MDPs and TD

Ajinkya Ambatwar
EE16B104
Dept. Of Electrical Engineering

March 23, 2019

1. The MDP for the given problem is as follows-

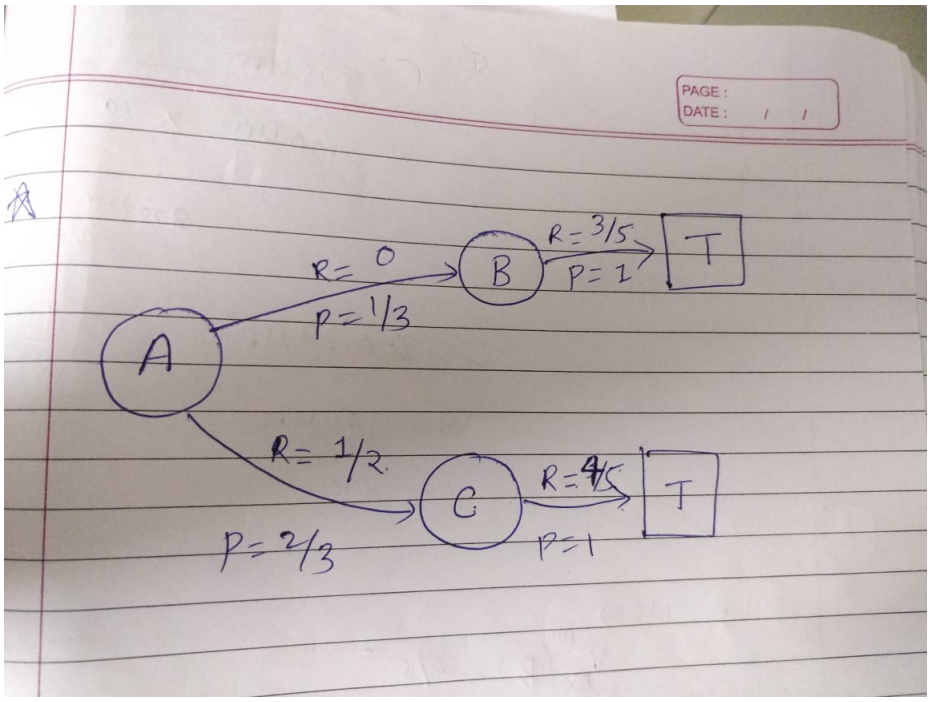


Figure 1: MDP for given table

So from the definition of batch MC the value function is as follows-

$$V_{MC}(A) = 1 * \frac{2}{3} + 2 * \frac{1}{3} = \frac{4}{3}$$
$$V_{MC}(B) = 4/5$$
$$V_{MC}(C) = 3/5$$

For the batch-TD(0) update

$$V_{TD}(B) = (3/5 + 0) = 0.6$$
$$V_{TD}(C) = (4/5 + 0) = 0.8$$
$$V_{TD}(A) = \frac{1}{3}(0 + 0.6) + \frac{2}{3}(0.5 + 0.8) = 16/15$$

The MSE values

$$MSE(TD) = \frac{1}{3}(\frac{16}{15} - \frac{4}{3})^2 + \frac{1}{5}(0.6 - 0.6)^2 + \frac{1}{5}(0.8 - 0.8)^2 = 0.023$$
$$MSE(MC) = \frac{1}{3}(\frac{4}{3} - \frac{4}{3})^2 + \frac{1}{5}(0.6 - 0.6)^2 + \frac{1}{5}(0.8 - 0.8)^2 = 0$$

For training data MC is better as it finds a solution that gives minimum error on training data. For the model, the TD method gives a better value as it finds the maximum likelihood estimate of V^{π}which also has the highest probabilty of generating data. Hence the error on future data is minimum for TD estimate given the process in Markov and hence this estimate is also called as certainty-equivalence estimate.

2. The time index will be taken t mod τ.
Hence the monte carlo return-

$$G_t = R_{t+\tau_1} + \gamma^{\tau_1} R_{(t+\tau_1+\tau_2)} ....$$

Similarly the TD(0) update can be given as -

$$V^{\pi}(S_t) = V^{\pi}(S_t') + \alpha[R(S, A, S', \tau) + \gamma^{\tau} V^{\pi}(S_{t+\tau}) - V^{\pi}(S_t)]$$

3. For given 3-step eligibility trace truncation, the $\lambda - \grave{return}$ is defined as

$$G^{\lambda}_{t:t+3} = \hat{v}(S_t, w_{t-1}) + \sum_{i=t}^{t+2} (\gamma\lambda)^{i-t} \delta_i^{'}$$

where $\delta_i^{'} = R_{t+1} + \gamma \hat{v}(S_{t+1}, w_t) - \hat{v}(S_t, w_{t-1})$
For general n the expression is as follows-

$$G^{\lambda}_{t:t+n} = \hat{v}(S_t, w_{t-1}) + \sum_{i=t}^{t+n-1} (\gamma\lambda)^{i-t} \delta_i^{'}$$

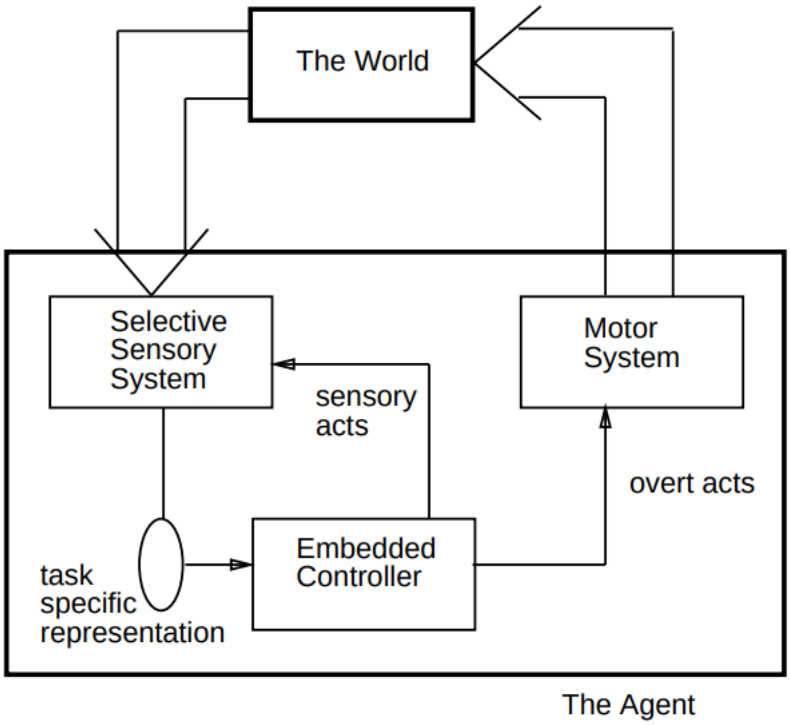


Figure 2: Active Perception model

4. One of the major reason for Non-Markov model is due to Active Perception. Active perception refers to the idea that an intelligent agent should actively control its sensors in order to sense and represent only the information that is relevant to its immediate ongoing activity. Tasks that involve active perception lead naturally to non-markov decision problems since improper control of the sensors leads to internal representation that fail to encode into relevant to decision making.
   In such case, a state in the internal representation maps to two or more states in the external markov model of the system. In such cases, the markov model of the system based learning methods can not find a reliable estimate for $V^{\pi}(S)$ for such perpetually aliased states.
   This leads to localization errors in the policy function. Use of TD methods in such case spreads the estimation erros throughout the state space, thus infecting even policy action for non-aliased states.
   [1]

[1] Ref: Reinforcement Learning in Non-Markov Environments(Whitehead and co.,1992)

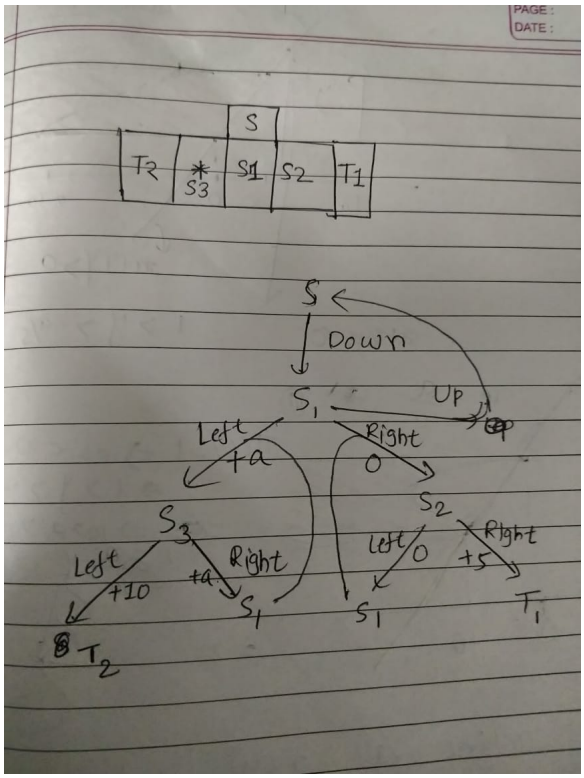


Figure 3: Gridworld and Tree Diagram

5. The grid world is as follows-

π	S_1	S_2	S_3	V_1	V_2	V_3
1	Left	-	Left	$a + 10\gamma$	0	10
2	Left	-	Right	$a + a\gamma + a\gamma^2 + ...$	0	$a + a\gamma + a\gamma^2...$
3	Right	Left	-	0	0	0
4	Right	Right	-	$0 + 5\gamma$	5γ	0
5	Up	-	-	0	0	0

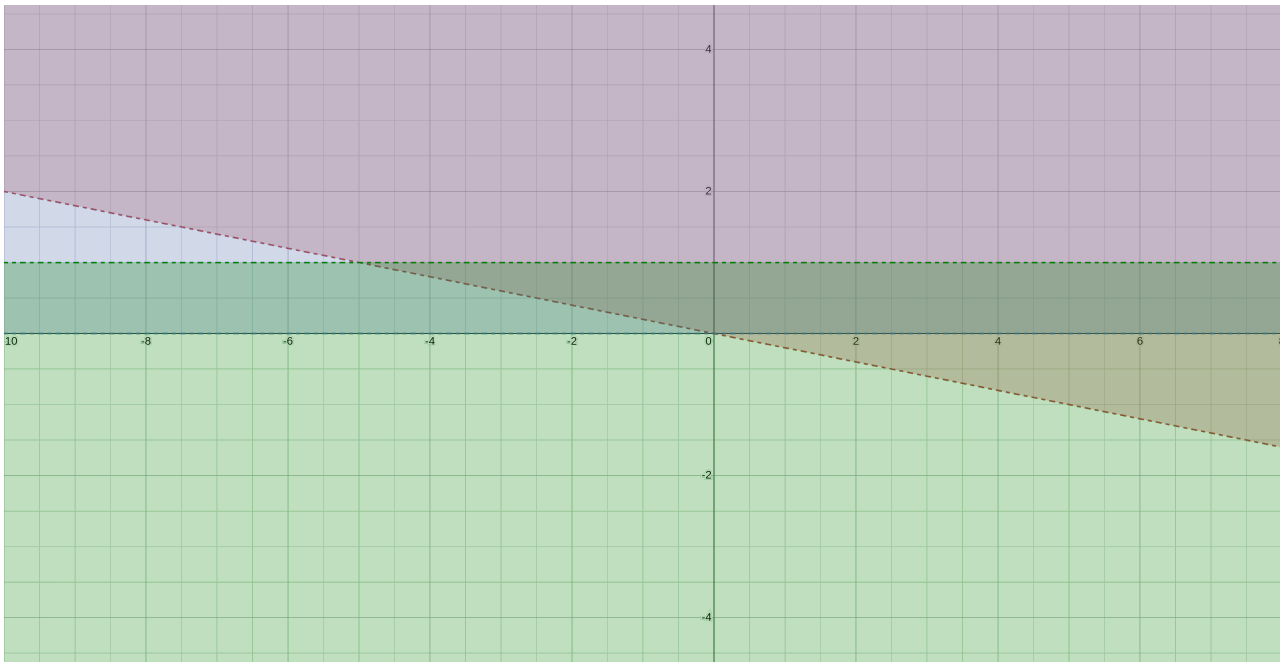
V	$\pi(1)$	$\pi(2)$	$\pi(3)$	$\pi(4)$	$\pi(5)$
V_1	$a + 10\gamma$	$\frac{a}{1-\gamma}$	0	5γ	0
V_2	0	0	0	5γ	0
V_3	10	$\frac{a}{1-\gamma}$	0	0	0

For calculation of $\pi(S)$ for various values of a and γ, let's take an example for S_1

(a) $\pi(S_1) = \pi(1)$

$$a + 10\gamma > \frac{a}{1-\gamma}$$
$$a + 10\gamma > 5\gamma$$

The given region can be found out as



The solution for this set of inequalities is the region of intersection shown in the graph. The solution is

$$\pi(S_1) = \pi(1) \tag{1}$$
$$\tag{2}$$

SImilarly we can solve for all other policies for all other states. The required condition is that

$$\pi(S_i) = \pi(j)$$

when V_i corresponding to $\pi(j)$ is highest among values correspondings all the $\pi(j)$s.

6. In this case we will define a new constant system dynamics $p_{new}(S'|S,a)$ which is not equal to zero for all {S,A,S'} pair. The new formulation for V^{π}in that case will b

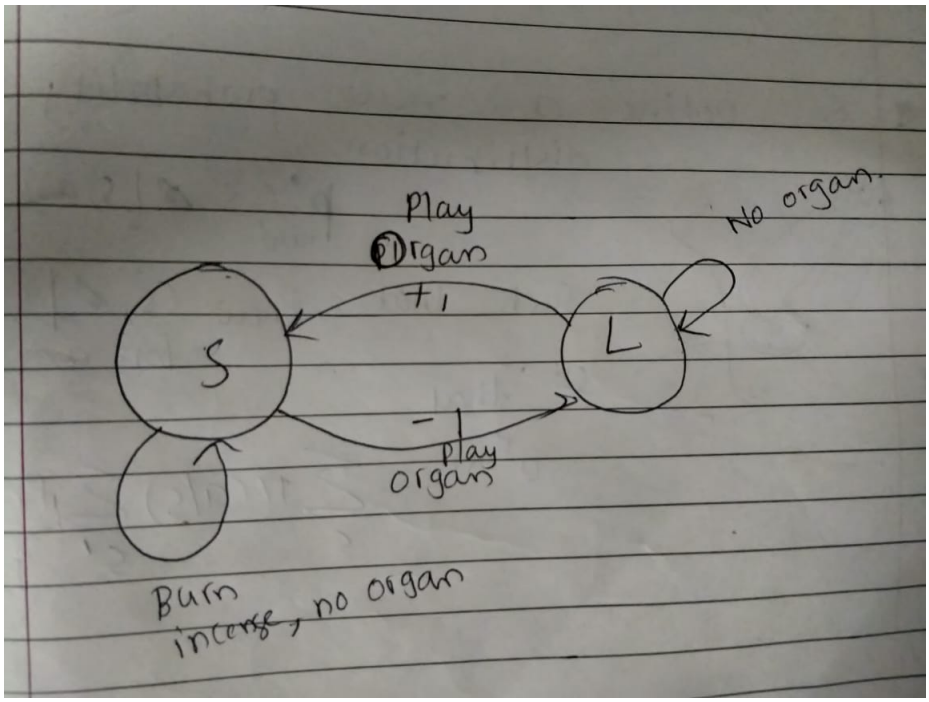
$$V^{\pi}(s) = \sum_{s} \pi(a|s) \sum_{s'} \{ \frac{p(s'|s,a)}{p_{new}(s'|s,a)} [R(s,a,s') + \gamma v^{\pi}(s')] p_{new}(s'|s,a) \}$$

So here instead of p the expectation is taken wrt p_{new} and is of $\frac{p}{p_{new}}(R + \gamma V^{\pi}(s'))$
In this way we can get rid of the changing system dynamics and make it markov. The only condition on p_{new} is that is should be non-zero for all possible s,a,s' pairs and hence the MDP should have a non-zero probable transitions from every state to every other state.

7. The on-policy version on Q-learning can be represented as-

   (a) Choose A from S using policy derived from Q
   (b) Take action A, observe R,S'
   (c) Choose A' from S' using policy derived from Q
   (d) $Q(S,A) = Q(S,A) + \alpha[R + \gamma Q(S',A') - Q(S,A)]$
   (e) $S = S', A = A'$
   This way we take action by policy π and also be change π toward greediness wrt q_{π}
   This will converge if all the state-action pairs are met often.

8. The MDP looks like this



The state set is {S,L} (representing Silence and Loudness respectively), Action set: {O $\bigwedge$ I, O $\bigwedge$ $\neg$I, $\neg$O $\bigwedge$ I, $\neg$O $\bigwedge$ $\neg$I}, where O corresponds to playing the organ, and I corresponds to burning incense).
Using policy iteration as the initial state is Laughing and initial action is (incense, no organ), there will be no change in the state as can be seen from the MDP.