

NPTEL
NPTEL ONLINE COURSE

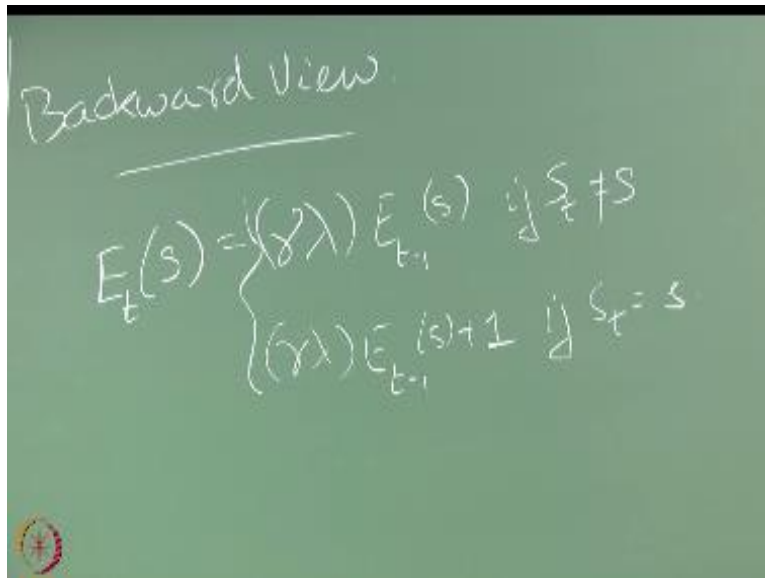
REINFORCEMENT LEARNING

Backward View of Eligibility Traces

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

So no questions on this so what is the backward view do okay.

(Refer Slide Time: 00:22)



Backward View

$$E_t(s) = \begin{cases} (\gamma \lambda) E_{t-1}(s) & \text{if } s_t \neq s \\ (\gamma \lambda) E_{t-1}(s) + 1 & \text{if } s_t = s \end{cases}$$

So that would be you saw something like this right so I said this were all stimuli right but you are not talking about supply now is a sake you going to talk about states right so whenever a state occurs let us say some s_1 has occurred here right I increase the eligibility of s_1 I introduce this

new variable called eligibility I will increase eligibility of s_1 and start decaying it right so here some state is to occurred is to occur again here so increase the eligibility of s to increase eligibility of s_2 and I keep decaying it and whenever some reward happens at this point.

I will update every state according to how eligibilities to receive an update all of them by the same rate, no right now I am just increasing everything by one I mean that is a option you could use how much ever you want to do right so this is one and this could be whatever this would be one right so here they show you a hedgehog like picture when a state occurred multiple times so it just keeps going up so what I am going to do is I am going to introduce this new variable called eligibility right.

I am going to say that is one for each state eligibility of the state is equal to $\gamma \lambda$ times the previous eligibility let us say DK part if it is not the current state if it is the current state shut up with all these abilities are 0 for the first time I visit a state it will go to one and then from there we will start decking, so this kind of are presentation is called we call this accumulating places because it just keeps on accumulating the eligibility as and when you revisit the state right there is also a version called replacing places.

Which essentially it is replaced by, so how will the replacing trace thing looks like so state occurs here when it occurs here again so here it will go up you start decaying it will go up the star DK right when it goes up the second time it will go up to 1 okay do not go beyond so if this takes a old trace and replace it with a new trace okay so this is called replacing traces so the first part is the same if it is not researcher you will decay it if it is visited instead of adding one you basically make it equal to one.

Okay this is accumulating this replacing it turns out that let me read this so let me define δT as small δT okay as this is the TD error we saw this way plus a TD error went define capital δT right this one no sir no not ΔT capital Δ oh cool α times hence so every step I am going to change the value of status or at least I m going to compute a change in the value for status right notice that the ΔT is only time indexed it is not state indexed right at time T whatever is the current state based on that I am going to compute the ΔT whatever is the current state.

Whatever the next state and whatever is the reward I will compute the ΔT and use that same ΔT for changing the values of all the states where does it make sense some sense the ΔT is where your current reward enters the picture right so whenever I get a reward here I will use the current reward and I will change all these states which have a non zero eligibility at that point so that is essentially what is happening here but we have to show that all of this plus key components here we'll all get telescoped into one another and we will end up computing $GT \lambda$.

That is the cool part about this right so now you can see how we can have an online versus offline version so online version is as soon as I compute that $\Delta V T$ of S , I play it offline version is will wait till the end of the episode and up all the $\Delta V T$ of s and then I will change it one go right now I can now the backward viewer can have an online and offline version and so that is what the graph in the picture is when the old edition they won't have a picture for the online version I think no they do okay.

Yeah they do it after they describe the online algorithm so the online version yeah here also Oh point six is the minimum not point eight it is the same graph I think point a 10.6 both are where but point six is lower than point eight point eight is lower than point six uh-huh you know why they do not have 0 point 6 in this graph at all this is the same random walk problem so I was really surprised when you guys said it is different but yeah so basically generated it for different steps of steps of λ .

So here they did it insteps of point 2 so you go from zero point two point four point six point eight and so on and so forth but there they have done some arbitrary jumps so there are no point six there anyway so yeah so what would your TD λ algorithm look like it looked like just like your TD algorithm in the TD algorithm if you remember I update the value of the current state I am in right or the right the current state I am inst I do an action I get s_t plus 1 then I compute the delta and I change the value of that state right the state I was in before making the transition.

Here what I do I do the same thing I come to a state I pick an action I perform the action I get the next state I compute the λ then update the eligibility trace try to update the eligibility trace right for all the states and update the eligibility trays for all the states and then I affect the value

function for all the states and then I keep quite right till they come to the end of the trajectory right if I am doing the offline version of this I will complete everything but I am not update the value function I will just compute my change in the value function for that instance I will keep adding it to my I have an accumulator I will just keep adding that at the end of the episode I will take everything and snake one update to the value function co-op bank why the microphones when asked to send how it decay very quickly bring up you might have to so there are multiple things here.

So if you can make sure that your eligibility trace would decay sufficiently before your expected return time to the stake right then you are safe there are whole bunch of different things you have to take into account here and if you choose your λ says that it decays very quickly right if you are if you suspect it is not going to be more important than more than one or two steps into the future then your λ should be very small and so very quickly it will decay to zero or close to zero right.

So there are a whole bunch of other tricks up so another reason you would want to make your λ 0 is because of efficiency in computation we have a very large state space right I do not want to be updating the eligibility trays for every state in every step I take right every step I take I have to keep updating this eligibility date so at some point if you make it zero then that will be efficient right so all I need to know do is have a mechanism for keeping track of the non-zero eligibility traces so if I have a fast way of accessing the non zero visibility trays.

I can only update the values for those states and whenever it comes below a certain threshold I can make it 0 think of the ED as some sort of a probability and update the rod is probabilistic so let us say your ED is point and then you have data Travis of that state probability point and you why would you want to do them responsibility of that state for this current reward is about point and maybe I will only have been in probability point why do you want to have the interpretation why you why you think that is a reasonable thing to have very random moves that though would not that happen over multiple trajectories.

So this trajectory Deckard next trajectory it might not occur I mean so if you say there is a random possibility of it happening or not right if at every tech tree I come right I see this and the

little while later I what happens right should not it get some benefit for that but every German only you would still be considering it as only point nine right so just because you can squeeze the number 200 to one does not mean there is a valid probabilistic interpretation for it right so I am not sure there is a valid probability interpretation for that quantity right.

There might be other ways of into coming up with the stochastic update but just not looking at the decade value of the crease okay, so any other questions on this okay so a couple of observations to make so what happens to my λ return if λ is 0 zero yeah for a conviction becomes td0 only GT one writes everything else will vanish myself make λ 0 I get td0 somewhere also it is fine to call that algorithm td 0 what happens we make λ 1 right so do not look at this if you look at this and it make a zero right then look at this right the first term will go to 0 if λ is one.

How will you be left with is GT was it a full return right so it becomes Monte Carlo do so what is the amazing thing about the TD λ algorithm I can implement this with λ equal to 1 right so I have a incremental way of implementing Monte Carlo right so earlier you are saying Monte Carlo algorithm got to wait till the end because you need to get the whole return but because of this backward view interpretation all I need to do is in to implement this with λ equal to one here right.

And it will give me an incremental way of doing water coral updates in fact it gives me a way of doing Monte Carlo updates even for infinite trajectories so should we do it and what not I said infinite non turban one episodic tasks come on depends on what is the problem with implementing it for longtime infinite trajectories it could go I am sorry yeah I am not going to in finite state spaces and this talked about infinite trajectories infinitely long physically give me too large.

So what ET males can become very little small yeah so then you can think of zeroing it out or something this make sure your γ is not very large man I mean or γ is one your doomed right so in episodic tasks you can implement Monte Carlo problems with γ 1 it is fine but if I am going to do t d λ with λ equal to 1 I have to make sure I have anon 1 γ alice is the trajectory run infinitely then my the Watts will blowup okay, so that simple thing okay good so we know the special

cases we have looked at the special cases so let me so I have to do the equivalence of forward and backward views of $3d \lambda$.

Which is not that in the new book they take in the to get out there is very cumbersome you know man just to page after page after math so very simple dumb math actually not even very clever stuff ok here is my way of getting back to you guys go read the first edition of the book read section 7.4 in fact the task to establish the equivalence between the forward and the backward view of tea $d \lambda$ like I saying that essentially if I do updates like that I said I compute this thing right and it is offline equivalent.

Because this is the offline return this is offline so I should not be looking at the online updates I should look at the offline updates essentially I have to accumulate all the λVG 's right that I will make in a single episode and show that is exactly equivalent to this okay so it is very simple it is just arithmetic just counting right so it is very cumbersome counting and since none of you want to read the book I am forcing you to read the book and I will ask you a question on this whenever the exams come.

So we will have to remember that so you better look at it now right I will have to remember that Oh mommy can we just push the exam syllabus till today the exam date today is it up push no exam date is not getting cushion so I can ask you this nice question no but I will ask you something that asked you to use the math in a more interesting way to make sure you tell people what the question is whether it is a pain even then right.

So we are asking about zeroing out eligibility traces after a while so that is the backward view right so what is the forward view of zeroing out eligible increases after a while it is non-trivial I usually give this only intake home exams not today or not give us an in class exam so I will ask you a different question than class exam but this is non-trivial so what is the forward view of eligible 0 you say dick keep it very simple.

Let us say I am going to be very drastic zero eligibility trace after three steps I do not care how much it has decayed I will cheer oh it out after three steps okay this is the backward view so

essentially if s is not equal to s I will decay this right once twice thrice so I had to keep a count right how many time steps have elapsed since I last encountered the state write a 0 or eligibility trace that is a backward view so what is the equivalent forward view to solve that you need to figure out how we do this arithmetic to establish the forward and backward equivalents.

So we will have to read that up and then figure out okay from there how do you derive this okay so someone can we add that as the homework question for this chapter okay so you must have the wording that I asked it in one of the earlier exams last year in self I said take home ray yeah so you can so is it takes too long to solve in class yeah I do not think anybody got it right also on take good what is the other thing I wanted to mention now ha.

So let us take TD one okay, so we look at two different variations of Monte Carlo right what are the two variations ah first we sit and every we sit right yeah so first to sit in every way CI a-- so off policy eligibility traces a little tricky right so we will come to that not much work has been done on off policy visibility traces so we will have to come to that in fact a very recently there are some work on something which they call to TD λ or exactly $d \lambda$ or something like that because $t d \lambda$ write the forward view talks about the offline updates but almost always when $T \lambda$ is used it is used in the online fashion okay.

Nobody uses an offline fashion so the equivalence of the forward and backward view has been established for the offline case what is online case the forward view is no longer that clean and you can imagine right online in case I keep changing the return as I go along right so I keep changing this δT right this VST that I will use here right that will keep changing every time so it is no longer as clean as this okay, it becomes very confusing and in the forward view so then people have come up with a way of implementing the backward view of lamp theory λ .

So that the forward view also stays neat o there is a established exact correspondence between an online TD λ versus it is I think it is called exactly $d \lambda$ or two TD λ something like that I forget the name of the algorithm now but I will probably give you a little bit of intuition into that ash win when you come to the next class okay yeah, so let us take accumulating traces that is that implement first you said Monte Carlo or every visit Monte Carlo seems to be like every visit

right why do you think that is the case there is a little bit of work you can show that because I keep bumping this up again and again.

So I will compute the return starting from here to here computed in starting from here to here again o I will be using each one of those right so you have to look at section 7.4 from the world book again to figure out how you can write this out right again simple arithmetic you can write it out so that you can show that the accumulating traces this is the implements every visit Monte Carlo now it is the interesting question.

What about replacing traces last visitor first visit it says something called last with Monte Carlo a dies latest visit any doubts any other thoughts what about first visit here is where the section 7.4 becomes your real best friend it implements first visit replacing eligibility trace implements first to set Monte Carlo okay it turns out that when you are showing the equivalence of these two things right so these terms telescope into one another right so therefore the first time you update you will have V_{St} then you will have $\gamma V_{St} + 1$ the next time.

You update you have $V_{St} + 1 + \gamma V_{St} + 2$ and so on so forth right so these things will start cancelling out some of these terms will start cancelling out and because you are the one here those terms do not cancel out right they actually each summation keeps going separately right so you get a visit but because I replace this with one here so whenever something happens again at some of the terms in between get cancelled out.

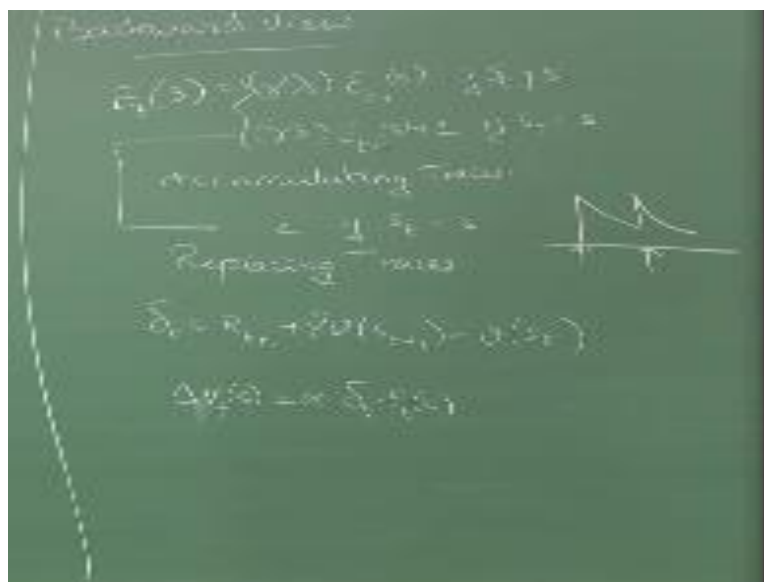
So the second time the trace will never start the summation will never starts it is just a first summation that will keep running till the end that this you understand what I mean only if you look at how the telescoping happens right so if more people had put up your hand this one this the beginning of the class when I asked you if you have been reading up chapter seven I would have done that so I am happy to do the derivation separately for you alone the rest of the class you have to read up from 7.4 then you will understand.

What I mean by the telescoping terms cancelling out so that whatever started from the beginning will continue till then right, so it is actually first visit it is very counter included and but we have

to look at that so then this is not that in the new book sorry a new second edition is not that you have to look at the first edition it is also available online and in fact available from the Model page both the first and the second editions are available but so I am actually when you stop here because the eligibility cases for control is a little painful thing so I will have to let me start it again in the in the next class okay.

So the next class we will finish eligibility traces for control and perhaps a little bit or off policy eligible resistance on that stresses that stresses yeah this place is something that came up recently right so it is something that sits between accumulating traces and replacing traces okay.

(Refer Slide Time: 27:33)



Essentially a Dutch trace I do the update as follows if beta equal to one I get accumulating trace if beta equal to 0 I get replacing trace rapid beta something between 0 and 1 I get Dutch traces. Because some vary from Netherlands objects the paper is called Dutch trace do not think too much about what is it just about it right and so it is a Dutch trace and it turns out that now it gives you an additional parameter attitude β right but then if you can figure out the right β for you to use in many applications turns out to have better empirical performance than either using

replacing or accumulate increased means basically setting $\beta = 1$ or 0 putting some intermediate value of β seems to give you a better performance okay.

So there are some hypotheses and reasons as to why this should be the case but largely the advantage is empirical Joanna have been established empirical wait so you really have a heretical reason as to believe one was better before we leave this one important thing I wanted to point out to you γ is a parameter of the problem why changing γ changes the answer changing γ changes the right answer λ is a parameter of the solution method changing λ does not change the right answer.

See the optimal policy is still the same the optimal value function is still the same λ only lets me have a different technique for measuring or estimating the value function okay, whether I take the expected value of $Z^T \lambda$ or whether I take the expected value of $G^T \lambda$ or $G^T \lambda$ it all converges to the same value function right but if γ changes the solution changes right I can plug in λ point three here and take the expected value truly it should be the same value as λ point six a λ point nine and so on.

So forth so what I showed you in those curves are because of finite sample sizes right if I actually take the two expectations they will all be the same and I showed you it means λ equal point eight this better point n breeze because we are working with a finite set of samples so if I take the two expectations the answer will be the same so this something you have to keep in mind so λ is a parameter of the learning process well γ is a parameter of the problem right.

And so do not get confused by looking at the backward back ward view λ always comes together right so I do λ that is because I want to make sure that when you do the telescoping things get cancelled outright I have $\alpha V^T \lambda + T V^T \lambda + 1$ here but in the next step I will have a minus $V^T \lambda + 1$ I want the 22 not cancel out but combine in the appropriate form so that two total when I sum up everything I will get $1 + 1$ return right.

So for that I need the gamma herein fact in the original version of eligibility traces of original TD λ algorithm that rich has been proposed there is no gamma they are just a λ and then later on they

put the γ n so that the forward and backward views could be merged when you are working with limited samples yeah because of the process of looking at I mean the α also affects a solution as we looked at right.

So the α is another parameter of the learning process right so this is a SEP size so that is also a fixed solution so likewise any parameter of the learning algorithm is going to affect the solution because it is going to influence the trajectory are going to take through the solution space so if you have finite sample sizes here λ will affect.

IIT Madras Production

Funded by

Department of Higher Education

Ministry of Human Resource Development

Government of India

www.nptel.ac.in

Copyrights reserved