NPTEL

NPTEL ONLINE COURSE

REINFORCEMENT LEARNING

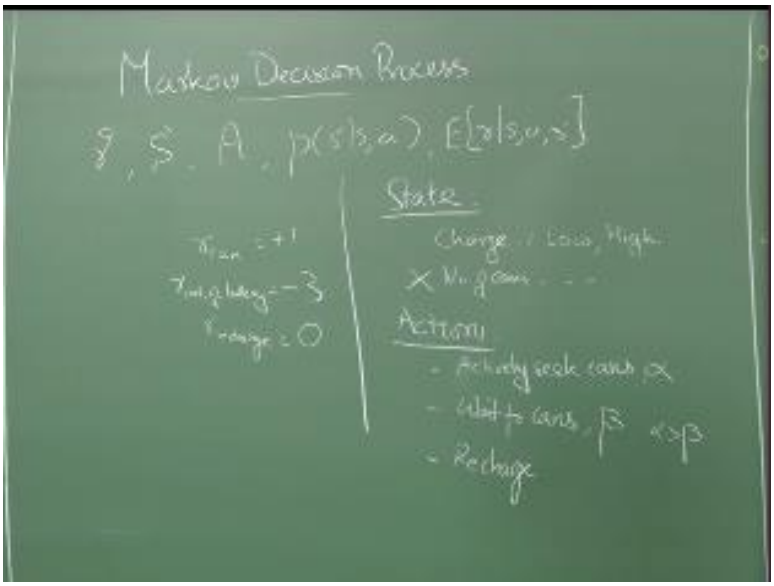
MDP Modelling

Prof. Balaraman Ravindran
Department of Computer Science and Engineering
Indian Institute of Technology Madras

So we will continue looking at the reinforcement the full reinforcement learning problem so we left off where we defined a Markov addition process right and so today's class will probably be a short one okay because the next thing next topic I need to start requires me to define a lot of rotations and stuff like that so if I start that today I will probably end up just defining the notation today and then when I start again next week I will ask you to I will have to ask you to recall the notation and stuff like that or spend time writing it out again so I will not do it.

So if I finish whatever I want to before the proofs start right I will just stop today so I will not start the notations okay so let us see how it goes so right.

(Refer Slide Time: 01:09)



So we started looking at Markov addition process plastic so we said that it is specified by set of states actions right transition probabilities and right and possibly a γ so if you look at the RL textbook right they leave out the γ right if you look at any OR textbook were also mdps are used widely right they would have the γ in there right so for them g is a very inherent part of the problem but in the end the RL books you would see that they would leave out the γ even though I mean gamma is inherent part of the problem and somehow they just do not acknowledge the fact that gamma changes the solution right licit Lee in the problem definition itself so this is essentially what an MDP is all about so what I will do is walk through one of the examples from the textbook right.

So I mean sticking is a textbook example so you can go back and read the book and also because it is kind of makes the course more related to the minor guys here okay, so let us take the example of a recycling robot from the book so how many of you have been reading the book okay the number went up by two since last time okay good Saudi remember the recycling robot so what was the thing about who is in any searches for the Cal State is where it wait for some someone to come and take an action between these three state based on the battery level reference ah based on that they will be reversed whenever it is where does one will be even okay.

Now it runs out of battery otherwise hmm okay so it is interesting they just turn on the mice or something an interesting that you used the description states okay so people heard it and if I will repeat it since I seem to have the Lord Mike so you know that those can also pick up but I seem to have the Lord of Mike so let me repeat it so that is this robot okay that has been created for the express purpose of switch b hearth okay so it goes around the building and finds all the empty cans and gathers it and puts it in some kind of a no trash recovery system recycling system or whatever.

So for every can the bot gathers is going to get a reward of 1 right this is a number I am pulling out of the Hat right so let us not talk about the reward in all right so the robot is supposed to gather Kemp's that is essentially what we wanted to do but being a robot it is going to have a limited amount of power that it can spend right it has a limited amount of battery power that it can spend so it has to choose carefully how it spends the battery right it can go around looking for cans right and if it is going to run out of battery then it can do one of two things it can recharge itself right.

Or it can to sit in a place like a trash can and expect people to drop the empty cans into that but okay just kiss sits in a place and the people can just use it like a trash can right so this is essentially the set up so I have this but that should I would like you to go actively gather cans but I also do not like it to run out of battery because if it runs out of battery somebody has to go physically either roll the robot or carry the lift the robot and take it to a charging station and charge it ok that is a very painful operation so I really do not want it to run out of battery.

But then I do not want it to just sit idle action as well because what it says is supposed to do is gather camps right so how would you model this problem so what are the states and what are the actions what should be the reward robot seems to be easy right so well what you wanted to do you wanted to gather Ken so whenever it gathers can whenever gather sir can you give it are ward right so never gather second to give it a reward of +1 right so I am going to call this our can +1 right then what you do not want it to do you do not want it to run out of battery.

So yes it is just yeah something large right so yeah so I would I would prefer to give it -10 or something but since I am walking through the example in the book okay so it is a - 3 right so it is not that bad I mean if we can collect 15 cans and then die is okay right so it is not too it is not that bad thing so what else if it charges if it chooses to charge right so what happens do we give you what do we do for it is thing isn't even a positive reward or why does no reward make sense he'll give yeah just the heck for pick up getting that reward will get recharged right but then recharging is also good we want to convey the fact that recharging also good.

So it is better than running out of battery and if you recharge you have the potential for getting more rewards in the future because you go out and seek more cans right so that way you have an indirect incentive for recharging because in the future it allows you to do something better if you just look at the immediate reward it looks like a bad thing but if you think about it so in the future you are making it more attractive for recharge I mean if you look at the future outcomes recharging is also very attractive okay.

What is below are so many do if they were charged no this is not this is if you think about it this is more to do with actions or something with outcomes yeah battery levels you could you could make it all the more detail you can make it as detailed as you want right now comes the crux how detailed do you want to make the representation to be do you want to make it a real number saying how much battery charge I have you could get a continuous state there right so what do you want to do you want to do some kind of dis cartelization right so if I want to do a dis cartelization so how do I discredited ah that resistant question are completely we forgot answers that question what is the state?

The battery charge level should be the stay right and then anything else the number of cans no let us just assume that magically it takes a can and transports it to well as an interesting if it has a finite carrying capacity then you probably need to throw in how many cans it has and so before it has to make it trip to the dump dumping ground or something like that so let us assume for the time being it so as soon as you put a can into that there is a teleported that takes it to the dumping ground okay.

Or the bot has an infinite carry I am just trying to make the problem simpler right if you want to make the problem more and more realistic right you have to think about a whole bunch of other things right for example I am kind of making it at a very high level I am not talking about the person I mean the post of the robot let the velocity which the bot is moving so all of these things you need right if you want to actually going to control the one and then this assuming all of those are somehow going to be magically taken care of right.

So we are just assuming that you are only worried about the charge of the bot right yeah we have been talking about only discrete States in discrete action so far we have not talked about continuous States and continues actions but you have continuous state contraction place okay, so the state is going to be the charge right and possibly number of cans right other things but for the time being I am just going to say I am not going to care about all of that we will just look at charges to keep it a simple example right.

I am also keeping it in sync with the book so and again so what do I do with charge I can treat it as a number in which case I will have to figure out how to handle the continuous value right all I can start discrediting it so how will I discredited it yeah so I again well percentage is also not a dis cartelization unless you round it off okay. So yeah so you could do something like that or it could be even more aggressive right I could do low medium high right and if you really stop and think about it when would you if I said if I do low medium high or something like that whenever I actually change my decision about charging versus not charging only when I low okay so in some sense you could even try operating with high and low and or sufficient and low okay.

Or whatever is it build is also probably actually I do not I am not having the location at all give it so assuming that there is always a convenient charging point located nearby okay just not like that a single base station I have to go there so as soon as I find out okay I am running out of search and let us assume that as long as the battery is low I can still get to a charging point in fact I am going to assume you in something stronger so I could with some probability keep operating even if my battery is low right.

And I still not run out of assumption I make is the following whenever I say recharge okay I will have enough battery power to reach the nearest recharge station and recharge so after I pick the option to reach other I will not run out of battery power is again very simplifying assumption right this makes it easier for us to draw pictures as I will in a minute right but all the points that are being raised to our very good points right so when you start trying to model a real system right this is all calls that you will have to make right.

So what I say that I would throw into my state right what I say that I can ignore right and so on so forth and yeah so Paulo we do you remember the homework assignment we are asking them yeah the third one we are asking them we are giving them a word problem when asking him to construct an MDP out of it right you think so you were saying Methodism yeah the one that was released today morning yeah press release yesterday then your time traveling okay released this morning.

Yeah so the new assignment and it is due on feb14 yes 14^{th} is what name Sunday yeah well plan ahead if it is important to you finish on the 13^{th} or as it is more likely for lot of people who have nothing to do 14^{th} only have a whole last question is that right so there is a word problem you expect you to correct MDP, so you can think of similar things but that is a more constrained word problem so you will see that there is not too many ways in which you can construct an MDP out of it.

So but then the idea is that a lot of design choices that you make when they give you a problem specification that the Lord of choices that you make when you go from the problem specification to an MDP right so when you are talking about designing a reinforcement learning solution for a problem the things that you have to decide are the states the actions and the rewards these are the three things that you have to decide and when you are constructing an MDP additionally you will have to figure out what this the transition probabilities are also going to be because we thought that you cannot solve the problem.

All right but in we are trying to use reinforcement learning techniques to solve these problems as we will see later on you do not need the transition dynamics I do not need to explicitly model the

transition dynamics but you still need to make choices for states actions and towards so that is where we are right now so after we finish this I will tell you some simple way of consulting the transition probability for the robot but that is fine so essentially what I am going to do charge this is going to be now low or high okay.

And then what about actions there is no pick up let us say when I actively we find that can I will automatically pick it up okay so I will go so explore essentially or as they say in the book actively seek cans because explore is overloaded right so we have explore exploit those kinds of things already so I do not want to have an action called explore okay, so you have you actively seek cans or wait for cans of recharge right so there anything else that we need to define here yeah so I will tell you how we can get around the day that is a good point so maybe you need a zero charge state.

But I also tell you how to get around that right so the Assumption we are going to make is if you run out of charge somebody is going to take you and charge you right then you are going to get this penalty of minus three so what does that mean so from low I can actually go to high with a penalty of minus three ok so that is that is equivalent to the zero charge state that you can make it more complex so right, now I am assuming once you plug it in to charge I do not worry about it until fully charged later right see this is another thing which you with an interesting point because I have discredited it low and high right.

So when will it run out of charge when it when it will go from high to low state so if I is one 100% high right it is probably not going to go to the low state when it take an action but if I say 40% high almost surely as soon as I take an action it is going to go into a low state let us say 40 is the way cutoff point all right so above 40 I say hi below 40 I say low let us percentage is right for 40% I will say hi below 40% I will say low, almost surely as soon as I take an action it is going to go too low.

But then I am not making this distinction I have made a dis cartelization here right so what will it look like to me and if I run the robot multiple times what will it look like to me sometimes it goes from high to low sometimes it goes does not go from high to low it stays in high so what is

going to happen here is I am going to translate this into the probabilities of transition right so what is the probability of going from high to low in fact if I you know if I knew all about my the battery dynamics if I know how much power each and every action takes and if I modeling this is a continuous valued state there might not be any causticity here.

But I might be able to tell you very exactly okay now at this point if this is the action you are doing this is so much power that you will consume and therefore the battery will go from high to low all right so I might be able to give that to you exactly so inherently this might not be a stochastic system I am only saying might not be because people have worked with robots so you know that he met he would specify as many initial conditions as you want right the bots will do what it wants to do right.

So it is still be a stochastic system but it could potentially not be a stochastic system if you know everything about the world right but we are making it stochastic by virtue of ignoring certain things that we could potentially know about the system right and ignoring this makes the system design a lot simpler maybe we are giving up some things in terms of how optimal or solutions can be but that is a trade-off that you will often have to make right so just the formulation itself is not an easy task way to think about it a lot to figure out toward it be the right place to draw the line okay.

So these are the actions that we have to take so I can actively seek can is wait for ken's recharge is that something else that we have to specify so I specified the stage they specify the actions we specified rewards for events right so if I collector can I get a plus one right if I run out of battery you get a minus three then if a recharge you get is you these are events that could happen in the world right so these are so I have to relate this to the states and actions only then I can tell you what is the expected reward.

I will not talk about time at all right so I told you for that at least for the initial part of the course will not talk about time with how long it run say so what is it that we need suppose I am actively seeking cans what is the probability I will get a reward of +1 if I am actively seeking cans right what is the probability that I will find a can correct it is I need to still specify that if I am actively

you are if I am waiting for cans to come like what is the probability that I will get a can what is the derivative of what rid of which one?

Yeah yeah was it Oh y-yeah maybe may be not I do not know why you chose to move from one to the other night I mean I could have just stopped seeking because of his running low on battery right so if I five min some facility where people are just taking cancer and throwing them the dustbin that I am a just happily go sit next to the can dispensing place and hacked us at the spin rate I will be happily getting rewards but I might not even worry about recharging or anything right.

So I might just choose to stop right so it depends on that depends on the kind of environment you are operating in right so even if I act as a dustbin if do I get enough can send I might not want to seek right so maybe there is some time of the day maybe between 12and 1 30 or something lot of cancel come if I just go stand outside the cafeteria I do not have to actively seek Ken's maybe after one 30 I have to go around seeking cans and I do not know it depends on the system it.

I am not trying to be too overly realistic here but I am just giving you what are the factors that you have to take into account right so if you start talking about what could be optimal policies right then you are going down a dangerous path that means you are going to start designing your problem itself depending on what you think the solution should be which is a bad thing right you should put in all the factors about the problem in the design and then you should go and let the algorithm figure out what the solution should be.

So you should not start thinking about what the solution should be and then try to create the rewards and the model itself I will give you an example of that after I finish this I will give an example of when that became a problem okay so is it clear so now I need to figure out if I actively seek ends what is the probability that I will get cancer right is there anything else I need to know yeah we will come to that.

Now it is related to that if I am actively seeking cans what is the probability that I will run out of battery because that also determines my reward right or if I am waiting for Kansas to come what

is the probability I will run out of battery may be 0 because I have turned off everything and just waiting for or maybe there is a leakage that is happening yes a very small probability that I might actually run out of battery right so. So I need both of these to actually figure out what will be the expected reward right for a state action path.

So I need to know what will be the expected or what so I have to actually think about both of these things if I am in a high state that a time actively seeking cans right so what is the probability I will find a can does it depend on whether I am in a high state or not no right so if I am actively seeking Ken's what is the probability I will get that can okay so that is one thing I need likewise if I am not actively seeking Ken's what is the probability I will get a can that is second thing you need on the other hand if I am in a high state and actively seeking cans.

What is the probability I will run out of battery there are the state matters right and likewise if I am in a low state and actively seeking cans what is the probability I run out of battery that state matters likewise for waiting for can see if I am in a high state and I am waiting for a can yeah probably of me running out of batteries very very minuscule almost you know but I can low state and I am waiting for a can so even with the leakage power drawn I met actually run out of battery so these are the things so you need to know all of these right.

The time being I will say α as the let the probability of you getting a can write and β is the probability of you getting a can if you are waiting and then $\alpha > \beta$ hopefully typically we were talking about probabilities I mean that might be momentarily where β is higher than α like I said stand outside the cafeteria during lunch time but then overall we are going to assume that $\alpha > \beta$ since we have not actually modeled that the time of day anywhere where is this going to say $\alpha > \beta$ right.

So now what else do we need yet we need to define p right so for that what do we need well if I am in state high what is the probability I will go to state low okay forget about going to out of battery that is what we needed for the rewards but for the test state transition so from high to low what is the probability that will happen if I am actively seeking again so that is basically what we

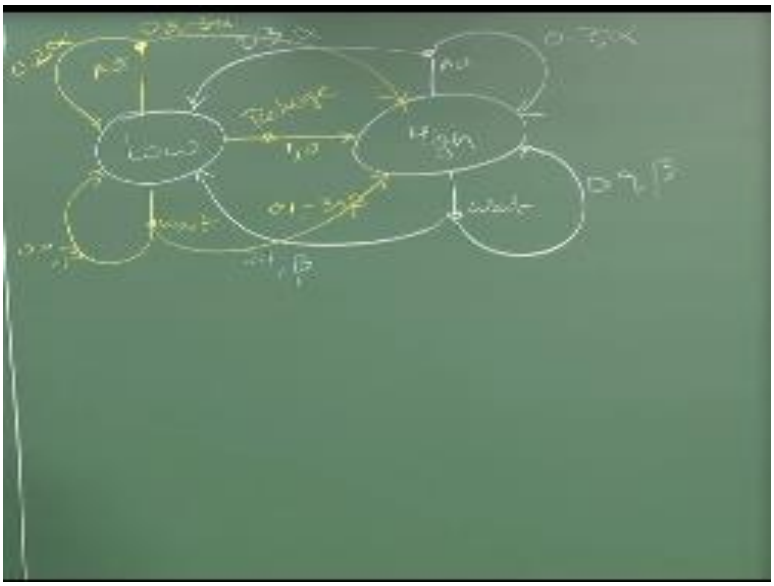
need right so I mean state hi I am actively seeking a can what is the probability I will stay in state high what is the probability will go to state low.

Right so like that and likewise if I am in state low and I am actively seeking a care what is the probability I will stay in state low what is the probability will go to state high and I am in state low and I am waiting for can what is the probability I will stay in state lower the probability will go to state high right and so on so forth so these are all the things that you need for defining the state transition probabilities of course if I am in state high and I do action recharge what we want? Exactly.

So if I am in state low I do action recharged with probability 1 I will go to state height right so we will actually prohibit recharge action instate Heidi well it does not make sense I am already fully charged I do not want to go plugging in again all right so I will fill diesel of that all depends right I could just have a stochastic policy I can say that ok my policy is that if I'm in state high with probability 0 point 8 I will search with probability 0 point 2 I will just stay so I mean so I am not talking about solutions here.

I am just talking about defining the problem this is what I was mentioning earlier also do not do not get into how the solution will look like right sometimes you will have to think about how the solution will look like because you need to know if you are giving the state sufficient expressiveness so that you know when the solution should change and things like that but do not worry too much about the structure of the solution. So now we will just draw this some kind of a state transition diagram.

(Refer Slide Time: 28:10)



Right so let us do the act state high first so there are two actions from state high one and one is actively seek cans right so active so what are the possible outcomes of active so what is the probability that I am going to be in high know some values equal is also possible well depends on how wide my high band is rate and how much power my actor action is going to consider I don't know this is to say some numbers here right what is other action I can take here right and what is the something some numbers it has now is there anything else I need to specify I need to look at what the reward is going to be for having done that right.

So what will be the robot here what yeah I will get plus one for a can but the expected number of cans I am going to collect this α right there is a probability of me getting a can so the expected reward will be α right likewise here the expected what is going to be so from low I again three actions no from low I have three actions or to read a three actions so what are the actions right so I will do the easy one first right from low if I do recharge with probability 1 I will go to high and I will get 0 to 1 right.

So now come see more interesting suppose I do wait here the rein chance that we will go to hide I could write so it is also possible but very low very low probability when I am not doing

anything so if I already set point one here is the probability of discharging I will say point 1 minus and I would still get a can write and this one is likewise if I am actively seeking the set mixes some small probability I met not discharged and they will keep accruing rewards at α right but then I might actually discharged with a high probability and then I go to the high state and I also get the additional minus 3 along the way.

I mean whether you do this + β + α or not is again a choice right you can say that as if I run out nobody is actually going to give me a can right so people might snatch away they can I already have I do not know what you can choose whatever I mean this is again a choice I just I just chose to add the α and β here we can just not do but you can you get the sense of what we are doing here right so is a very simple problem it took us a while to identify all the factors right and we chose to ignore a whole bunch of factors because you identified a lot of useful factors right we chose to ignore a whole bunch of factors or in order to come up with a simple tractable MDP right okay.

IIT Madras Production

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

www.nptel.ac.in