

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

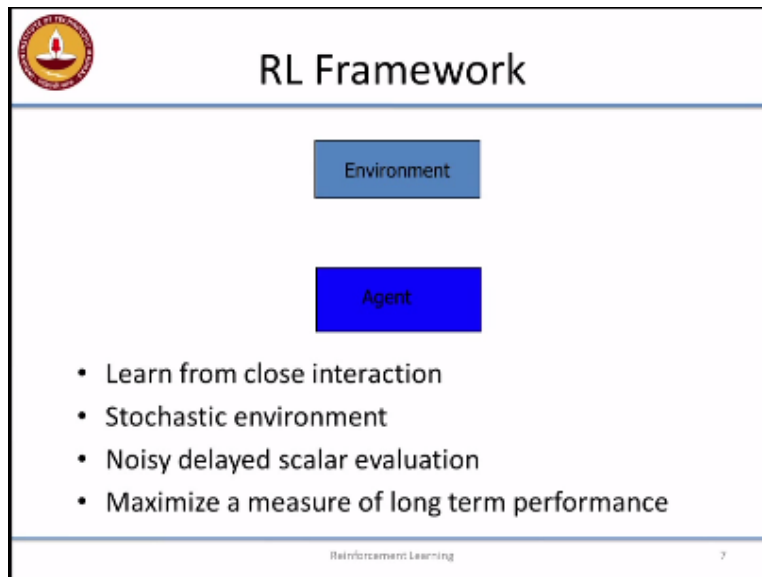
Introduction to Machine Learning

Lecture-84

RL Framework and TD Learning

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

(Refer Slide Time: 00:15)



Right so the Kirk's in the reinforcement learning is that the agent is going to notice is learning agent right it is going to learn in close interaction with an environment right so the environment could be the helicopter it could be the cycle right and or it could be your backgammon board and your opponent all of this could constitute environment a variety of different choices so you sense the state in which the environment is in right you sense the state of the environment and you can figure out what is the action that you should take in response to the state right.

So in apply the action back to the environment this causes a change in the state right so now comes into keeper right so you should not just choose actions that are beneficial in the current

state but it should choose actions in such a way that they will put you in a state which is beneficial for you in the future right just capturing the queen of your opponent is not enough in the chess that might give you a higher reward but it might put you in a really bad position so you do not want that right you really want to be looking at the entire sequence of decisions that you are going to have to make.

And then try to be have optimally with respect to that right. So what we mean by behave optimally in this case right we are going to assume that the environment right is giving you some kind of an evaluation it is like falling down hurts when or capturing a piece maybe gives you a small plus point five or winning the game gives you like hundred right so every time you wine very time you make a move or every time you execute an action you did not get a reward or you did not get an evaluation from the environment right.

So it could be just zero it could be it could be nothing so I should point out that this whole idea of having an evaluation come from the environment is just a mathematical convenience that we have here but in reality if you think about biological systems that are learning using reinforcement learning right all they are getting is the usual sensory inputs right so there is some fracture in the brain okay that sits there and interprets some of those sensory input as rewards or punishments right.

So you fall down you get hurt I mean that is still a sensory input that is coming from your skin right or somebody pats you on a back that is still a sensory input that comes from the skin and it is just another kind of an input right so it could choose to interpret this as a reward right or this as a collision with an obstacle something is brushing against my shoulder let me move it right or you can just take it as somebody is patting my back so I did something good right.

So it is a matter of interpretation so this is a this whole thing about having a state signal and having a separate evaluation coming from the environment is a friction right there is created to have a clear cleaner mathematical model but in reality things are a lot Messier you do not have such a clean separation right and like I said so you have a stochastic environment.

(Refer Slide Time: 03:18)



RL Framework



- Learn from close interaction
- Stochastic environment
- Noisy delayed scalar evaluation
- Maximize a measure of long term performance

You have delayed evaluation noisy so the new term that we have added here is scalar the new term we have had a daily scalar so that is one of the things with the classical reinforcement learning approaches he said I am going to assume that my reward is a scalar signal right so have we talked about getting hurt and having food and so on so forth what do all of this will happen mathematically is I will convert that into some kind of a number on a scale right.

So getting hurt might be minus 100 right getting food might be plus 5 winning the game might be plus 28 capturing a piece might be +0.5 or something like it so I am going to convert them to a scale right and the goal is now know that I have a single number that represents the evaluation the goal is now to get as much as possible of that quantity over the long run okay, make sense right. So if you have questions doubts stop me and ask.

So mathematically a scalar which is easy at optimized not necessarily right I am just talking about so it us like a cost function if you want to think about it in terms of in terms of control systems right so this is like a cost and I am trying to optimize the cost all right and so for the cost is going to be vector value and then I have to start trading off one direction of the vector against the other so which component of the vector is more important so then it get into all kinds of super at optimality and of questions so it is not really clear what exactly is optimal in such cases so here again let me emphasize it is not supervised learning right.

(Refer Slide Time: 05:11)



Not Supervised Learning!



- Very sparse “supervision”
- No target output provided
- No error gradient information available
- Action chooses next state
- Explore to estimate gradient – Trail and error learning

In supervised learning this is essentially what you are going to see there will be an input and there will be an output that you are producing and somebody will be giving you a target output okay so this is what you are supposed to produce and essentially compare the output you are producing to the target output right and we can form some error signal right and you can use that error in order to train your agent right.

You can try to minimize the error you can do gradient descent on their work into variety of things you can try to train the agent so here I do not have a target I do have to learn a mapping from the input to the output but I do not have a target and hence I cannot form an error right and therefore my trial and error becomes very essential see if I have errors rate I can form gradients of the errors and I can go in the opposite direction of the gradient of the error right and then that he gives me some direction in which to change my parameters and that constitute the agent all right major is going to be described in some way right the error gives me a direction right but now since I do not know a direction right so I just I do something I get one evaluations I do not know that the evaluation is good or bad right so think of writing an exam right I do not tell you the right answer I just tell you three right and so what do you do now do happy with answer should we change it should it change it in one direction or should he change it to the other direction.

See what makes it even more tricky is I do not you do not even know how much the exam is out of so when I say 3 it could be three out of three it could be three out of 100 right, right so it could

be any of these things right so you don't even know whether, three is a good number or a bad number so you have to explore to figure out A if they can get higher than three right or three is the best the second thing is if I can get higher than three how should I change my parameters to get to become higher than three.


Let us I have to change my parameters a little bit that way okay I have to change the parameter rise a little bit this way right so if I am cycling wait I have to push down a little harder on the pedal okay I will have to push down a little softer on the pedal to figure out whether I am staying balanced for a longer period of time or not I do not know that otherwise unless I try these things I would not know this is why the trial and error part.

So if I pushed on a little harder and I stay balanced maybe I should try pushing down even more harder next time right so maybe that will make it better and then there might be some point where it too poor so I need to come back so this is how things which you have to try unless you try that you do not even know which direction you have to move in right so this is much more than just the psychological aspects of trial and error there is also a mathematical reason if you want to adopt my parameters right I need to know the gradient okay so that you need to yeah.


The reward is the one that you know that gives you the evaluation for the output right so herein the supervised case the error is the evaluation for the output of the error is 0 then your output is perfect right but then the way of gauging what the error is because you have a target which you can compare right and from there you get the error so in the reinforcement learning case the evaluation is directly given to you as the evaluation of the output right it is not necessarily comparing against a target value or anything you do not know how the evaluation was generated that you just get an evaluation directly so you just get some number corresponding to the output and so maybe I should have done put an arrow from the top saying evaluation comes in from there.

But that is exactly where it is coming let us substitute for the error signal but it is just that you do not know what the evaluation is of course the way differs from the error is minor differences you typically tend to minimize error but you tend to maximize evaluation right it is also not unsupervised learning so unsupervised learning has some kind of an input right.

(Refer Slide Time: 09:07)



Not Unsupervised Learning




```
graph LR; Input --> Agent; Agent --> Activation;
```

- Sparse “supervision” available
- Pattern detection not primary goal

Reinforcement Learning 9

That goes to the agent and then it figures out what are the patterns for thee in the input right here you have some kind of an evaluation and you are expected to produce an action in response to the input it is not simply pattern detection right so you might want to detect patterns in the input so that you know what is the right response to give but that is not the primary goal right but in ref unsupervised learning the pattern deduction itself is the primary co so that is the difference right. (Refer Slide Time: 09:41)



Temporal Difference

- Simple rule to explain complex behaviors
- Intuition: Prediction of outcome at time $t+1$ is better than the prediction at time t . Hence use the later prediction to adjust the earlier prediction.
- Has had profound impact in behavioral psychology and neuroscience!

Reinforcement Learning 10

So here is one slide which I think is kind of the soul of reinforcement learning right it is called temporal difference so I will explain a little more detail and in a couple of slides but the intuition here right so if you remember the Pavlov's dog experiment right what was the dog doing it was

predicting the outcome of the bell you know if the bell rings there is an outcome that is going to happen it is predicting the outcome which is food is going to happen and then it was reacting appropriate to the outcome right.

So most of reinforcement learning you are going to be predicting some kind of outcome that is going to happen since I am I going to get a reward if I do this or if I am I going to not get a reward I am I going to win this game if I make this move what am I not going to win this game all right so I am print always trying to predict the outcome the outcome here is the amount of reward or punishment I am going to get right this is essentially what I am trying to predict at every point right.

So the intuition behind the what is called temporal difference learning is the following right so the prediction that I make at time $T+1$ okay of what will be the eventual outcome let us say I am playing a game right I am going to say I am going to win now I am very sure I am going to win down, right. So I can say that with a greater confidence closer to the end of the game then I can at the beginning of the game right so I have all the pieces set up and if I am going to sit there then and say I am going to win the game right it is most probably visual thinking right but then you have played the game for like 30 minutes or something and there are like five pieces left on the board.

Now I am going to say I am going to win the game now I say I am going to win the game that is a much more confident prediction then what I did at the beginning right so taking this to the extreme right so the prediction I make at $t + 1$ is probably more accurate than the prediction I make at t right, the prediction I make it $t + 1$ is more accurate than the prediction I make it t , so if I want to improve the prediction I make at t , what can I do?

I can look go forward in time then basically go to the next day let the clock tick over and see what is the prediction I will make a time $T+ 1$ with additional knowledge I am getting right, I would have moved one step closer to the end of the game so I know I know its little bit better about the game right I do not know how the game is proceeding I know I can may now make a prediction about whether i will not lose right.

And use this go back and modify the prediction I make it time at time T , and T I think there is a possibility of say probability of 0.6 of me winning the game okay and then we make a move then

I find out that I am going to lose the game with a very high probability then what will I do is I will go back and reduce the probability of winning that I made at time T so instead of 0.6 I will say okay maybe 0.55 or something right.

So next time I come to the same state as I was at time T, I would not make the prediction of 0.6 I will say 0.55 that is essentially the idea behind component difference learning right so it has a whole lot of advantages we will talk about it a couple of slides down but one thing is that the significant impact in behavioral psychology and in neuroscience right so it is widely accepted that animals actually use some form of temporal difference learning.

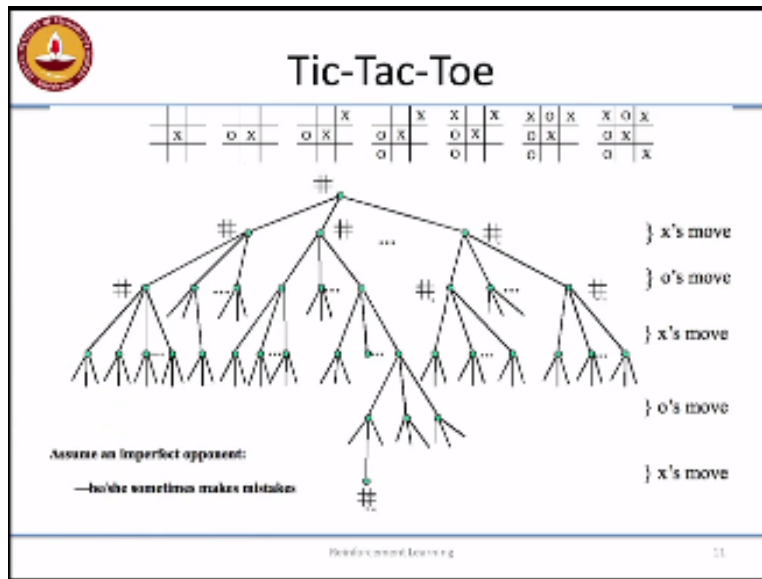
And in fact there are specific models that that have been proposed for temporal difference learning which seemed to explain some of the neuron transmitter behaviors in the brain yeah, no see at this point I will be making a prediction about what is the probability of winning and it could be for each of the moves right if I make this move what is the probability opening if I make this move what is the probability of pending.

Let us say I make move two okay and then I go I see a new position right the opponent response to it and then I decide oh my god this is a much worse move than I thought earlier, so what I do is I will change the prediction I make for move two in the previous state you see that the other moves will not be affected because the only move I took was 2 only about that move I have additional information therefore I can go back and change the prediction I make for move 2 alone.

So you can still have the 10 moves so they are not changing any of that yeah, seeing the prediction is not like I mean if an ideal world you should be able to take back a bad move right, except if it is a parent playing with the kid I do not think those things are allowed right, in fact I when I play with my son we have sometimes had to remain back all the way to the beginning it will probably be me asking to do the remaining not him, because he will be dropping me in someone of those games but yeah, otherwise you cannot you just make the change the prediction.

So next time you play the game it will be better at it, it not for that game well basically you have missed up or you did well I mean so whatever it is yeah, okay I hope I was not too boring in somebody fell over right, that is known to happen yeah, so people sleep and I actually had a person sleep and they fall off the chair once.

(Refer Slide Time: 15:24)



Yeah, I still cannot get over this okay, there is one time was going to teach a class right, and I said was entering the class one person was leaving the class I said hey, what are you doing I suppose me in my class he said no, no, I feel very sleepy I cannot and I do not care if you are going to sleep this get back to the class right, and he looked at me for a minute shortly I said okay, and they walked into the class went to the last place actually lie down on the bench on and I going to sleep okay, and he recently sent me a friend request okay, coming back to looking at the RL right.

So listen looked at tic-tac-toe right, how many of you have played tic-tac-toe good, even you put your hand up okay, good so in tic-tac-toe so you have these board positions right, and so you make different moves, so in the first this is what I have drawn here is called a game tree right, so I start off with the initial board which is empty right, and there are how many possible branches there for people making moves nine possible branches right, for excess move there are nine possible moves I have nine possible branches and then for each of these I will have like eight

possible I am not sure this is the right TV and for each of this i have eight possible branches and they keep going, right.

So what we are going to be doing is essentially trying to this formulate this as a reinforcement learning problem so how will you do this as a reinforcement problem right, so I have all these board positions right, let us say x is the reinforcement learning agent and o is the opponent right, so initially given a blank board I will have to choose one among nine actions right, so there state that I am going to see is this the accessory goes on the board right, and the moves I will be making are the actions right. So in the initial position I have nine actions I make that do I get any reward not really there is no natural reward signal that you can give.

Essentially the reward that I am going to get in this case is if at the end of the moves if I win I will get a 1 if I do not win I get a 0, and if I win I get a 1 if I do not pin I get a 0 right, so what is going to happen is I am going to keep playing these games multiple times right, and at each point right yeah, okay, so there is a note here so what is it note say. You have to assume it is an imperfect opponent right, otherwise there is no point in trying to learn tic-tac-toe why.

We will always draw and the way we have set up the game you are indifferent between drawing and loosing so you learn nothing I mean basically, so you will not know even learn to draw okay, you will just learn to nothing, basically we learn nothing because you can never win right, so you are never going to get a reward of 1 so you will just be playing randomly, so it is this is a bad, bad idea so let us assume that you have an agent that is imperfect right, that makes mistakes so that you can actually learn to figure out where the agent makes mistakes, where the opponent makes mistakes and learn to exploit those things, okay, right.

So your states are going to be this board positions as you can see we give you see a game that has been played out on the top of the slide right, and the actions you take or in response to those board positions and finally at the end of the game and if you win you get a 1 if you do not win you get a 0 right, does not clear. Sir, in case like I mean does it have to be a binary sort of a reward system I mean could you have a scale whether there are three parameters you lose a 0 if you draw 1 by privilege 1.

Sure, you could even know other things like if you win it is 1 if you lose its -1. Yeah, you possibly could but you probably have to pay a lot of games because the perfect opponent it is

almost impossible for you to start getting any feedback in the beginning right, you will always be losing so it is going to be hard for it learn but you will eventually learn something yeah, it will take a lot of mozi level to learned something go back, so if I say that at every point, so we are learning like at a particular stage the probability of winning and like it is what I am going to US state so you are storing information for each and every state that you have entered right.

So how will it be different from exploring the proper next state space every time because after you have done let us say a thousand games or million games you would rather explored a lot of states I will have to store for each state the probability of you winning at that point. Yeah. And all that so I will that be different from exploding it again.

I know the probability of winning program why would I have to close, this still I am not even totally how you are going to solve it okay, let me explain that and then you can come back and ask me these questions, okay if you still have that okay, quit. So what the way we are going to try and solve this game is as follows right, for every board position I am going to try and estimate the reward I will get if I start from there and play the game to the end right, every board position I am going to look at the word I will get if I start from there and play till the end.

Now if you think about it what will this reward connected right, so if I win from there I will get a 1 if I lose from that or if I do not win from that I will get a 0 right, when I say what is the reward I expect to get starting from this board position right, it is essentially this average over multiple games, it some games I will win some games I will lose, or I will not win like some games I win some games I will not win so what will this expected reward represent after having played many, many, many games.

The probability of winning right, the reward is going to represent the probability of winning in this particular case right, if the reward had not been 1 right, if it had been something else if it had been +5 that you would have been some function of the probability of winning right, half it has been +1 for winning -1 for loosing and 0 for draw well it is something more complex is no longer the probability of winning right, it is the gain I expect to get right, how what fraction of games I expect to win over the fraction of games I expect to lose or something like that, right so it becomes a little bit more complex.

So there could be some interpretation for the value function but in general it is just the expected reward that I am going to get starting from a particular board position okay, so that is what I am trying to estimate right, that is assume that I have such an expectation well defined for me right, as you say I have such an expectation well defined, right. Now I come to a specific position let us say I come to this position here right, let us say I come to this position.

How will I decide what is the next move I have to make sorry, whichever next state has the highest probability of winning so I just look ahead to see okay where if I put if the x here right, if I put the X here what is the probability of winning, if I put the X here what is the probability of winning, if you put takes here what is the probability of winning right, I do this for each one of these right, and then I figure out whichever has the highest probability of ending and I will put the x there, right.

So that is how I am going to use this function does it make sense, yes, it is very important so this is this is something which issued understand this is the crux of all reinforcement learning algorithms right, I am going to learn this function that tells me if you are in this state right, if we play things out to the end what will be the expected payoff that you will get right, whether the rewards or punishment or cost whatever you want to call it what is the expected value you are going to get and I want to behave according to a this learnt function.

So when I come to a state I look ahead figure out which of the next states has the highest expectation and then go to the state okay, great how do I learn this expectation. What is the simplest way to learn the expectations, this is especially keep track of what happens, essentially keep track of the trajectory through the game tree right, you play a game you go all the way to the end right.

So you keep track of the trajectory and if you win right, you go back along the trajectory and update every state that you saw on the trajectory you update the probability of winning right, it just increase it a little bit or you come to the end of the game and you found that you have not 1 right, you go back along the trajectory decrease the probability of winning a little bit, right.

Alternatively you can keep the history of all the games you are played so far right, after every game has been completed you can go back and compute the average probability of winning

across the entire history of all the games in which you saw that particular position right, make sense thus easiest way of estimating this probability, right.

But the problem with this is a you have to wait till the game ends right, or you have to store the history of all the games you have played try to means all of these could be potential drawbacks okay, you can get around the history part by coming up with an incremental rule but the main difficulty here is you will have to wait all the way to the end of the game right, before you can change anything along the way so tic-tac-toe was easy is like how many moves can you make in tic-tac-toe at best 4 right, the fifth one is determined for you, right.

So it is basically four choices that you can make right, so and that is easy enough to remember right, you can always wait till the end of the game and then you can always make the updates right, what if it is a much more complex situation right, what if you are playing chess, maybe you can wait till the end, so what if you are cycling maybe you can wait till the end exactly he we do not know right, this it depends on where you are cycling if you are cycling learning to cycling 90 meters it is fine, when you are learning to cycle somewhere on the Surdhal Patel road you do not want even think about what end is there right, so this is there are some tasks for which you really like to learn along the way, right.

So this is where TD learning comes into comes to help right, I do not think I have it slide anyway and I am not using the fancy thing where I can draw on the projection, so let us see if I can do it here right. Suppose I have come here right, and from here I have played at this point I know the probability of winning is say point 4 right, so I came here by making a move from this position, so I said here late and we made a we know that the probability of winning from here is a point 3 right.

But I made the move from here to come here right, but here I had thought my probability of winning was let us say point 6 then I thought my probability of winning was of point 6 right, but then I looked at my next states and I found that the best one was point 3 somehow right, so I went there right. But now since the best I can do from here this point 3 me saying point 6 here there is something wrong right, so I should probably decrease the probability of winning from here right. So why could it be, why could it have happened that I thought that was point 6, but the best among the next was point 3, the thing is.

So that whenever I came to come through this part right maybe I one before right it so happened that when I went through like this right initially I would have gone through like this and played the game and in the examples I drew I might have actually won some of those games right. So I would have change this 26 right but it is possible for me to get here by playing a different sequence of moves also right.

So for example to come here right I could have put the X first here and then here or I could have put the X like I did here I put the X first here and then here, that either way I would have reached this position right, so there are many combinations in which I could have reached the same positions right it just to be nice to these guys right. To reach here there are different orders in which I could have put the O and the X right here we have showing a cell specific order the O first 1st put here then put here that the x was first put here. Then put in it could very well be I put the X first here in the war first here and then I put the X here in the O here right.

The multiple ways in this thing goes reach right so sometimes when they play those games right I lost, all right sometimes and it played these games I won therefore it turns out that for due to some random fluctuations right, so sometimes i will when I go through this place pacific point and that is what I have a higher evaluation of winning right but when I went through the other paths I had a lower evaluation of winning right. But we know that really does not matter what path you went through in tic-tac-toe right. Once you reach that point what is going to happen further is determined only by that point.

So what i can do now is take this 0.3 right you should update that 0.6 down, so I am very confident here I think I will win with the probability of 0.6 right but the best probability I have from the next stage is 0.3, therefore here I should not be so confident right good point, that depends on the house sarcastic your game is right. So if you are game has a lot of variability then you do not want to make a complete you know commitment to a 0.3 so you might want to say ok now let me move it little bit towards 0.3 right. But if it is a more or less a deterministic game then you can say okay 0.3 yes sure let me go to all the way to, in the difference on the yeah it is misleading it is called game tree actually.

But it is a game graph in this case yeah, so as I said kid when this is this is an instance of temporal difference learning, so how while I use the thing to update this is called temporal differential learning okay. So there is one other thing which I should mention here right if I

always take the move that which I think is the best move right now, right let us talk about it I start tab I have never played tic-tac-toe before right, so I play the game I play it once I get I get to the end I win. So now what I do I go back right whether I am using temporal difference learning or waiting till the end up dating whatever it is I change the value of all the moves I have made in this particular game right.

So the next time I come to a board position what am I going to do? I look at all possible outcomes everything except the one that I have played will have a 0 right and the one that I have played will have something slightly higher than 0, I am going to take that, then in fact it will be like how many of you watch the movie Groundhog Day? It will be like Groundhog Day I will be playing I will be playing the same game again and again because that is what happened to give me a win in the first time around right. That but that might not be the best way to play this right.

So I need to explore right, so I need to explore multiple options right so I should not be always playing the best festival right I should always be paying the best move I need to do some amount of exploration, so that I can figure out if there are better moves than what I think is currently the best move right. So tic-tac-toe there is inherently some kind of noise if your opponent is random right but if an operand is not random and if operand is also playing a fixed rule and if you are playing greedy, then you will be just plain a very small fraction of the game tree and you would not have explored the rest of the outcomes right.

So you have to do some amount of things at random so that you learn more about the game right. So here is a question for you, when I am estimating this probabilities of winning right, let us say I have reached here I look downright and the action that gives me the highest probability of winning say gives me a probability of say 0.8 right what I want to explore right so I take an action that gives me a probability say 0.4 okay. So I will go from here to another action that has a probability 0.4 that another boat positions that has a probability of 0.4 of winning. So should I use this 0.4 to update this probability or not.

No why? that you are questioning the whole TD idea and you are exploring you should probably wait for the or not just ignore it okay, any of any other answer because you are good or a bad move will be found out I have to update the value of that new I agree. Do I update the value of winning from the previous board position was the question so that 0.4 I will have to change right but do i change the 0.8, that was the question the 0.8 was a probability of winning from here

right I look or whatever. So probably say I had a probability of winning of 0.6 from here I look at the bottom and the best probability of winning says 0.8.


But then I take because I am exploring I take an action that has a probability of winning of 0.4 all right the question is do I go back and change the 0.6 towards 0.4 or do I leave the 0.6 as it is? Sorry that one where I am exploring rate, I mean this is we will be necessarily be less than 0.8 this will be 0.4 will be 0.6. So the question here is 1 way of arguing about this is to say that, a if I am playing to win right I will play the best action from here and then the best action says 0.8 therefore I should not penalize it for the bad action which is 0.4.

Which I did to learn more about the system and that is one way of thinking about it another way of arguing is to say that hey, no this is how I am actually behaving now right. So I should give you the probability of winning about I about the current behavior policy right, this should not be some other ideal policy should be about to what I am behaving currently and therefore I should update it right. So which one is correct first or the second questions? But this is something these are this is like I said ask you to think about the whole tic-tac-toe thing and many of these answers have relevance later on.

In fact there are 2 different algorithms one does option one does option to write, so there is no right answer or wrong answer right answer is depends yeah, so yeah so this is a different things that you can think about in this but I told you about 2 different ways of learning with tic-tac-toe one wait till the end and figure out what the probabilities will be, the other one is keep adapting this as you go along right and both cases you not explore that is it to keep out here in both cases you have to explore otherwise will not learn about the entire game.


So this is where the Explorer exploit thingy comes in okay yeah. Great question different algorithms deal with indifferent way that is one of the crucial questions that you have to answer in MRL. So it is called the explorer exploit dilemma right, so you have to explore to find out which is the best action right and you have to exploit.

(Refer Slide Time: 37:19)



Explore-Exploit Dilemma

- One key question - the dilemma between exploration and exploitation
- Explore to find profitable actions
- Exploit to act according to the best observations already made
- *Bandit problems* encapsulate 'Explore vs Exploit'
- Chapter 2



Reinforcement Learning 32

Whatever knowledge you have gathered right and you have to act according to the best observations already made right, so this is called exploitation right. So the key one of the key questions is when do you know you have explored enough right should I explore now or should I exploit now, this is called the explorer exploit dilemma right and a slightly simpler version of reinforcement learning called the Bandit problems okay. Some carefully called bandit problems they of course he is an expert on bandit problems here you can the Bandit problems encapsulate.

This explore exploit dilemma my god lot of people are turning and looking at a noticeable but, so this will ignore a whole bunch of other things like the delayed rewards you know the sequential decisions and other things. Even in the absence of all of these other complications that even if I say that you are all your problem is you have to take an action and you will get a reward okay your goal is to pick the action that gives the highest reward. I will give you 10 actions you have to pick the action that gives you the highest reward right, but the problem is you do not know what is the probability distribution from which these rewards are coming right.

So you will have to do some exploration I have to actually do every action, at least once okay to know what will be the reward even if they are deterministic right. So I cannot say which is the best action before I try every action at least once, if it is deterministic it is fine I can just try every action once and I know what to seek payoff right. But if it is to scars tic I have to try every action multiple times right how many times you have to try it depends on the variability of the distribution.

IIT Madras Production

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

www.nptel.ac.in

Copyrights Reserved