Lecture 4

→ No MDP given, although the environment is an MDP like system.

→ We will discuss model free prediction →
i.e estimate the value function of an unknown MDP.

Methods

→ Monte Carlo Learning → Learns directly from the episodes of experience.
↳ we take sample returns from the episodes and average over them
↳ It only works for episodic experience. i.e we need to terminate to be able to work
↳ Value = mean return over many episodes

Goal: Learn V_π from episodes of experience under policy π.
↳ We will look at the total reward in each time step (G_t) and then take an average of it to calculate $V_\pi(s)$ i.e $E(G_t)$
↳ Use emperical mean returns instead of expected return.

→ 1

→ First visit monte carlo policy Estimation →

↳ to evaluate state s

↳ Given the series of episodes, we look at the ~~t~~ number of time ~~it~~ we visited the state s the first time

↳ $N(s) \rightarrow N(s) + 1$

↳ Increment the total return as we visit it $S(s) \leftarrow S(s) + G_t$

↳ Value is estimated by mean return

↳ $V(s) = S(s) / N(s)$

↳ By law of large numbers

$V(s) \rightarrow V_\pi(s)$ as $N(s) \rightarrow \infty$

→ Every visit mate carlo policy Estimator →

↳ Same as the above except that the visit to the episode need not only be the first to be included in the mean

If we visit a state multiple times in an episode, it will be include in the $N(s)$ and $S(s)$ ~~those many times~~ every time

→ Incremental mean → Update the mean after every time step and not necessarily after the run is over.

$$\mu_k = \frac{1}{k} \sum_{j=1}^{k} x_j$$

$$= \frac{1}{k} \left(x_k + \sum_{j=1}^{k-1} x_j \right)$$

$$= \frac{1}{k}\left(x_k + (k-1)\mu_{k-1}\right)$$

$$= \mu_{k-1} + \frac{1}{k}\left(x_k - \mu_{k-1}\right)$$

$$\mu_k = \mu_{k-1} + \frac{1}{c}\left(\delta\mu_k\right)$$

μ_{k-1} → old mean

$\frac{1}{c}$ → constant

$\delta\mu_k$ → Error term

→ Incremental Monte Carlo Updates → ~~Update~~ Update $V(s)$ incrementally after episode $S_1, A_1, R_2 \cdots S_T$

↳ For each state S_t with return G_t

$$N(s_t) \leftarrow N(s_t) + 1$$

$$V(s_t) \leftarrow V(s_t) + \frac{1}{N(s_t)}\left(G_t - V(s_t)\right)$$

If it is a non ~~stationdl~~ stationary problem i.e the MDP changes with time steps, we need a method which gives more weight to latest observation in the mean.

$$V(s_t) \leftarrow V(s_t) + \alpha\left(G_t - V(s_t)\right)$$

$$0 < \alpha < 1$$

→ Temporal Difference Learning

↳ learn directly from actual experience
↳ TD is model-free
↳ learns from incomplete episodes
We can learn while the episode is still running and doesnot require the episode to end to update the value functions. It utilises "bootstrapping", we update the value function using the value functions of the other states.

Goal: learn V_π online from experience under policy π.
Online implies that we will update our value function while the episode is still running.

↳ we update our value function towards the estimated return
$R_{t+1} + \gamma V(S_{t+1})$

In TD(0) we take one step from state S to reach state S' and get a return R. The estimate value of S is updated using the R and the estimated value of S' i.e $V(S')$ using the bellman equation.

Then an ~~mean~~ incremental mean is performed on the V(s) using the new bootstrapped value of V(s) from R and V(s')

$$V(s_t) = R_{t+1} + \gamma(S_{t+1})$$

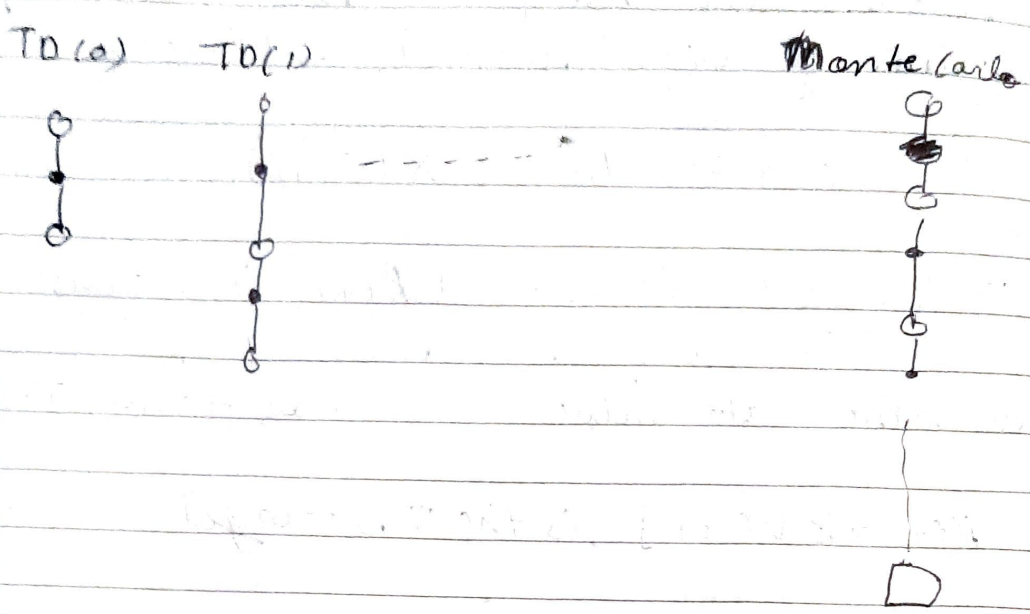
$$\Rightarrow \underbrace{V(s_t)}_{\text{New Value}} \leftarrow \underbrace{V(s_t)}_{\text{old value}} + \alpha \underbrace{[R_{t+1} + \gamma V(s_{t+1}) - V(s_t)]}_{\text{Error term (TD error)}}$$

$R_{t+1} + \gamma V(s_{t+1})$ is the TD target

→ Bias / Variance trade off →

↳ TD return $R_{t+1} + \gamma V(S_{t+1})$ is biased estimate of $V_\pi(S_{t+1})$, it would be unbiased if $V_\pi(S_{t+1})$ is the actual value of the value function. But the variance is low because the ~~random ness~~ randomness is only due to the reward after the State S. # Compared to monte carlo which considers all the rewards till the episode terminates and therefore will have very different values everytime we sample it.

→ n steps Prediction →



$n=1$ (TD) $G_t^{(1)} = R_{t+1} + \gamma V(S_{t+1})$

$n=2$ $G_t^{(2)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 V(S_{t+2})$

⋮

$n=\infty$ (MC) $G_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} \cdots \gamma^{T-1} R_T$

$$G_t^n = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V(S_{t+n})$$

nstep temporal difference Learning

$$V(S_t) \leftarrow V(S_t) + \alpha \left(G_t^{(n)} - V(S_t) \right)$$

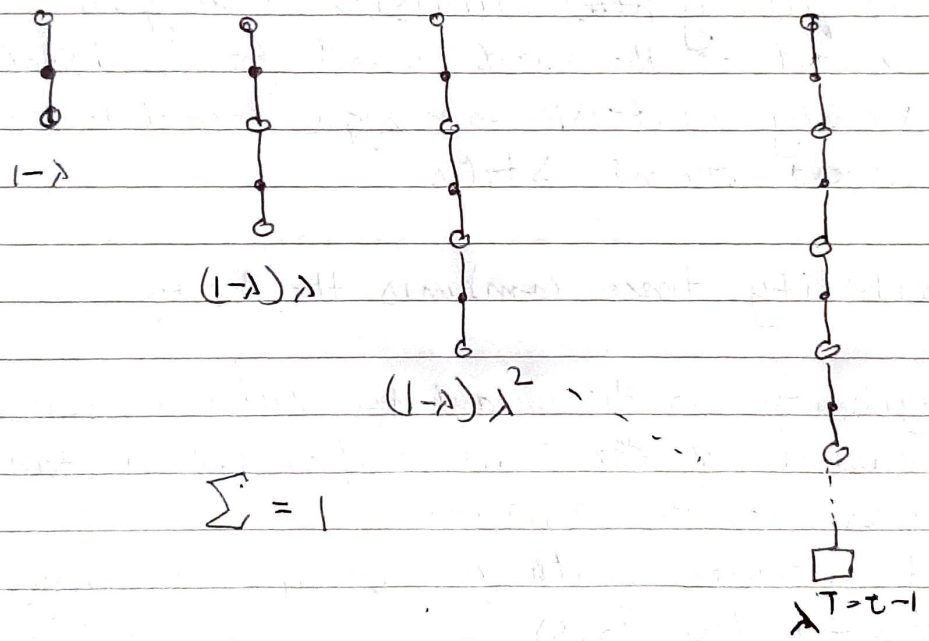
→ Average n step Returns

Eg we can agvage 2step and 4 step returns

$$\Rightarrow \frac{1}{2} G_t^{(2)} + \frac{1}{2} G_t^{(4)}$$

can we do it for all n

→ TD (λ)



$$G_t^{\lambda} = (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} G_t^{(n)}$$

Forward view

$$V(s_t) \leftarrow V(s_t) + \alpha \left(G_t^{\lambda} - V(s_t) \right)$$

→ Forward-view TD(λ)

Like MC it can only be computed from complete episode

→ Backward-View TD(λ)

What caused a reward ?
or which state was most responsible
→ Frequency heuristic → assigns credit to the most frequent state
→ Recency heuristic → assigns credit to the most recent states

Eligibility trace combines the both

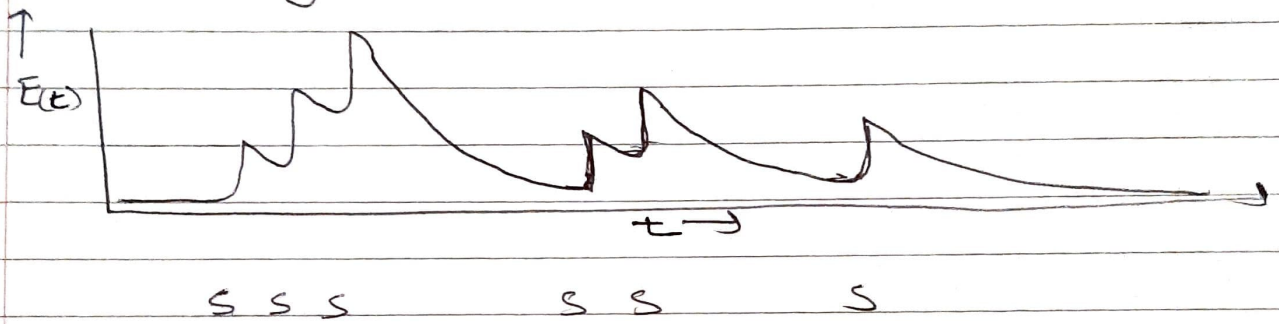
Eligibility is the quantity which is a measure of how responsible a state was for the reward
It increases with frequency and recency
→ denoted by $E_0(s)$

default value = 0 ; $E_0(s) = 0$

$$E_t(s) = \gamma \lambda E_{t-1}(s) + 1(s_t = s)$$

when ever we encounter the state s we increase its $\gamma \lambda E_{t-1}$ by 1 and make it the new $E_t(s)$

Otherwise $E_t(s)$ is updated by multiplying with λ and γ both are less than 1. Therefore as the time step increases, the eligibility of the state decrease until it occurs again when its value is increased by 1 after reducing it by γ and λ time.

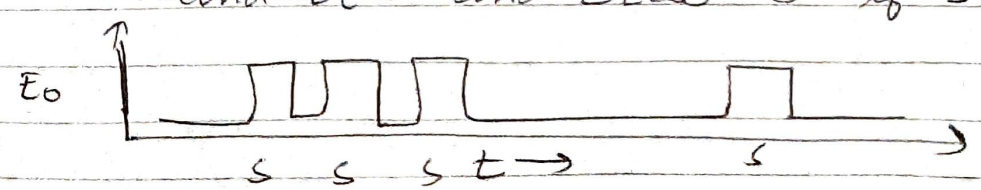


→ Keep an eligibility trace for every state s.
→ Update value V(s) for every state s in proportion to the TD error and the eligibility trace $E_t(s)$

$$\delta_t = R_{t+1} + \gamma V(S_{t+1}) - V(S_t)$$
$$V(s) \leftarrow V(s) + \alpha \delta_t E_t(s)$$

when $\lambda = 0$, it reduces to TD(0)

$$E_t(s) = 1(S_t = s)$$
~~and E_t~~ = and $E_t(s) = 0$ if $S_t \neq s$



$$V(s) \leftarrow V(s) + \alpha E_t(s)\, \delta t$$

as $\lambda = 0$.

$$\Rightarrow \cancel{V(s)} \; V(s_t) \leftarrow V(s_t) + \alpha \delta t$$

If $\lambda = 1$; we get same behaviour as MC