go along in when this not be explicitly talking about each and every one of these bullet points but everything that we look at all the algorithms all the methods that we look at as we go along in this course okay, we'll have all these aspects as part of it okay.

(Refer Slide Time: 16:32)



So reinforcement learning has been used fairly successfully in a wide variety of applications right, so you can see a helicopter there okay, so that is not a cut and paste error the helicopter is actually flying upside down, so the this group at Stanford and Berkeley which have actually use reinforcement learning to train after to fly all kinds of things not just upside down an aural agent can do all kinds of tricks on the helicopter so, I will show you a video in a minute and it is an amazing piece of work right I mean it's considered it was considered the showpiece application for reinforcement learning getting such a complex control system to work and it actually could do things at a much finer levels of control than a human being could.

Relax after all a machine so we would expect that but the tricky part was how it learn to control this complex system from without any human intervention right, and in the middle right so they I have a couple of games there so that is can you see that is too small an arrow that's a game called backgammon right so how many of you know about backgammon, 1, 2 plus one maybe. How many of you know about Ludo?

Okay fine so backgammon is like a two player oh okay so you throw the dice you move pieces around and you take them off the board right, so it's a fairly easy game but then you have all kinds of strategies that you could do with it but it's also a hard game for computers to play

because of the stochasticity right and also because of a large branching factor that is that in the game, so at a certain point there are many combinations in which you could move the board pieces around and then there is the role that adds an additional complexity.

So people are not really you know getting great results and then there's this person Jerry Tesaro from IBM who came up with something called neuro gammon thing was called neuro gammon and that was trained using supervised learning and the neural network and, so if it had if he had done it recently it would have been called a deep learning version of neuro gamma or something because he did it back in the 90s early 90s it was just called neural network question of backgammon and I played really well for a computer program right so it was essentially the best computer program backgammon player at that point.

And then Jerry heard about TDI we heard about reinforcement learning he decided to train a reinforcement learning agent to play backgammon so what he did was set up this reinforcement learning agent which played against another copy of itself, let them play hundreds and hundreds of games rather thousands and thousands of games so essentially what they did was so you train one copy for like 100 games or something then you move it here right freeze it and then continue learning with this. So essentially what was happening as you learn you are playing against better and better players gradually your opponent was also improving right.

And then so this was called self play right, so he trained the backgammon using self play and he came to a point where they TD gammon as he called it was even better than the human player of backgammon at that point in the world right so they actually had head-to-head the challenge with the human champion that is a world championship of backgammon you know it's apparently very popular in the Middle East and people actually have World Championships that's the world championship of backgammon and so, he challenged the human champion which IBM seems to do a lot right i mean they challenged a Caspero matches and things like this so, he also challenged just this out of work for IBM.

You should realize right people who spend a lot of resources getting computers to play games, I will probably be working for IBM so, Jerry had this thing and it beat the world champion so we

have reinforcement learning agent that's the best backgammon player in the world not no more best computer player or anything right so we could actually make that game right and there's another game that this is a snapshot from the game of Go right so people have let go, oh come on if that leaves one or two people who have played go okay, people have played the hotel oh okay that's not a very few number isn't it one of those free games on ubuntu I thought everybody plays that because I mean at some point or other well you would rather play Othello then watch paint dry you know.

But anyway it goes a like a more complex version of Othello few with right it's again a very hard game for computers to play because the branching factor is huge right and it is actually a miracle that humans even play this because the search trees and other things that's really complex right so this is 11 case which clearly illustrates that humans actually solve problems in a fundamentally different way then we try to write down in our algorithms because, they seem to be making all kinds of intuitive leaps in order to be able to play go right.

So there's this person David silver who currently works for Google deep mind and but before that he spent some time with Jerry tesoro at IBM and at some point along the way he came up with this reinforcement learning agent called the TD search that place go at a decent level still not, not like master human level performance but it performs plays at a pretty decent level ok so he what I'm pointing out here is things that are typically hard for traditional computer algorithms or even traditional machine learning approaches to solve a as had good success right.

And here is another example of them to point emphasis this or that I forget which one right, they told me I should use only one of those screens for pointing because it's hard for them to record on the other one I forget which one okay, forget it there are some robots on the bottom left of the screen right and so that is a snapshot from the UT Austin Robo soccer team called Austin Villa right, and they usually first we're learning to get their robots to execute really complex strategies so this is really cool. But the nice thing about the Robo socket application is that they don't use reinforcement learning alone right they actually use a mix of different learning strategies and also planning and so on so forth, which is going on the other studio right so they use a mix of different kinds of AI and machine learning techniques in order to get a very very competent

agent is very hard to beat and they I mean the Champions I think two or three years running now in the humanoid league right.

And again hard control problems things like how do I take a spot-kick, you know those were the things for which they used reinforcement learning which is a really hard balancing products we basically have to balance the robot on one leg and then swing the other leg so that you can take the kick and it turns out to be a hard control problem right. So they used RL to order to solve those right and then up on the top right ok is an application which will probably the one that actually makes money of all these three all the others right that is on essentially on using read first learning to solve online learning right.

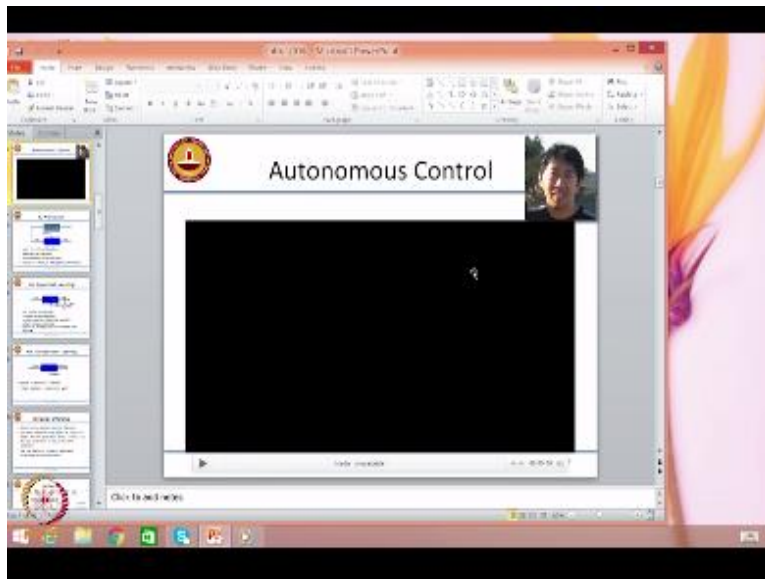
So online learning is a use case where I do not get I don't have the feedback available to me a priori right, so the feedback keeps coming piecemeal right so for example that is a case where we are having new stories that need to be shown to people that come to me webpage right, and when people come to the page I will have like some editors will pick like 20 stories for mean d from those 20 stories i have to figure out which are the ones i put up there prominent and what is the feedback i am going again. Nobody tells me what stories that user is going to like I mean then I cannot have a supervised learning algorithm here right.

So the feedback I am going to get this if the user clicks on the story i am going to get a reward the user does not click on the story I will not get a reward that's, essentially the feedback i am going to get nobody tells me anything beforehand. So i have to try out things i have to show different stories to figure out which one is going to click on and i have very few attempts to do this in so how do I do this more effectively so, people have done a supervisor approach for solving this and it has worked fairly successfully. So it has worked fairly successfully but what reinforcement seems to be a much more in a natural way of modeling these problems.

So not only in these kinds of a new story selection people use reinforcement ending ideas even in ad selection right, so how do I see some of those ads that you see on the sides when you go to Google or some other page right so how are those ads selected so, so there might be some very

basic economic criterion for selecting a slate of ads okay, here are these 10 ads which would probably give me the right payoff right and then you can figure out which of those which three of those 10am I going to put here and things like that you could use a reinforcement learning solution for selecting those right.

(Refer Slide Time: 27:12)



(Refer Slide Time: 27:45)



Of course the this is this whole field called computational advertising right, it is a lot more complex than what I explained but RL is a component in computational brutalizing as well right. okay here is the video courtesy Andrews webpage the people recognized again there huh, okay which is not a human-sized helicopter but still it is a fairly large amazing all of this is being done by a aural agent this goes on for a while we stop.

**IIT Madras Production**

**Funded by**

**Department of Higher Education**

**Ministry of Human Resource Development**

**Government of India**

**[www.nptel.ac.in](http://www.nptel.ac.in)**

**Copyrights Reserved**