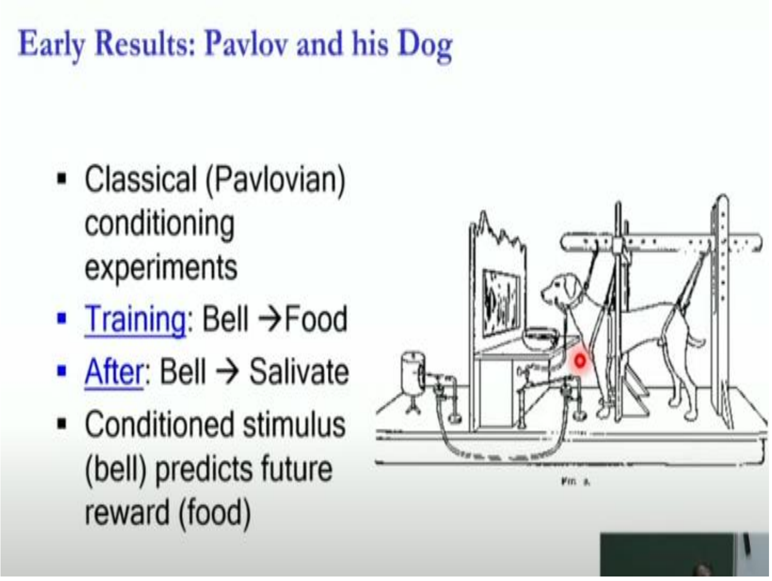
Artificial Intelligence
Prof. Mausam
Department of Computer Science and Engineering
Indian Institute of Technology-Delhi

Lecture - 80
Reinforcement Learning: TD Learning and Computational Neuroscience

But let me show you the value of TD learning in the case of Pavlovian dog.

(Refer Slide Time: 00:23)



Early Results: Pavlov and his Dog

- Classical (Pavlovian) conditioning experiments
- Training: Bell →Food
- After: Bell → Salivate
- Conditioned stimulus (bell) predicts future reward (food)

These kinds of works come from the field of computational neuroscience. So what is computational neuroscience? It says that people, animals whatever behave in a certain way. So in a given state, they take a certain action right, something happens in their brain. Now can we give a computational model to explain the behavior of the brain? So I am trying to build a computer which mimics the animal and that is computational neuroscience.

It is somewhat different from artificial intelligence. AI says I want to build an intelligent machine, it can be superhuman behavior, it can be subhuman behavior, depending upon where the state of the world is. We do not have to follow much humans, we do not have to give the emotion of humans to the machine. We do not have to give many such things. We have to, you know solve AI, independent of how humans work.

We can take insights, but we have to solve independently. Whereas neuroscience is the fundamental discipline which studies the brain, which studies how the brain works, which studies how the behavior works. And computational neuroscience says, can I create a computational framework to replicate the behavior of the animal? So what do they do? They do experiments of the following form.

They get a monkey or a rat or you know whatever, and they give them some tasks. And the monkeys or the rats do those tasks and they see how they do the task. And then they try to figure out an algorithm or a model, which explains that behavior, okay. So computational science by definition starts the data from the actual animal, and then tries to come up with the computational model that is closest to that data, right?

So it is a closely aligned field, but still different from AI. So let us see how TD learning shows up in the context of computational neuroscience. So now we all know about the Pavlovian experiment. There is a bell. You give food, ring the bell, the dog slowly learns association between the bell and the food. It salivates whenever it sees the food and later even if you ring the bell and do not give food the dog still salivates and they have learned the association with bell and the salivation right.

Now, let us take it to one step further. Let us say I ring the bell and I give food 100 seconds later. What is going to happen? Okay. So let us forget TD learning for a minute. What do you think is going to happen? We do this many times we ring the bell and give food 100 seconds later. This is for let us say monkeys. We ring the bell and then give food 100 seconds later. What do you think slowly is going to happen?

“Professor - student conversation starts” Yes. Harsh. It will start salivating after almost 100 seconds after bell rings. **“Professor - student conversation ends”.** So Harsh says that the dog would learn to salivate 100 seconds later or the primate would learn to salivate about 100 seconds later. Okay, that may be I do not. I am not sure about this though.

In particular, if 100 seconds is 2 hours maybe but if 100 seconds is small number is possible that they start salivating in anticipation of the food that is going to come. But let us think about what do they know about the world, right. What reward structure do they thing is going to happen?

“Professor - student conversation starts” Yes, Raghav. It will associate the ringing of the bell with the offering of the food and they will only salivate at the presence of the food and not in the ringing of the bell. **“Professor - student conversation ends”.** Ah, Raghav says that they will not salivate until they see the food. I am not sure that is going to happen. So I am not sure that you can be a good dog.

So well, we do not know, right? So this is not a salivation experiment. So I cannot predict per se, but at least this is my hunch, this is my understanding that they would start behaving as if they know that the food is going to come. Right, whatever that means, right? But how would we know that they know, right? And so what computational neuroscience folks do is that they put some electrodes on the brain and they try to look for a chemical called dopamine.

You know this right. Dopamine is that chemical which tells us that the primate or the brain is feeling happy, right? The brain feels that something good is happening. And now the question that we are trying to ask is when would dopamine actually get secreted? Is it going to be when you ring the bell? Is it going to be when you give the food? Is it going to be somewhere in the middle? When?

So let us think about this question for a minute. This is the right question. I did not frame the earlier question right? So I ring the bell, I give you food after 100 seconds. I ring the bell I give you food after 100 seconds. I do this you know several times. Now of course when you see the food you know you feel happy of course. This is the basic behavior that is known.

Now after this training, what is going to happen when I ring the bell? When would I start feeling happy from and until what point? Guesses. It is a guess, it is a guess question. So there is no answer that you can know until you work with that primates. But let us say yes. **“Professor - student conversation starts”** So Kirti. When the bell rings. **“Professor - student conversation ends”.**

So you feel happy when the bell rings and when the actual food comes, you are still happy that the food has come. So this is so Kirti’s guess that you feel happy when the bell is rung and you know you continue to feel happy for the next 100 seconds of your life. Right, anybody else or everybody believes in this? There are other theories.

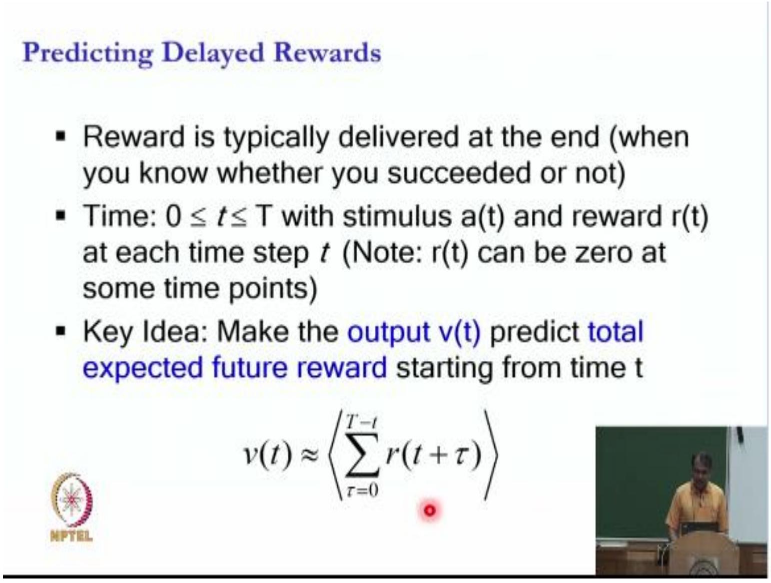
“Professor - student conversation starts” Parth Yes. It can be that once you hear the bell your dopamine concentration increase and then may slowly decrease. **“Professor - student conversation ends”.** So Parth says when you hear the bell, your dopamine increases, then for the 100 seconds, it starts to decrease but then when it sees the food, you know it feels happy all over again.

So in this world for one food you feel happy twice. This is Parth’s suggestion. Any other theories? Give me a wilder theory. **“Professor - student conversation starts”** Yes. What is your name? Shahin, yes Shahin. I think that it will not feel happy when it hears the bell but when the food actually comes probably it will be happy. **“Professor - student conversation ends”.**

So Shahin says you feel you do not feel happy with the bell. There is no conditioning, you know social conditioning everything is a myth. When you ring the bell, nothing happens to you. When food comes, you feel happy. You are completely detached of any baggage of history in your life. This is what Shahin feels. But there is one theory that still remains.

“Professor - student conversation starts” Yes. What is your name? Saigwan. You feel happy when the bell rings and your happiness increase until you get the food. **“Professor - student conversation ends”.** Oh, okay. So Saigwan says you feel happy when the bell rings and you feel happier and happier and happier until the food comes in anticipation of finally seeing the food. Okay, very good. So let us see how many of you are correct okay. So let us work with this.

(Refer Slide Time: 08:08)



Predicting Delayed Rewards

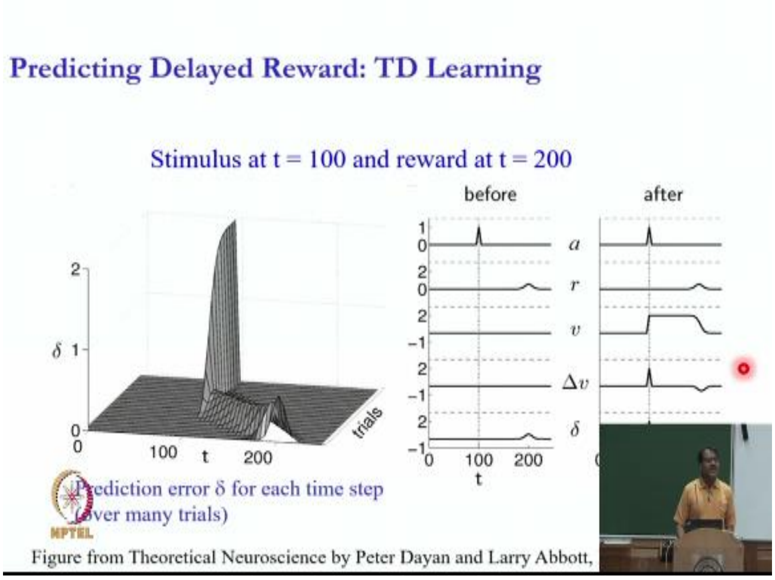
- Reward is typically delivered at the end (when you know whether you succeeded or not)
- Time: $0 \le t \le T$ with stimulus a(t) and reward r(t) at each time step t (Note: r(t) can be zero at some time points)
- Key Idea: Make the output v(t) predict total expected future reward starting from time t

$$v(t) \approx \left\langle \sum_{\tau=0}^{T-t} r(t+\tau) \right\rangle$$

Let us also see what TD learning gives us right, because eventually we are trying to work with TD learning, right. So TD learning is a difference between what I thought world to be and what I estimated the world to be after seeing some observation right. So that is the difference and that difference tells us something about our value function, right. So let us see. So this is the experiment, predicting delayed rewards.

Reward is typically delivered at the end when you know when you succeed or not. That is the food. Now we will work from 0 to capital T. Let us say we give stimulus a t and give reward r t. So stimulus is the bell and reward is the food, okay. And now the key idea is to make the output prediction v t, v t is the expected long-term value starting in time T right. And it is therefore given by all the rewards I will get starting in time T up to a time capital T right.

(Refer Slide Time: 09:22)



Predicting Delayed Reward: TD Learning

Stimulus at t = 100 and reward at t = 200

Prediction error δ for each time step (over many trials)

Figure from Theoretical Neuroscience by Peter Dayan and Larry Abbott,

Now, this is our experiment. So let us look at each of these curves. This is before training on the left and this is after training on the right. So this is before the primates started training and this is after they have been trained many times on the same experiment. So I give stimulus at let us say time equal to 100. This is when the bell rings okay. And this is when I give the reward. So this is when I give the food, actual food.

Now initially they have not been trained so their value of life is zero. In the long run, I am not going to get anything in life. The primates do not know that the human is playing with them and they are going to get rewards. Now let us think about what is the long-term value going to be. Long-term value if you have learned that this is what happens all the time, the bell rings, and after some time I get food.

So then, as soon as the bell rings, the primate knows what does the primate know that they are going to get food. So their long-term value includes the reward, the value that they are going to get. And so that becomes high. Because right at the point they hear the bell, they know that I am not going to get food. So their value has increased the long-term value over life has increased.

And eventually when they have gotten the reward then it comes down to 0 because now nothing else is going to happen. Their life is just now plain old. So this is after training, this is what the value function that TD learning is going to give us, right. We should also look at v t + 1 – v t. What is the difference in the value at time t versus time t + 1. So of course, initially the difference is 0.

Then something happens and the value suddenly increases. So your difference becomes very high and then because we do not have a discount factor here, then my difference stays constant. So value still remains zero and then later my value comes down because I have gotten the food and so my delta will becomes negative, okay. This is the intermediate quantity. But we are interested in the TD error.

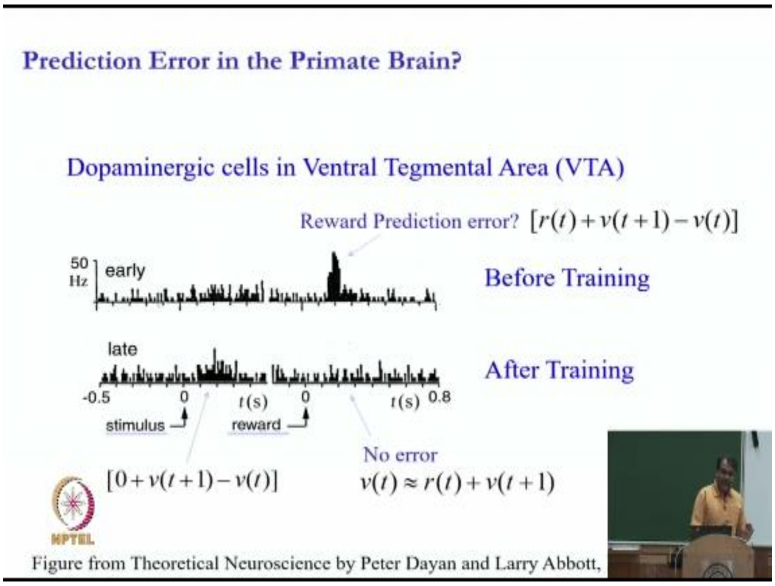
What is the TD error r + v t + 1 – v t, right r + v t + 1 is my new sample and v t is my old prediction. So this is my prediction error, this is my temporal difference error, this is the difference right? So what is the difference? What does temporal difference suggest? Initially after training, there is a lot of temporal difference. So this is we will be adding delta v with r, r plus delta v.

So initially I have a peak because I was not expecting anything, neither did I get the reward but now my expectation has increased. So there was a sudden jump in my value function here. And so this is reflected in the TD error. On the other hand later when this was going to go down because I have actually received the reward, then this reward that we got and the negative reward gets canceled out and so there is no TD error whatsoever, okay.

Now here is what people observed. That your happiness level is completely dependent on this delta, the TD error. So your TD error tells us how happy you are going to be or how sad you are going to be. And where is the TD error? Right at the time the bell is rung. So in practice what happens? None of your four theories are correct believe it or not, or at least fully correct.

Some people thought that there will be double happiness. Some people thought that there will be no initial happiness. What actually happens is very interesting.

(Refer Slide Time: 13:45)



Prediction Error in the Primate Brain?

Dopaminergic cells in Ventral Tegmental Area (VTA)

Reward Prediction error? $[r(t)+v(t+1)-v(t)]$

Before Training

After Training

No error

$[0+v(t+1)-v(t)]$

$v(t) \approx r(t)+v(t+1)$

Figure from Theoretical Neuroscience by Peter Dayan and Larry Abbott,

So before training, nothing is going on, nothing is going on. Bell is rung, nothing is going on. You give reward, the primate is very happy. Yeh, I got reward. Lots of dopamine gets generated. But after training you ring the bell and suddenly there is more activity. Why? Because you are happy, I am going to get reward. But when you actually get reward nothing happens. No happiness. This is very interesting.

Why is there no happiness when you actually see food? Because you already knew you are going to get food. What is the big deal? I have to get food. I am entitled to food. There was a bell. The fact that I got food is yeah, obvious stuff. So this is super exciting if you think about the ramifications. It says your happiness is not really governed by the reward that you get. It is governed by the mismatch in the expectation of your reward that you are getting.

You are happy because you did not expect to get something and you got it. Now there are so many different ways this manifests, okay. I can go on and on and on about it. So first of all, I can tell you a story of my good friend who was very unhappy after he got a JEE rank of 90. And why JEE in our days was not as competitive in as JEE in your days. It is still 150,000 students, 1.5 lakh students used to take JEE.

Why was he unhappy at the rank of 90? Because he expected more. Have you been in a situation where your friend comes to you and says, that is a great movie, you should go see it. And you were like, yeah, it is a good movie. You know this friend is always right. Let me go see it. And then you feel like, okay, what is the issue? You expect you are expecting too much from the movie.

On the other hand, if it is an average movie, but you expected it to be terrible, you will be very happy. Here is the other thing. Now think about what is happening. Suppose I play with the primate, this is a very interesting experiment. Suppose I play with the primate, ring the bell, and at 200 do not give them food. What is going to happen? They are going to feel sad.

They are not going to feel happy at getting food, they are going to feel sad at not getting food. Food was their birthright at that point. This is the beauty of prediction error. If we were really that rational that we became happy and sad with with our with actual rewards you know then all the poor people will be sad and all the rich people will be happy. That just does not happen.

We all adjust to the mean that we are in, the mean reward that we are used to having. And we are only interested in the deviations. This is very interesting also. Somebody who makes a lot of money, somebody who has lot of friends, somebody who is this, somebody who is that begins to expect this. So then one friend who leaves this person, this person feels very sad. How can this friend go? It is okay you can feel sad.

But you know at that point, you are not also thinking about the you know 99 friends you have so much money that you have. You are also not thinking about think about those other people who do not have many friends, etc, etc. You are only thinking about your expectation and the new outcome. Your happiness is only dependent in your world.

That is why a lot of you know philosophers say you should always look at you know other people, they have more problems than you. But you do not realize this, you do not do this. In fact, even the math says that your happiness has nothing to do with their problems. It has only to do with where you think your mean is, and what the delta happens on top of it. I will tell you one more example.

And then we will move on because I can go on and on as I said. I had a good friend who got divorced after what 9 or 13 years or some large number of years being in a marital relationship. So far so good. So once I was having a you know close conversation with her and I asked her why after so many years, she said something really beautiful. So she said her husband was a, whatever abusive.

She says she said when my husband abused me for the first time I was very perturbed. I was very angry. I was, you know I wanted to do lots of things. You know I wanted to run away, this, that. But he calmed me down. He said, you know I have made a mistake. I am sorry. You know they made up everything, right? So she forgave him because it was the first mistake.

But then the second time he abused her it was not abusing from 0 to 2, it was abusing from 1 to 2. There was always the already the context that was built that yeah, he can abuse once in a while. So that difference of experience was not that much. And then over time, the frequency kept increasing and everything.

But one was always looking at the delta, one was always seeing in a where one was and a new experience that happened and it was never too big to make a big decision and you know leave the relationship. Only when she says that she took a step back and looked at in how so many years where she started from and where she has come, she realized that she was, you know living in an abusive relationship and expecting the abuse to be part of her life.

This is when she sort of took the decision. This is very interesting. I did not know TD learning at that time. Otherwise, I would have explained her to her. I may not have been friend anymore if I had done that, that is a different story. But notice, this is what TD learning is saying. TD learning is saying it is all about the mismatch of expectations.

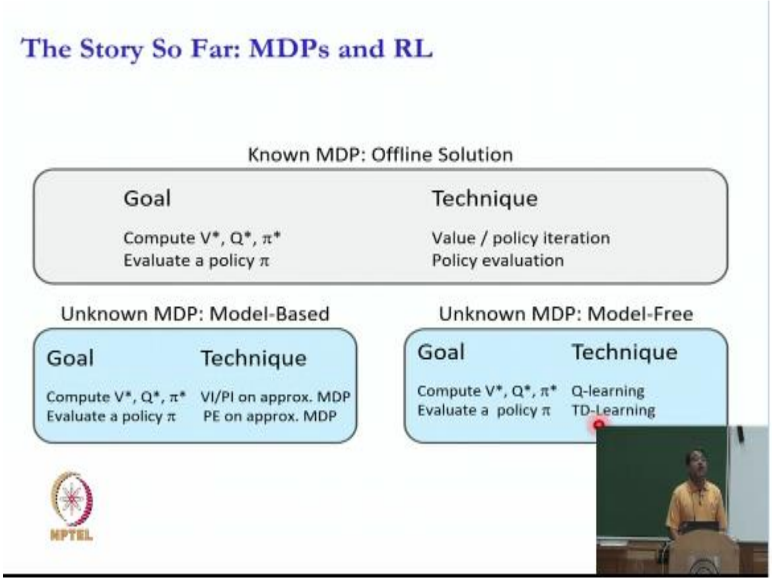
You may over time come in a very bad place or in a very good place your happiness does not matter where you are. It only matters with the difference in what we expected and what you get. Therefore, in one of our philosophical you know sessions, somebody said some very beautiful line, behind every upset, I do not know how many of you have heard this, but you can live by this by the way if you what.

Behind every upset there is an expectation behind every upset there is an expectation. It cannot be any otherwise. It cannot be otherwise. TD learning tells us this. You can laugh. These are beautiful ideas. You know when you think more and more about them, you would appreciate them. That you can only get upset because you expected more and you got less. There is just no other way to be upset.

So if you can reason back and think about, okay, why am I upset? I am upset because I expected this, was it a fair expectation, blah blah blah. If you start reducing your expectation, you will be very happy in life. Like for example, you always you know you did the exam to the best of your ability and say, I am not going to get good marks. I am going to get really bad marks you will be very happy because you will get slightly better marks than you expected.

And then you will be happy. On the other hand, he said I did everything right. You do not even know. The professor is smarter than you sometimes. Then you will get lower marks then you will say oh, I thought I did well. Even if your marks are much better than the other person who expected less and is very happy, okay. So we will stop that philosophical discourse right now. But you see how beautiful the model is.

(Refer Slide Time: 22:18)



The Story So Far: MDPs and RL

Known MDP: Offline Solution

Goal	Technique
Compute V*, Q*, π* Evaluate a policy π	Value / policy iteration Policy evaluation

Unknown MDP: Model-Based

Goal	Technique
Compute V*, Q*, π* Evaluate a policy π	VI/PI on approx. MDP PE on approx. MDP

Unknown MDP: Model-Free

Goal	Technique
Compute V*, Q*, π* Evaluate a policy π	Q-learning TD-Learning

So where are we? The story so far is that we have been trying to do Markov decision processes and reinforcement learning on top of it. We said that, if you are given the transition function and reward well, you can do value iteration, policy iteration or policy evaluation depending upon whether your goal is to evaluate a policy are to compute the policy.

On the other hand, if our goal is to evaluate a policy and we are not given transition and reward, then in the model-based setting, you will first compute transition and reward, estimate not compute, first estimate, transition and reward and then do policy evaluation on this approximate MDP and that is called and that would be model-based. Or we can do TD learning and that will be model-free, right.

Now we are still left with two more steps. The steps are what if I am not giving you a policy? What if my job is to actually find the policy and then what do I do and in that case we still have model-based and model-free. And in model-based learning, we will basically estimate the full transition function and the reward function and then we will do value iteration or policy iteration on the approximate MDP.

On the other hand, if you are doing model-free, then we will do an extension of TD learning called Q-learning, okay. So this is the rest of the part of this lecture.