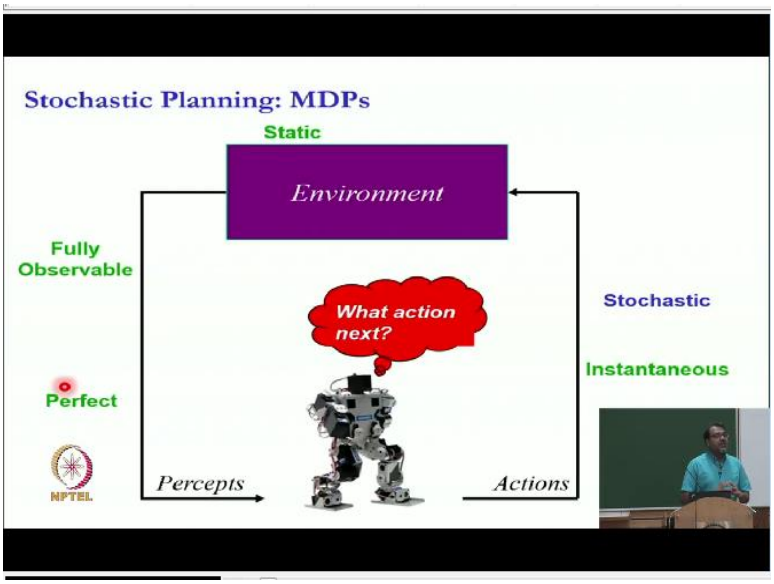
Artificial Intelligence
Prof. Mausam
Department of Computer Science and Engineering
Indian Institute of Science Education and Research-Delhi

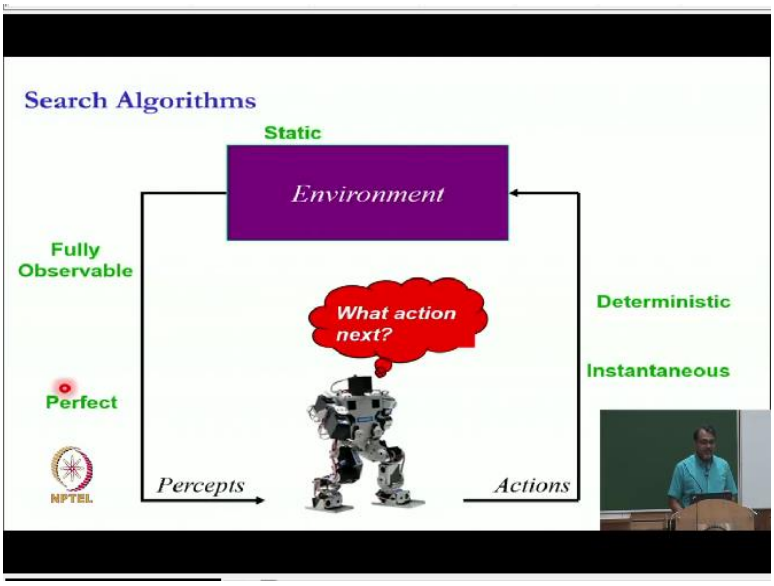
Lecture-71
Markov decision processes:
An Example of a Policy
PART-2

We have started talking about Markov decision processes.

(Refer Slide Time: 00:26)



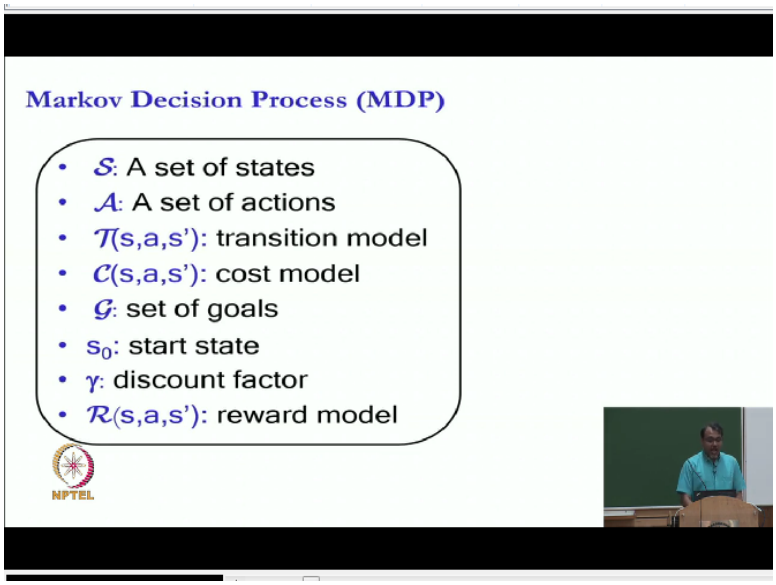
(Refer Slide Time: 00:29)



And if you remember, the main point of change that we did from the standard search algorithms where the world was deterministic is that we will now live in a world which is stochastic. In particular, the stochasticity would be in the agents own actions. So, we are still assuming that the world is fully observable and the sensors are perfect. So, that means I know exactly which state I am in. But when I take an action, they may be probabilistic outcomes.

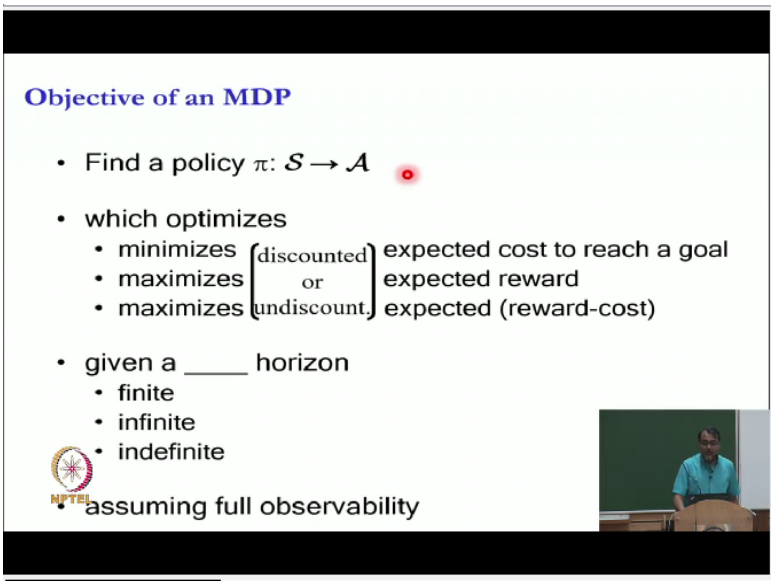
But once there is not a probabilistic outcome after having taken the action, I would know exactly where I reached okay, so this is sort of what the model says.

(Refer Slide Time: 01:08)



And we also defined a Markov decision process as with a set of states or set of actions, a transition function, a cost function. And then we said that something that determines the objective function, so it could be a set of goals or the reward function. And we also said that we might be given a discount factor, okay.

(Refer Slide Time: 01:27)



And the solution that we are looking for is that we are looking for the policy right. Now, why did we say we are looking for the policy because depending upon what happens as an outcome of my action, I may want to do a different action. For example, if I am trying to pick up this cell phone, and with some probability, I pick it up. So I will move on with my task, with some probability I miss it. So then I will try to pick it up again, with some probability it falls down. So then I will try to pick it up from slow which is a very different action I am picking up from table.

So basically, depending upon what happens in my outcome of the action, I might take very different actions downstream. Therefore, I am not looking for the path or the plan, which is a set of actions, a sequence of actions well known in advance I am going to take an action and then depending upon what happens I will be taking the next action. So, there may be many ways to represent this, is one way to represent this is in a policy graph, which we will talk about, but we said that we will define a simple output function which is going to be a policy pi, which is a mapping from states to actions.

Now, this works because we are going to assume full observability which means that once I have taken the action, I know exactly where I landed. Because of that, I would always have access to the state, because I have always have access to the state. I can always check this policy pi and figure out which action do I do okay. So in conjunction with the fact that I am assuming full observability a policy makes sense as an output of the Markov decision process.

The next thing we have to figure out is how long am I allowed to move. If I am allowed to move a fixed number of steps like a fixed time point. For example, most of you are going to graduate in 4 years, right. Let us say everybody was going to either graduate in 4 years or fail and not graduate. Let us say that was what the university system we are living in, and then you can optimize your reward which is your CGPA.

That will be called a finite horizon problem, that your horizon of in you know being in this educational Markov decision process is only 4 years long, and now you have to make the best decisions possible. On the other hand, you may be told that you can go on living in this college for as long as you want, until you finish your credits, until you get your goal. So, if you will live life like this, that would be called an indefinite horizon.

Indefinite means that there is a fixed horizon but I do not know what it is. I cannot even give a bound on it. I cannot say 4 years or 5 years or 7 years, you can live here forever. And a lot of PhD students are like this. That should motivate you to do a PhD okay, bad joke. So, no, that is not true exactly. I mean, in some colleges, there is a fixed limit on how much you are allowed to spend time.

For example, at university of Washington, where I did my PhD, the limit was 10 years. So after 10 years, you had to request for permission to stay, until 10 years, you could just say, nobody asked you anything right. So an IIT Delhi has a different policy, so and not as liberal as yesterday. So in other words, you may not be told that this is your finite number of steps, it may not be 20 40, 4 years, 8 years, it may be any number of steps until you reach the goal.

And that model will be called an indefinite horizon problem. On the other hand, we could give you a world where you live here forever, right. And you do not even die. And that will be a world which will be called an infinite horizon world where you keep on going and there is no limit on anything right. So that is the horizon, how long am I allowed to act. And then the next question is, what am I optimizing.

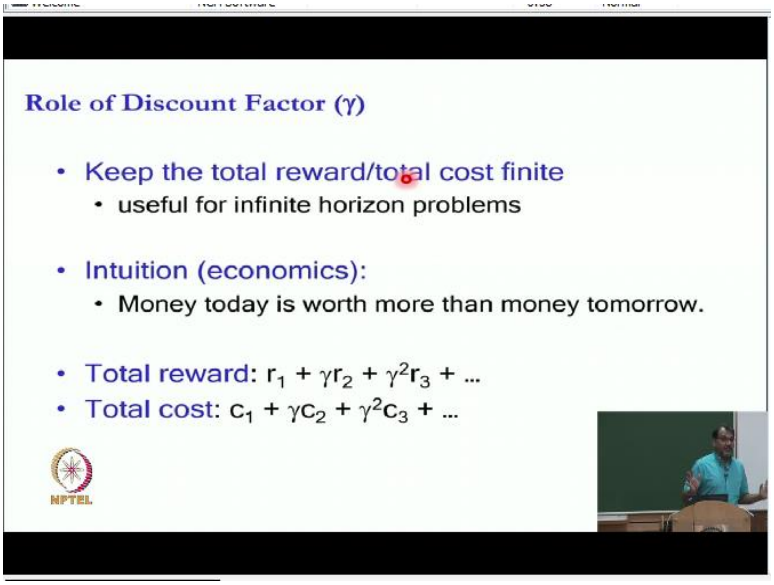
And now that depends on the input that is given to me. So, if I am given cost and I want to minimize costs, if I am given rewards and I want to maximize rewards, if I am given both rewards and costs, maybe I want to maximize reward minus cost, okay. But the problem is my actions may have stochastic outcomes, if the actions of stochastic outcomes, then they can lead to different parts, right.

So, you take an action, you get to some state, you take a different action, you get to a different state, and you run this and this will be called a trajectory, a particular sequence of state action pairs, but in a different world when you take the first action, you may lead to a very different state. And then you may have taken a very different action. And that would have lead you to a different trajectory in life.

And there will be many, many, many, many such trajectories and different trajectories would have given you different rewards. So, what do you want to maximize. Well, each trajectory comes with a probability, which is the probability of the outcome for each action multiplied. So that probability allows you to create an expected reward. So, you would be optimizing either the expected cost or the expected reward or the expected reward minus cost.

And the last thing we discussed last time, is that there is also a notion of discounting because if you want to run for infinite time you want this total reward to be bounded, otherwise, things do not remain well formed. And to make them bounded, there are many, many models one is called the average time model which we are not going to study, but the nicer model that is easier to understand is the discounted reward model, which says that reward r today is worth more than the reward r tomorrow. And the reward r tomorrow is sort of worth gamma times r. And this gamma keeps multiplying.

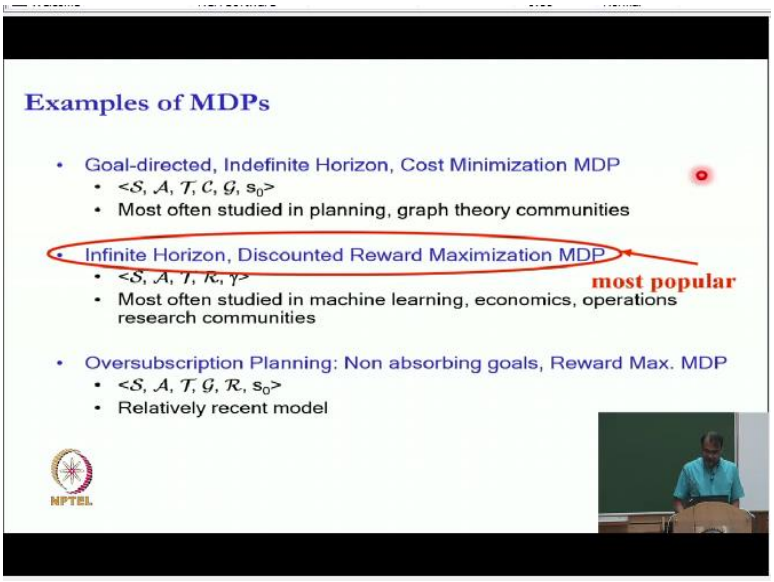
(Refer Slide Time: 07:11)



So when we want to compute the total reward starting today, we will say that the reward that I get in time step one is r 1, but the reward that I get in time step 2 is gamma times r 2, the reward I get in time step 3 is worth gamma squared times r 3, and so on so forth. And this is the total reward that I am going to get. Therefore, my reward, total reward from any state will be bounded by the max reward divided by 1 - gamma, which is a finite quantity.

This is where we were and we stopped, okay, and our job today is to build on it and learn some algorithms that allow us to figure out which action do I execute in which particular state but before I go on, I want to mention that the way I have generalized and define a Markov decision process for you, it can encapsulate a large style of problems.

(Refer Slide Time: 08:01)



So, one very common natural problem that you will see in AI is goal directed indefinite horizon cost minimization MDP, in such a situation you will be given a set of states actions transition and costs with a set of goals that you want to achieve and you are given a start state and this is closest to the atomic agent world. So, if I take away the transition functions stochasticity and make it into a deterministic version.

Then this would be the search problem that we had been studying right. I have goals that I want to achieve a set of goals anyone is fine and I want to achieve it with the minimum cost or in this case minimum expected cost and that would be the closest okay. This model believe it or not, is also called a stochastic shortest path problem. So shortest path is you know, get to a goal with a minimum costs.

Stochastic shortest paths get to a goal with a minimum expected cost, right. And this is what is typically studied in the graph theory, communities and AI. Communities which comes from the classical AI world, then there is a version that people often study in the more modern MDP world and those are infinite horizon discounted reward maximization MDPs. If it is infinite horizon, it will always be discounted, right.

Because we want this to be well formed, there the input will be actions, transitions reward states and a gamma. And this is commonly studied in economics, because I want to maximize the

money, this is commonly studied in the reinforcement learning community. And you can safely say that this is the most popular version of a Markov decision process today, even in the AI world right.

But for our practical purposes, these are all interchangeable, minimize costs, maximize reward does not matter right. And then there is a model that, you know, came out for little bit of time in the middle and sort of has lost steam. But is there an interesting model called over subscription model where the goals are achieved, they give you reward, but then they do not stop your agent from going further.

So think about a mars rover. And it has many scientific experiments it can do. And let us say it has some fixed length of time it is allowed to add a fixed length of battery or whatever it is. And each experiment gives it some reward in terms of scientific money, right. And so the robot has to figure out should I do this goal first, or that goal first or that goal first and if once I have done this goal, which goal do I do second based on how much battery I have remaining.

And how far I have to go for the next experiment and so on so forth. That kind of a model would be called an over subscription planning model. And here goals are not absorbing so you do one goal, you cannot achieve the same goal again. So you have to somehow keep track of what you have achieved and what you have not achieved. And then you keep moving, right. It is a relatively recent model came out in only 2004 2005 timeframe.

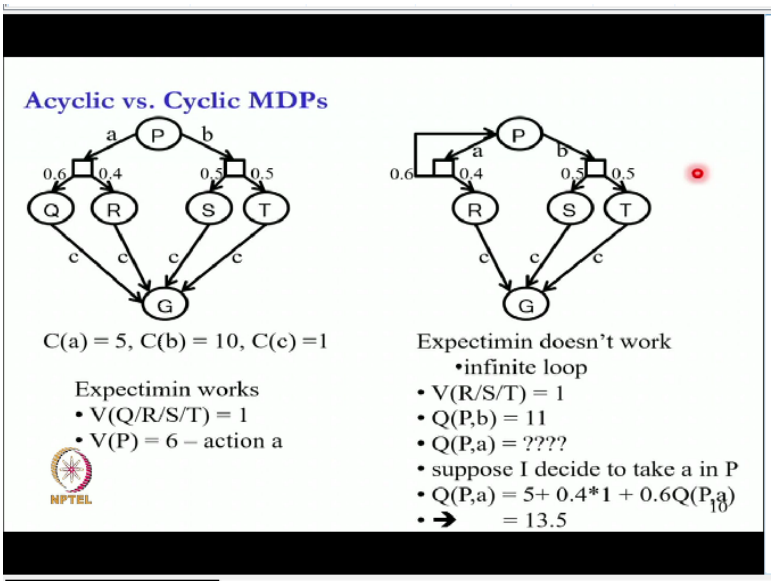
And again MDPs can model this scenario as well. So, the main point I want to make is MDP is a very general model and can make can formulate a large number of scenarios. And then you have to do some practice on different scenarios and how would you capture it in a Markov decision process. So to capture anything in a Markov decision process, you will have to define a set of states carefully. So that your state has full information about what has happened so far.

And what is needed for the long term set of actions what is the agent allowed to do. What will be the transition function and optimization cost of reward. Now let us get started, okay, with this background, let us get started and let us solve our first MDP. And let us first answer the question

how is it very different from expectimax, because we have done expectimax when we did or we have done expectiminimax.

But we can get rid of the mini part and let us just simply do expectimax and then let us see how is it very different from expectimax or expectimin.

(Refer Slide Time: 12:09)



Right if we want to minimize. So, in this case I have a minimization problem. So, let us say I have this model, I will start in state P and I can take action a which leads to Q and R with different probabilities and I can take action b which leads to S and T with different probabilities and from those I can reach to the goal by taking action C and action C costs one action bcost 10 and action a costs 5, very simple problem right.

And if I have to ask you, what is the minimum expected cost to reach the goal starting from state P, what would you say okay, which is a better action, action a or action b. Action a can everybody see this. Why is action a better. It is cost is 5 whereas the cost of b is 10. And moreover everything leads to the goal with downstream same thing. So, if I have to ask you what is the expected cost to reach the goal starting from state Q, what would you say.

Okay, one, this is something where I want everybody to come in because if you get this and the next few slides, then you get the intuitions if you do not then you know life is hard. So, to reach

the goal starting in state Q, I will have to pay an expected cost of 1, to reach the goal starting state P I would have to pay the best expected cost of 6 right, because Q R S T all have costs 1 and then some state P we know a is a better action.

And so then we can compute the expected cost there, which would be 0.6 times 1 + 0.4 times 1 + the cost of a and the cost of a is 5 that gives me the value of P as 6. Everybody with me, any confusions on this small example. So then yes question Ruba. So, the question is to find the minimum expected cost from path some state P. Why did not we consider the action b, and we did not consider the action b because both actions a and b are in our control.

And we can consider action b and that path will lead to a value of 11 and now I will have a decision problem do I pick the action that takes me to the goal with cost 6 or do I take the action that takes me to the goal with cost 11 and what would I say cost 6. Therefore, I will take action a, and all this processing I assumed you did this in your brain, but we will come to exactly how to do this, you know, explicitly good.

So now comes to the real question, because if this was the only problem we were solving expectimin or expectimax would work. The problem is we will have a state space like this or we can have a state space like this. So use your notebooks take 30 seconds and tell me which action is better action a or action b. And moreover, what is the expected cost to reach the goal starting in state P okay.

So let us come back. Let us see what happens. So the first thing we realized that expectimin does not work. And the reason expectimin does not work is because expectimin assumes a search tree. And it can probably handle a search graph, which is acyclic, but it cannot handle cyclic graphs. The reason it cannot handle cyclic graph is because what happens when you take an action a is that with some probability, you come back to the same state.

And if you come back to the same state, you know, how would you back the values up. It is not going to be straightforward. Now, here is a simple thing we can recognize, what is the value of R

S and T 1, because there you still have to do action C and you get to the goal by paying cost 1. So value for R S and T is 1. Moreover, what is the second thing you realize.

Is that the value of taking action b in state P is 11 right. So, this is what I believe pretty much everybody would have recognized by now, that a value of R S T is 1, D value of taking action P in state P would be 11, because I will pay 10 and then you know, I will have to pay 1 downstream and then it will be total of it. Now, what we do not know is what is the value of taking action a in state P if a was the optimal action.

We do not know whether a is the optimal action or not, but let us say it was the optimal action. If a was the optimal action, then I would be paying how much to reach the goal that is the question and it is a very simple answer. The simple answer is that I can write down this problem as a system of equations. So, we can think about suppose I decide to take action a in P, then what is the equation for Q of P, a.

The Q of P, a would be first I will pay cost of 5 then with 0.4 probability I will pay a cost of 1, but with 0.6 probability I will come back to P and decide to take a again. So, the cost I will pay afterwards after one step would be again Q of P, a recursively defined. Therefore, my equation would be Q of P, a is nothing but 5 + 0.4 times 1 + 0.6 times Q of P, a. This is the key step. If I say that Q P a is the expected cost to reach the goal if I take action a in state P.

And if action a is the best action, right. Let us say I have defined it like this. This is not exactly how I am going to define it in the future in the next few minutes, but let us say this is how I define it. If I decide to take action a in state P because it is the optimal, then Q of P, a would be, first I pay 5, then I pay 0.4, but that is for one outcomes. For the other outcome, I pay 0.6 times myself. So therefore, if you solve this, it will become 0.4 times Q P a is equal to 5.4.

And therefore Q P a is equal to 5.4 divided by 0.4, which is going to be equal to 13.5. And now you can ask the question is a better or is b better which one is better, b is better because we can get me to the goal in 11, whereas a will take 13.5 to get me to the goal. And therefore, we will

say that V of P, V that is the value of P would be equal to 11 and now 13.5. This is the kind of reasoning we want to do in a general Markov decision process.

I took some liberties by creating this example and by doing this maths, so, we will do it the right way now, but this is what differentiates a Markov decision processes, general Markov decision process with the other algorithms that we have studied in the course so far okay.