INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

NPTEL

NPTEL ONLINE CERTIFICATION COURSE

REINFORCEMENT LEARNING

Options

with

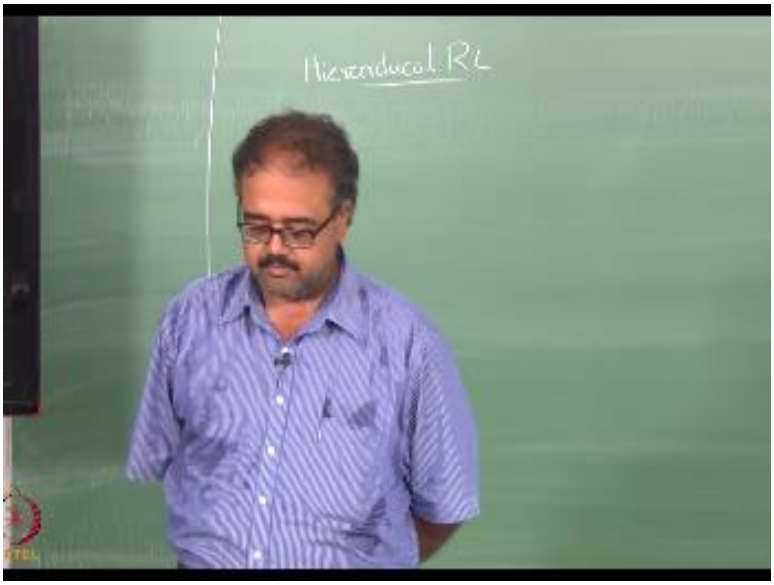
Prof.Balaraman Ravindran

Department of Computer Science and Engineering

Indian Initiation of technology, Madras

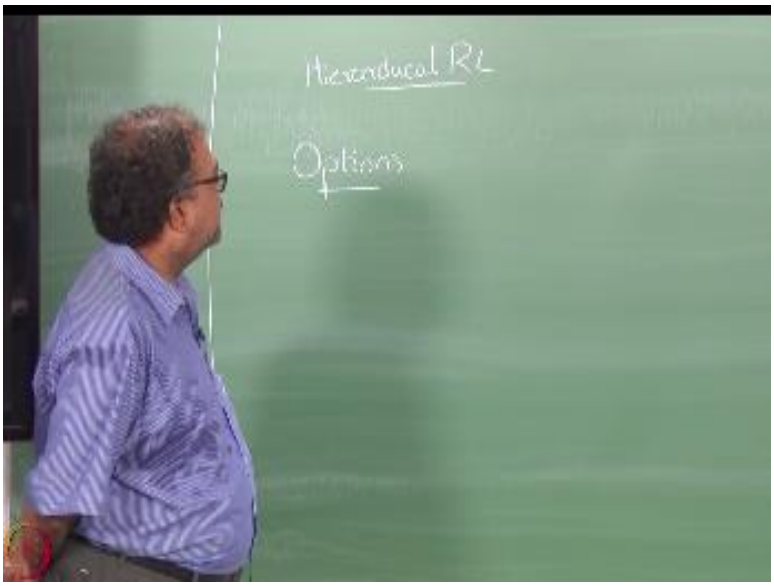
So we are going to continue looking at hierarchical reinforcement array and so.

(Refer Slide Time: 00.24)



What we saw in the last class was essentially a very long drawn motivation for why you want to do hierarchies right and then I spoke about a lot of issues that might creep up right so it is more at a very, intuitive level right. And finally towards the end of the lecture we actually looked at SMD piece right so hat which is a model that many of the hierarchical learning architectures use as the underlying, mathematical formalism okay.

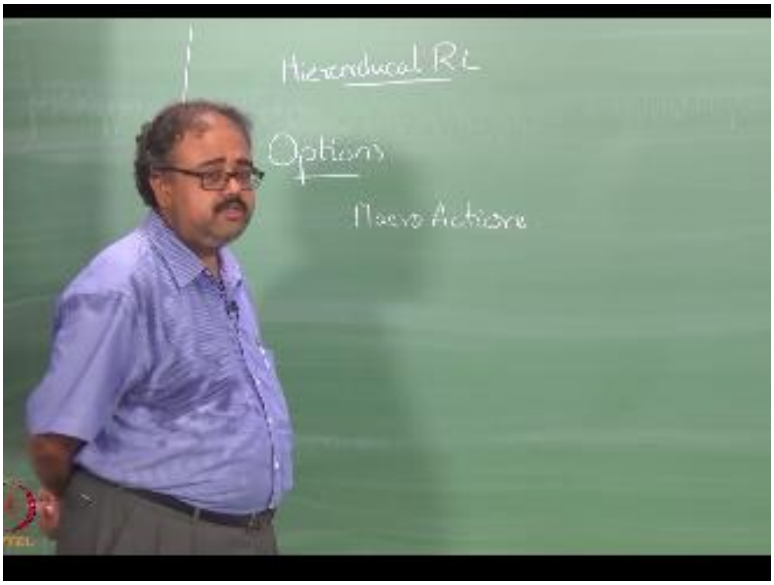
(Refer Slide Time: 1.18)



So what we will do today is start looking at some of those some of those hierarchical frameworks okay the first one we look at this is called the options framework. Is chronologically this is not the earliest but in some sense it is the simplest of the hierarchical frameworks. And therefore we will start there so the basic idea behind options is something very, very, very simple.

So I am going to take the solutions to these sub problems that we are defining right so I just take the solutions to those right I will encapsulate them as a single action so the idea comes from the planning literature actually where people talk about something called macro actions.

(Refer Slide Time: 2.00)



Right so in the planning community macro actions are essentially this is sequence of actions put together and then we talk about it as a single action right so for example we spoke about many such examples in the last class right and so what we get down the go out of this building I can have a macro action that says climb down the stairs.

Right that consists of a lot of smaller actions but then as far as I am concerned all I need to think of it is a single action that is climb down the stairs right so and in the planning literature you just have a sequence of actions right but then when we move to the reinforcement learning space there is something else that we need to take care of what is that I can't just blindly specify a sequence of actions let us take the go down the stair case.

Can you just say okay standing front of the stand at the top of the stairs okay then put your right foot out down once put your left foot down once and keep doing this. I don't know120 times and you will reach the ground floor right. So can I specify a macro action like that, what could go wrong? However okay let's assume that I get the floor right there anything else that can go wrong balance let's assume I get the balance right so direction this you will get the direction time but

Where can the direction go wrong there is a step side steam direction yeah I'm assuming I've learnt all of that people could just becoming up the stairs right I would have to stop move to the

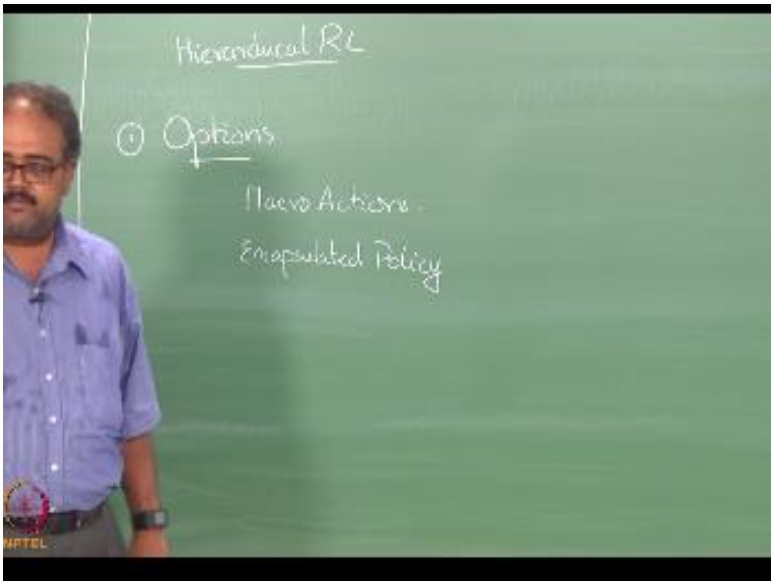
side let the more I can't just keep bulldozing and all that is already there no these are things are dynamic right see the direction the staffs turns and all those things are static so I could I could plan for all of that and then say okay.

This is the day act direction in which I will be taking my steps and so on so forth right so all of that I could plan for but then there are other dynamic components for the world right and suppose there is some kind of slipperiness associated with the steps right so. I keep going down and then suddenly the homestead I could skin and then I met actually fall down a few steps and we not really fall down but then like skip ahead a few steps right or suddenly there could be no this happens.

In the other day it happened to me when I was climbing down the civil engineering department thoroughly as this one, one step there where this is big chunk of concrete taken out of the step right. I was talking to somebody and then I was actually holding a cup of coffee in my hand as volcanoes suddenly there is no step there okay skip the step and actually had to step on the next, next table this kind of put throughput so it's right off right and then I had to actually go down a few more steps very quickly to regain my balance so this kinds of things happen but

These are all dynamic things you do not know small changes that are happening and so on so forth we are actually all this while when we are talking about solving the reinforcement learning problem we have been assuming we live in a stochastic world right so, so things can go wrong right so what do you need to specify to get a macro action right, not a sequence of actions but short and then the end I need the solution to the problem right close observations right.

(Refer Slide Time: 5.53)

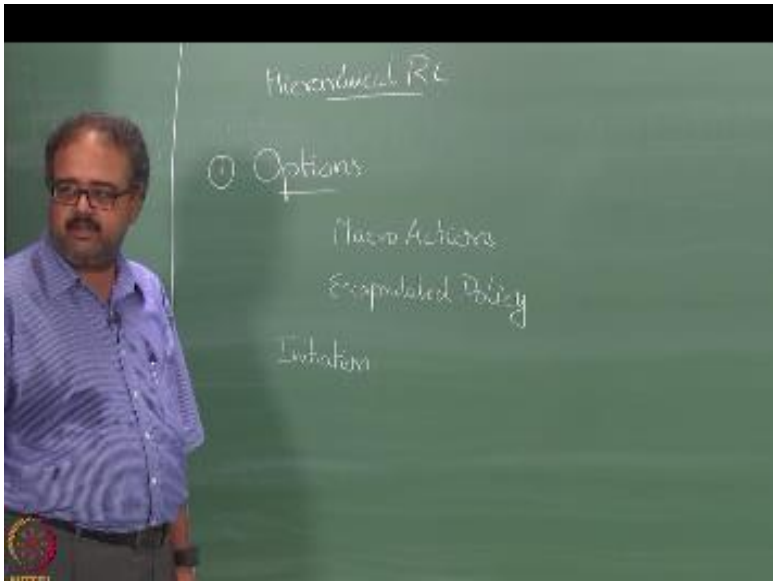


So what you need to specify is a policy right so an option is essentially some kind of a I'm going to call it some kind of an encapsulated policy okay great so now let's go back to the example I was actually giving you already so how did I describe this macro action for getting downstairs. How did I start the description I say stand in front of the step so I need to be able to tell you where these macro actions are applicable right I need to tell you where they are applicable icon just say or you get down the stairs now right so if I say get down the stairs here.

I mean it really it won't work right so I have to be actually there this is not a new, new idea even for simple actions I need to be able to specify where the action is valid right so I cannot just say no playing chess and the first move cannot be captured the king right so, so I have to set it up that has to be there has to be a mate condition before I can say go cut the king right so, so those kinds of things I need to tell you where to start and then I need to tell you about to end right.

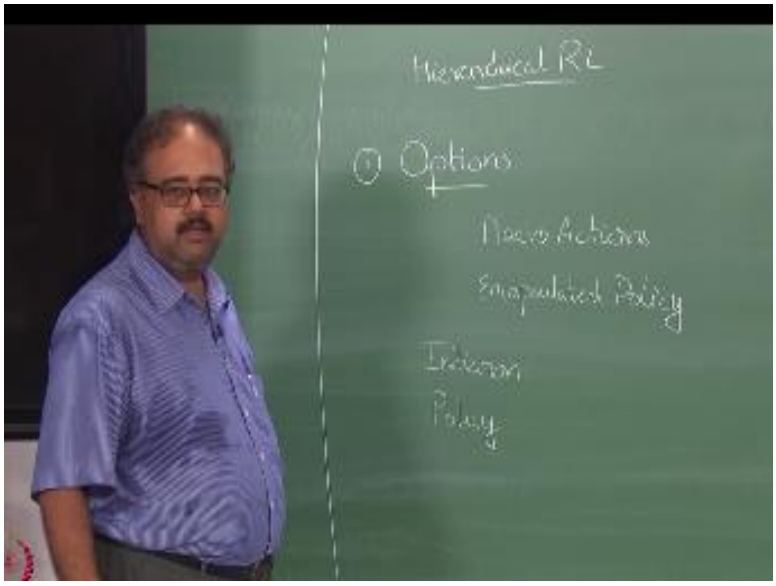
If I if I with exactly are selling you do you do this 130 times and then you will end at the ground floor right so I need to tell you at to end so I put all of these things together so if I want to specify a useful macro action.

(Refer Slide Time: 7.22)



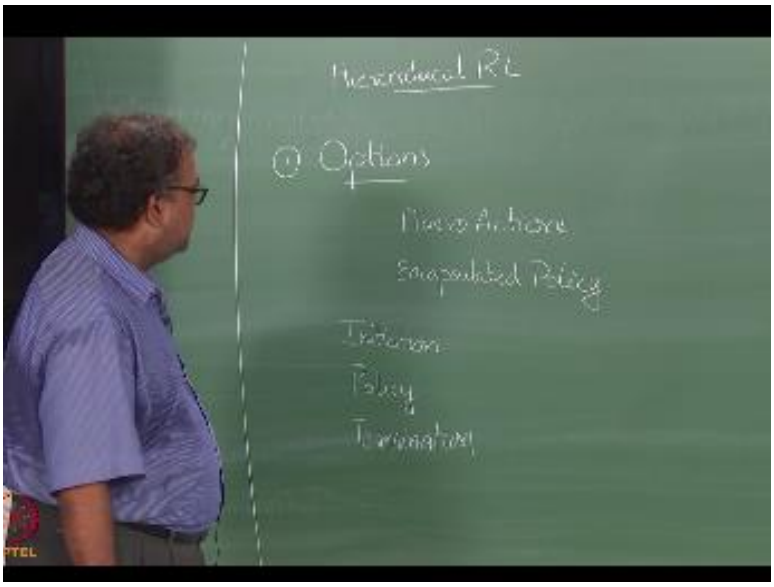
So I need something that tells me where the initiation can happen right when

(Refer Slide Time: 7.39)



I need a policy to use while the macro action is executing and then

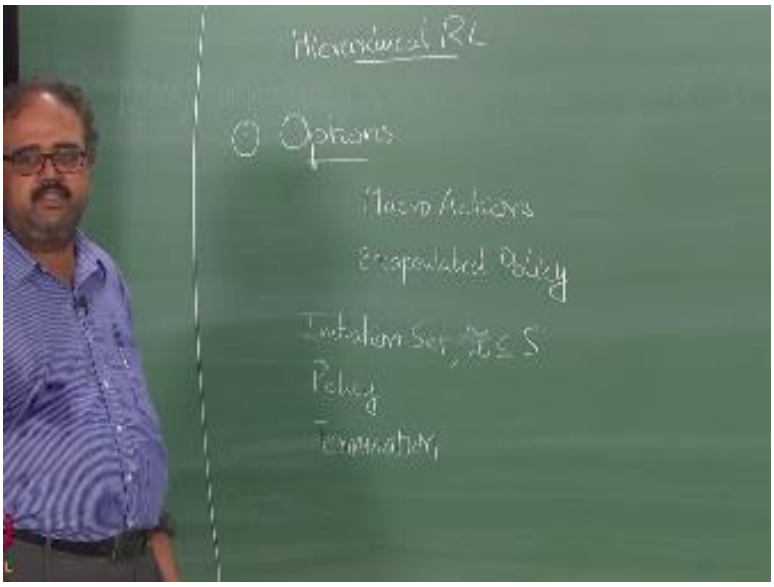
(Refer Slide Time: 7.45)



I need I need a termination ok sorry it can be started in states I agree but why what do you mean we within the particular encapsulated cancellation is just a policy right just a policy so assume I assume that wherever I can start the policy I will have some action defined for that policy right so let's listen talk about that little further as we go along it so,

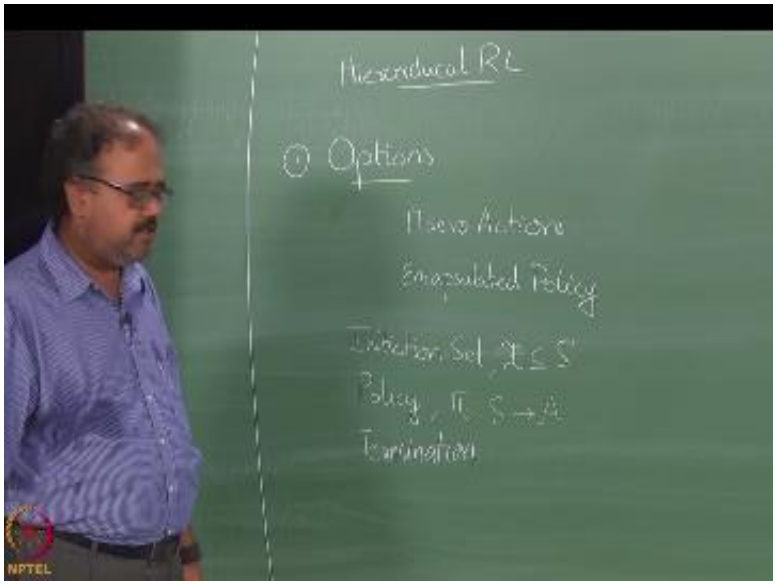
So the initiation if you look at the, the classical options framework right I mean is proposed in 99 proposed we can suppose we can start calling it classical now if you look at the classical options framework they usually give this as

(Refer Slide Time: 8.54)



A set of states right so initiation set if noted by I is a subset of the state space okay.

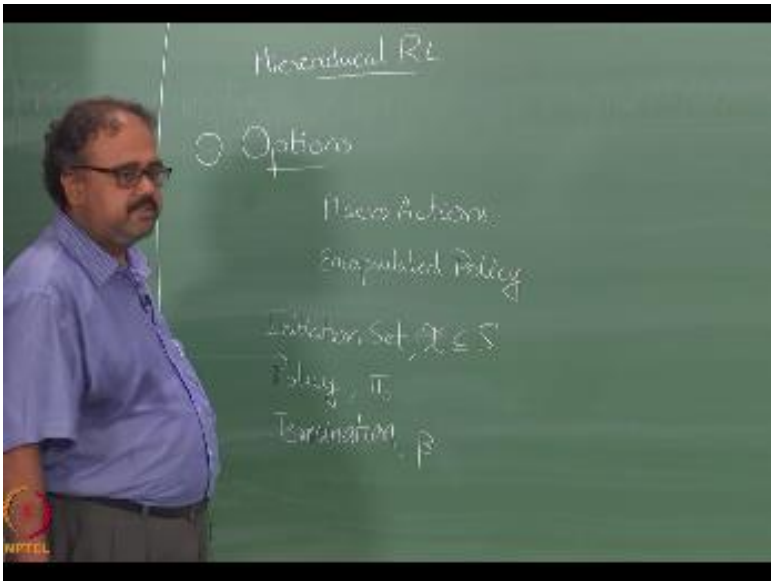
(Refer Slide Time: 9.05)



And then the policy is given as if it is a deterministic policy it could be this right or it could also be given as a mapping of yeah or it also be given as a mapping of history elaborate a little bit more but can be given as a mapping of a history of things that have happened since I started doing this to the next action right so that allows us to define options like we go three steps forward and then, then right.

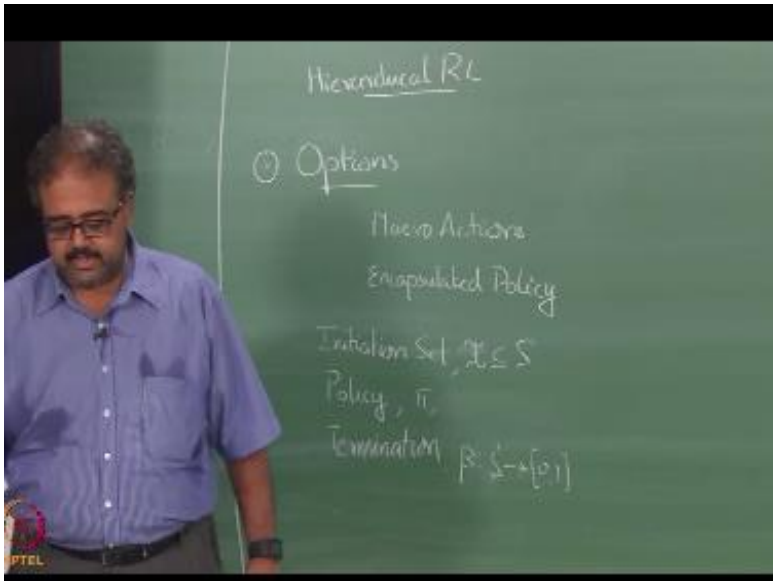
It might be useful right so what happened there when I said go three steps forward and turn right automatically added history to the policy so I need to count okay how what was my last action was it go forward okay what was the last two actions where they are both go forward what are the last three actions were they both they all three go forward right and then if that is the case then I do a right action right so I could define it like it I will come back for it and

(Refer Slide Time: 10.32)



The termination is typically specified by beta right

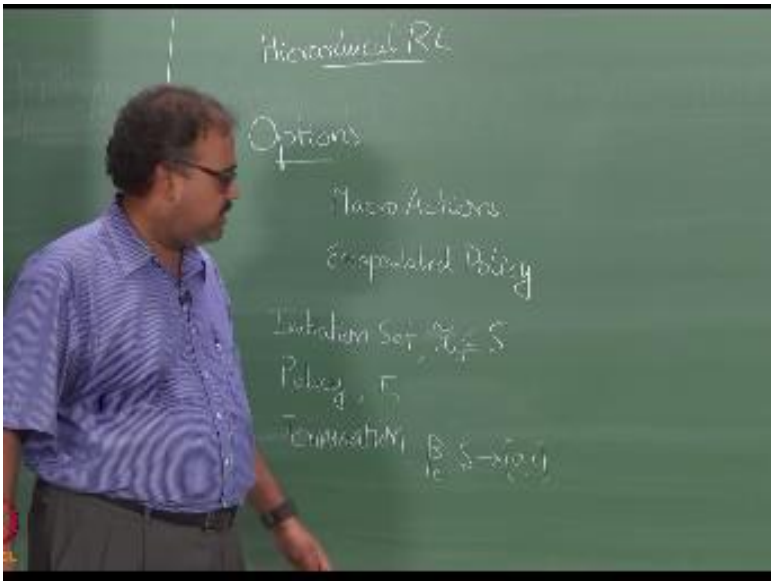
(Refer Slide Time: 10.32)



Is in the original definition of the option is given by a probability. So beta of swill be the probability that that option will stop in status right so for many states the probability could be one because that means on all those states the option if you reach that state option will not continue. To the agent is it in our air we started as we never space with a goal to the engine but you're specifying a termination point to be yeah but this is not the RL problem and solving it this planning no, no it's just an action I'm defining it's not a problem so suppose you say.

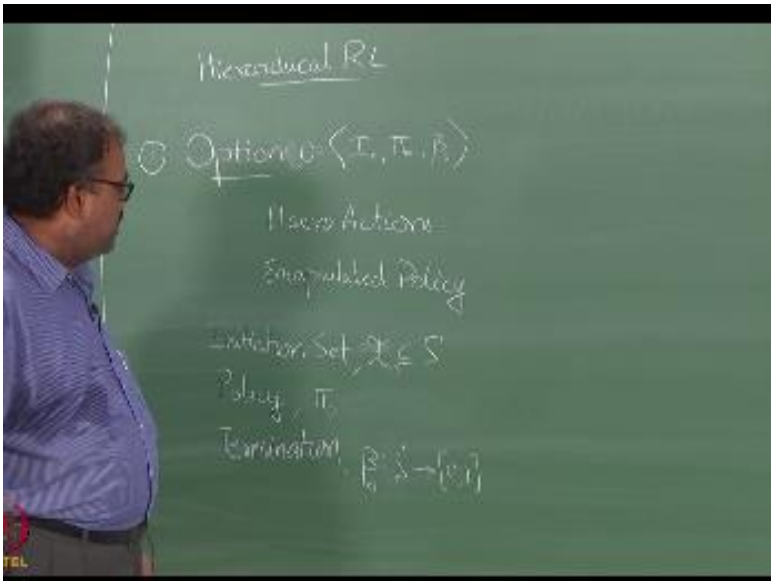
You can't tell me that Oh in the agent if it goes north you can't you shouldn't say that it's actually move upon step right I mean that's the dynamics of the problem I'm just defending the dynamics here so right so these are the three components that make up an option right so for a particular option.

(Refer Slide Time: 11.43)



I can probably add a coat everything so I have the basic MDP which some states actions and so on so forth and an option is going to consist of

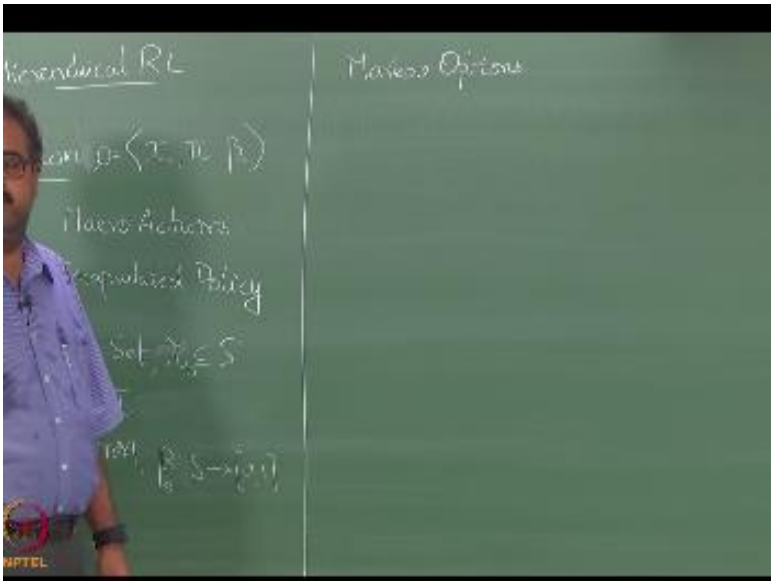
(Refer Slide Time: 11.59)



These three things so this probabilistic definition of options was actually introduced to for more of a mathematical convenience right so, so that they could show some proofs more elegantly in practice. When you define this options you typically make the probability of termination 0 or 1 so 0 means it does not terminates, 1 means it terminates right so you can think of it as an indicator function that indicates whether the, the set belongs to the termination set or not.

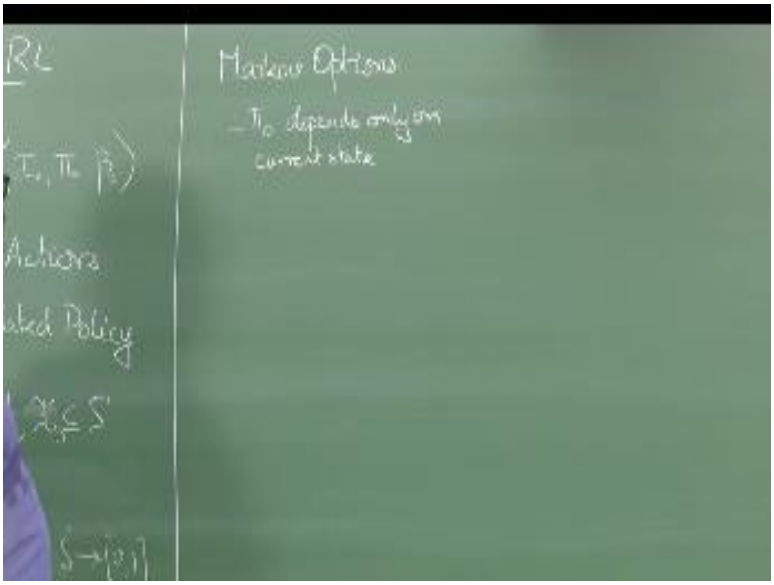
Let us think of the initiation and the termination be sets okay.

(Refer Slide Time: 12.52)

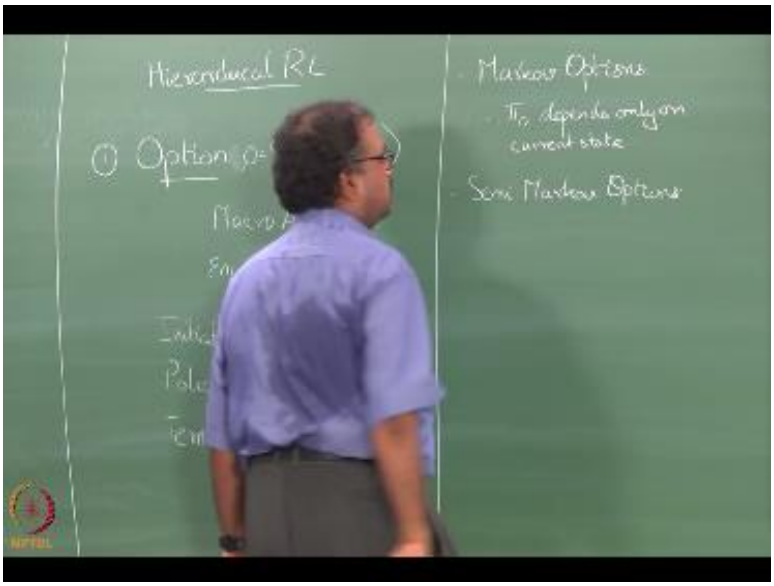


There are two kinds of options right so which mark of options which are easy enough to define so what are mark of options. Sorry, not exactly so Π is not defined only on the current stage right so these are mark of options so the nice thing about Markov options is I do not have to worry about where is started the option it is a wherever it is it will be the same action right so if I am in a particular state and I am executing that option then I will take the same action.

(Refer Slide Time: 13.45)

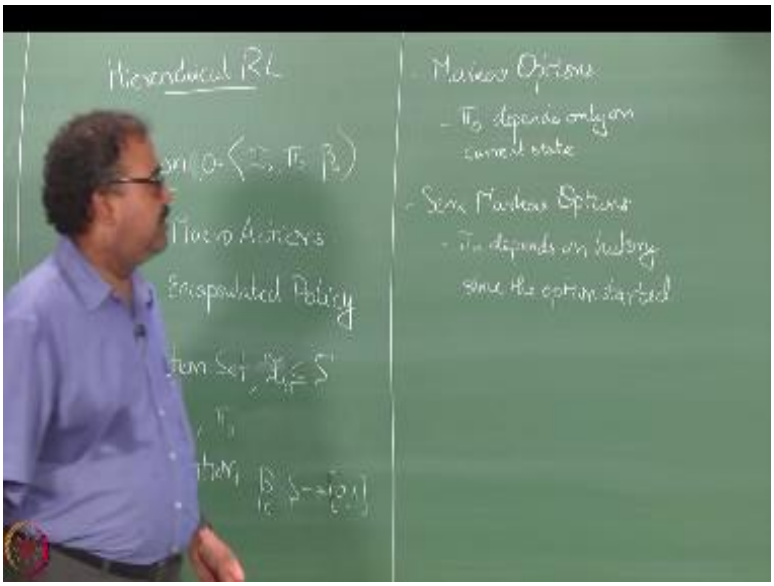


(Refer Slide Time: 14.11)



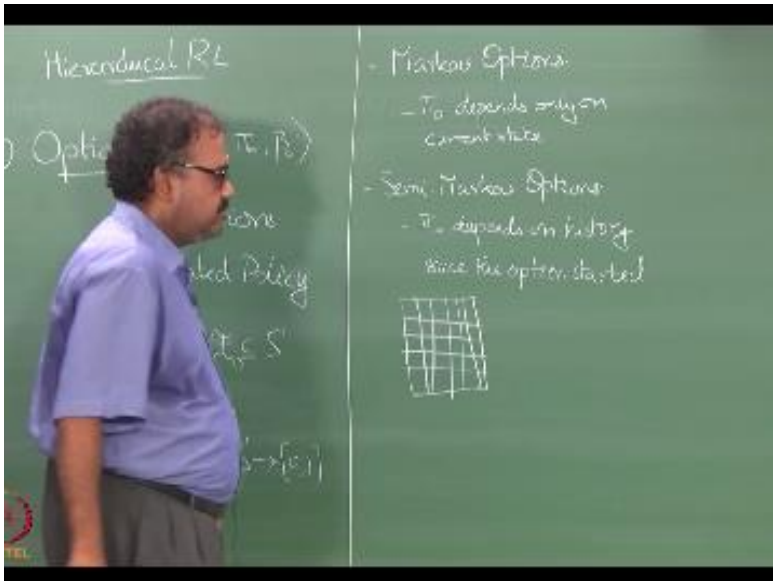
In the second class of options called semi Markov options. So what are some of our coefficients where's my second rotation sequence of states since the option was initiated right if you remember I was telling you why is a semi Markov because during the execution of the option there is a dependence on history or could be a dependence on history we do not know right now I was telling you at the I mean when the holding time is an operation you really do not know how the system is behaving.

(Refer Slide Time: 15.03)



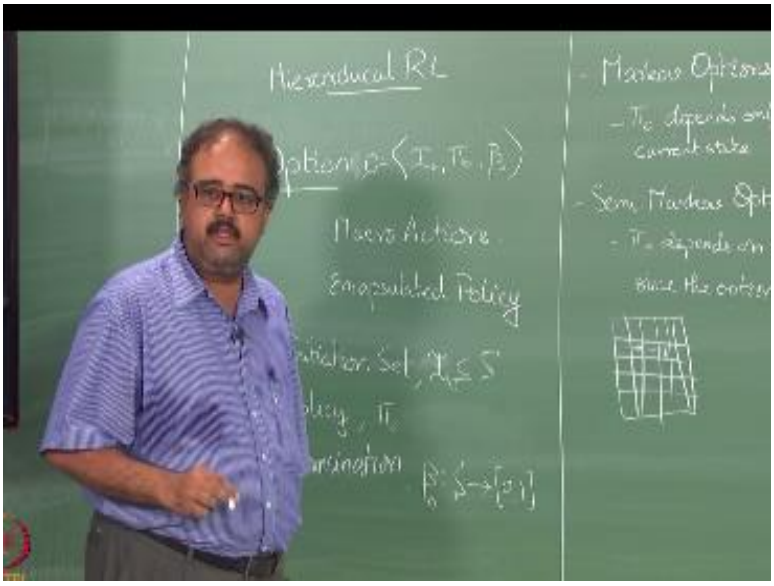
So and I say that is the reason we are calling it semi Markov so in a semi Markov option so Л not depends on another personal history since the option started so this is since let's say

(Refer Slide Time: 15.28)



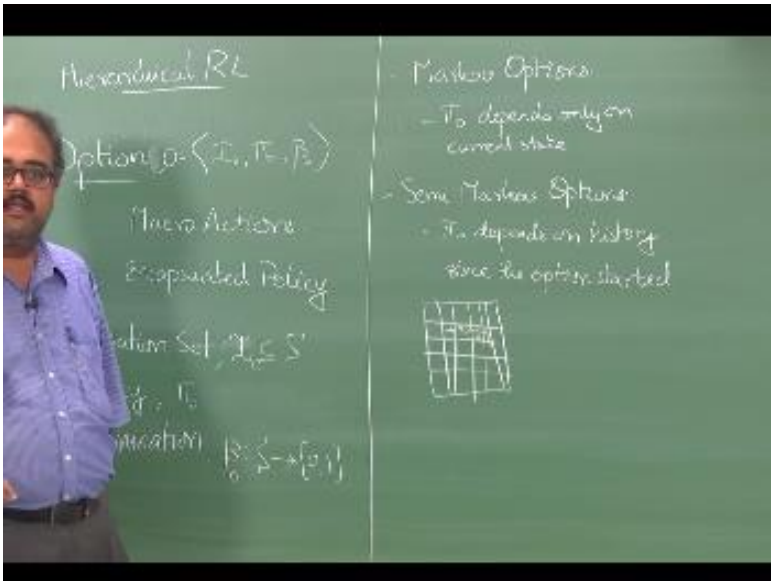
I have a if you start here and say go right two steps and down one step right

(Refer Slide Time: 15.25)



So it will be this, this like 1 2 and then this right so in this state I go right in this state I go down right but if I start at the option.

(Refer Slide Time: 16.13)



Here and set go right two steps and one down one step it will be this, this and then that right so what I do in this state depends on where I started the option right. Whether I go right here or that I go down here depends on where I shorted the option if it has been a mark of option right it shouldn't matter where I start the option when I come to the state I should be taking the same action.

So there are many instances where we want to allow your agent just to do a larger amount of exploration okay so sure how many of you have started finish all your programming home works Q learning right if you make me think of how hard it would be if you make the puddle world 10,000 x 10000 puddle world these people try it this is the part of the assignment what is for us

assignment then well we didn't give them a function approximation assignment enough on the 12 by 12.

What's the fun in that when did you change the film's problem segment I thought I thought I get a thousand a thousand at least a thousand two thousand twelve at all there is no fun what do you want to do function approximation on a 12 x 12 it for atleast a thousand a thousand be dual. What is it that's a fun, and it is a fun part what do you miss all the policy what is the need to store the policy really the Q function right.

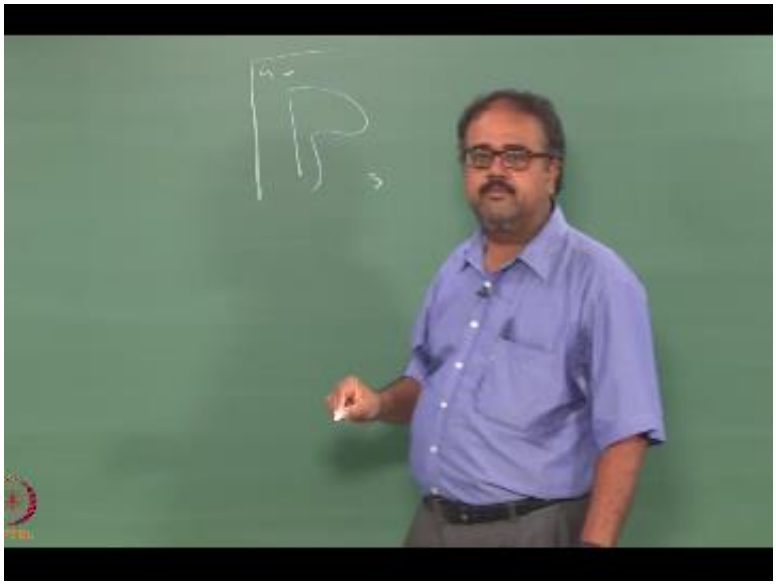
That’s book keeping, I won't care, I mean learning curves. Computer is a very smart comparisons kids can you figure out how to do that automatically yes we just look at it blow up the puddle and figure out rules for where the puddle should be read and then write a small function that I initiates the puddle one that is the easy enough to do until we've gotta work anyway how long would Q-learning take to work in the thousand a thousand good work or 10,000 x 10000 gradually.

Saw how long it takes right in the text even in a 12 x 12 grid world it just takes awhile to do all the exploration and soon so forth in situations like that you really want your agent to have the ability to do crazy things in one shot right so for example imagine a thousand by thousand great well where I define an option that says go in this specific direction for fifteen steps so I can just go north 15 steps or go east 15steps.

See in the puddle world is not really clear to me what are useful options to begin with right I mean what are the useful options, there are no doorways what are useful options in the puddle world now if the puddle rupees but the kind of a puddle world I give given the RLX that one big puddle in the middle of the world really no bottlenecks or anything that.

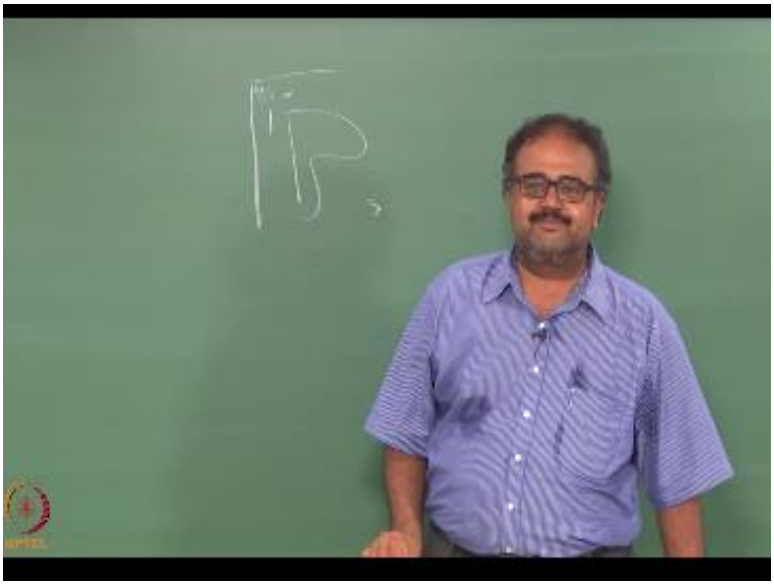
I mean there are bottlenecks but it is not immediately obvious to you what the bottlenecks are the bottlenecks would be the, the edges of the puddle before it ends.

(Refer Slide Time: 21.03)



The so essentially you have the puddle like this right, right and then there is a goal here and they start state here that this would actually be a bottleneck if you think about it. Right so this could be a bottleneck so you need to get there before you can actually get to the goal okay.

(Refer Slide Time: 21.25)



So that will be that will act that those states will act like bottlenecks so but then it is not that is not useful really that interesting and option right so what you could do is just define or one was here wherever you know you kind of can guess where the option bottlenecks would be right so you can kind of define options that they essentially make it more efficient explore by saying that go up 15 steps right now go right 15 steps right.

So that way instead whenever you pick this action by accident when do you pick it back accident when you are doing exploration right your epsilon greedy thing when you pick go up the 15 state suddenly you break out of your random walk loop so you move away 15 steps and then you start doing a random walk there right and then you say you take another action then you will move away another15 steps and you start doing a random walk there that way you essentially can explore lot more of the state space then you can if you are just working with primitive actions so in such cases having these kinds of options are helpful.

IIT Madras Production

Funded by

Department of Higher Education

Ministry of Human Resource Development

Government of India

www.nptel.ac.in