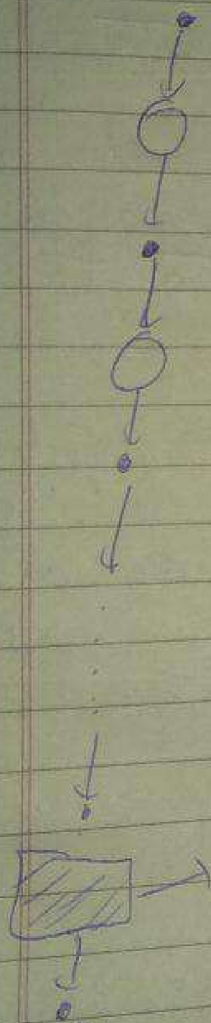
Date

DELTA Pg No.

Q8 Yes, as when picking a' for our policy we will get same result as when taking argmax. Hence, was Q-learning will become on-policy.

Q2

Q2



Terminal state

$Q(S_{terminal}, a) = 0$

Q5 TD depends on bootstrapping. ~~we~~

If we already have statevalues for a model, & then make a minor change to ~~[illegible]~~ some ~~[illegible]~~ states of the model, ~~[illegible]~~ when ~~[illegible]~~ finding q values for this modified model, if we initialise its q values with the q values of initial model, we will see significant gains compared to MC, since TD within the episode ~~to~~ will use these correct q-values to estimate the ~~correct~~ value, giving it ~~a very significant increase~~ in accuracy modified state values. ~~[illegible]~~ allowing

~~giving very good est~~

~~to let's consider the~~

Q

Q ~~We went~~ all the way left & stopped.

We terminated on the left side. Initially all $v_i$'s are same hence only change due to updates is due to the reward which is

Q6 i We terminated on the left side
For all non-terminal states $v_i$'s are same
therefore & rewards are also 0 on all non-edge states, hence no update have any effect on these states.

It changed by $\alpha(v[s'] + r - v[s])$

$= 0.1(0+0-0.5)$

$= -0.05$

ii ~~The algorithm~~ If large number of iterations are used then smaller alpha is more beneficial, as it ~~lowers variance due~~ ~~to~~ it lowers effect of recent events and accounts for entire experience better.
~~However this~~ ~~be~~ Thus there is no perfect alpha that will always work.

Q1 Initialize a $count(s, a) = 0 \ \forall s \in S, a \in A(s)$

& change update of $Q(S_t, A_t)$ to

$$Q(S_t, A_t) = Q(S_t, A_t) + \frac{(G - Q(S_t, A_t))}{count(S_t, A_t)}$$

$count(S_t, A_t) \mathrel{+}= 1$

Q3 $$Q(s,a) = \frac{\sum_{t \in \tau(s,a)} \rho_{t:T(t)-1} G_t}{\sum_{t \in} \rho_{t \in T(t)}}$$

In MC, since the episode is already simulated $\therefore G_t$ is effectively already conditioned on $A_t = a$, since we know that $A_t = a$ was the action taken.