Course 2, Module 5 Planning, Learning & Acting CMPUT 397 Fall 2019

Finishing the demo

2. An agent is in a 4-state MDP, $S = \{1, 2, 3, 4\}$, where each state has two actions $A = \{1, 2\}$. Assume the agent saw the following trajectory,

> $S_0 = 1, A$ $S_1 = 1, A$ $S_2 = 2, A$ $S_3 = 2, A$ $S_4 = 3, A$ $S_{5} = 4$

and uses Tabular Dyna-Q with 5 planning steps for each interaction with the environment.

Once the agent sees S_5 , how many Q-learning updates has it done with real experience? (a) How many updates has it done with **simulated experience**?

$$A_0 = 2, R_1 = -1,$$

 $A_1 = 1, R_2 = 1,$
 $A_2 = 2, R_3 = -1,$
 $A_3 = 1, R_4 = 1,$
 $A_4 = 1, R_5 = 100,$

 $S_0 = 1, A$ $S_1 = 1, A$ $S_2 = 2, A$ $S_3 = 2, A$ $S_4 = 3, A$ $S_{5} = 4$

Which of the following are possible (or not possible) simulated transitions $\{S, A, R, S'\}$ (b) given the above observed trajectory with a deterministic model and random search control?

i.
$$\{S = 1, A = 1, R = 1, S' = 2\}$$

ii. $\{S = 2, A = 1, R = -1, S' = 3\}$
iii. $\{S = 2, A = 2, R = -1, S' = 2\}$
iv. $\{S = 1, A = 2, R = -1, S' = 1\}$
v. $\{S = 3, A = 1, R = 100, S' = 5\}$

$$A_0 = 2, R_1 = -1,$$

 $A_1 = 1, R_2 = 1,$
 $A_2 = 2, R_3 = -1,$
 $A_3 = 1, R_4 = 1,$
 $A_4 = 1, R_5 = 100,$



3. Modify the Tabular Dyna-Q algorithm so that it uses Expected Sarsa instead of Q-learning. Assume that the target policy is ϵ -greedy. What should we call this algorithm?

Tabular Dyna-Q

Initialize Q(s, a) and Model(s, a) for all $s \in S$ and $a \in A(s)$ Loop forever: (a) $S \leftarrow \text{current}$ (nonterminal) state (b) $A \leftarrow \varepsilon$ -greedy(S, Q)(c) Take action A; observe resultant reward, R, and state, S'(d) $Q(S, A) \leftarrow Q(S, A) + \alpha [R + \gamma \max_a Q(S', a) - Q(S, A)]$ (e) $Model(S, A) \leftarrow R, S'$ (assuming deterministic environment) (f) Loop repeat n times: $S \leftarrow$ random previously observed state $A \leftarrow$ random action previously taken in S $R, S' \leftarrow Model(S, A)$ $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma \max_{a} Q(S', a) - Q(S, A) \right]$



4. Confident and the state of t transitions the deterministice are surections or changerifthe actions. "Maigh the state switches if the action switch, comespending neighboring states are randomly distributed. movement is blocked by an obstacle or the edge of the maze, in which case the agent remains where it is. Reward is zero on all transitions, except those into the goal state, on which it is +1.5 P(tRr|SeachingAthesgoal)state (G), the agent rP(tRnsStorthesforthesThe main part of Figure 8.2 shows average learning curves from an experiment in W.P. 0.5 Whith Dyng, Qlagentayyere applied to the mate (Rask In the part of the second se the step-size parameter was $a = W \cdot P$, and the exploration parameter was $\varepsilon = 10.1 W h eh^5$ selecting greedily among actions, ties were broken randomly. The agents varied in the (a) How might of planarnet the ps. war the operformed: pt indate by the part of the properties are stimated. Forthex ampler affrate pift taken where the agent interactive adhet be goability and episoide, laved aged over adds? If your on stational and the instant of the second state of the state of the state of the probabilities by confirting: was held constant 0 grass algorithms. Because of this, the first episode was exactly the same (about 1700 steps) for all values of n, and its data are not shown in Modifyighter the first episode it performance in the state of the st more rapidly for larger values. Recall that the n = 0 agent is a nonplanning agent, using



(b)

Tabular Dyna-Q

Initialize Q(s, a) and Model(s, a) for all $s \in S$ and $a \in A(s)$ Loop forever:

(a) $S \leftarrow \text{current}$ (nonterminal) state (b) $A \leftarrow \varepsilon$ -greedy(S, Q)

(c) Take action A; observe resultant reward, R, and state, S'(d) $Q(S, A) \leftarrow Q(S, A) + \alpha | R + \gamma \max_a Q(S', a) - Q(S, A) |$ (e) $Model(S, A) \leftarrow R, S'$ (assuming deterministic environment)

(f) Loop repeat n times:

 $S \leftarrow$ random previously observed state $A \leftarrow$ random action previously taken in S $R, S' \leftarrow Model(S, A)$ $Q(S, A) \leftarrow Q(S, A) + \alpha \left[R + \gamma \max_{a} Q(S', a) - Q(S, A) \right]$

Modifyigheetakelartheynesteepilseritherteenhandlenteisvelderwithateeshastichewarden more rapidly for larger values. Recall that the n = 0 agent is a nonplanning agent, using