CMPUT 365: Introduction to Reinforcement Learning, Winter 2023
Worksheet #8: Planning, Learning, and Acting

Manuscript version: *#80c831* - 2023-04-08 23:26:26-06:00

Question 1. An agent observes the following two episodes from an MDP,

$S_0 = 0, A_0 = 1, R_1 = 1, S_1 = 1, A_1 = 1, R_2 = 1, S_2 =$ `<terminal>`,
$S_0 = 0, A_0 = 0, R_1 = 0, S_1 = 0, A_1 = 1, R_2 = 1, S_2 = 1, A_2 = 1, R_3 = 1, S_3 =$ `<terminal>`.

and updates its *deterministic* model accordingly. What would the model output for the following queries:

1. Model$(S = 0, A = 0)$
2. Model$(S = 0, A = 1)$
3. Model$(S = 1, A = 0)$
4. Model$(S = 1, A = 1)$

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

$$
\begin{aligned}
&S_0 = 1, A_0 = 2, R_1 = -1,\\
&S_1 = 1, A_1 = 1, R_2 = 1,\\
&S_2 = 2, A_2 = 2, R_3 = -1,\\
&S_3 = 2, A_3 = 1, R_4 = 1,\\
&S_4 = 3, A_4 = 1, R_5 = 100,\\
&S_5 = 4.
\end{aligned}
$$

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

1. Once the agent sees S_5, how many Q-learning updates has it done with **real experience**? How many updates has it done with **simulated experience**?
2. Which of the following are possible (or not possible) simulated transition tuples (S, A, R, S') given the above observed trajectory with a deterministic model and random search control?
   (a) $(S = 1, A = 1, R = 1, S' = 2)$
   (b) $(S = 2, A = 1, R = -1, S' = 3)$
   (c) $(S = 2, A = 2, R = -1, S' = 2)$
   (d) $(S = 1, A = 2, R = -1, S' = 1)$
   (e) $(S = 3, A = 1, R = 100, S' = 5)$

Question 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 $s \in \mathcal{S}$ and $a \in \mathcal{A}(s)$
Loop forever:
- (a) $S \leftarrow$ current (nonterminal) state
- (b) $A \leftarrow \varepsilon\text{-greedy}(S,Q)$
- (c) Take action A; observe resultant reward, R, and state, S'
- (d) $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$
- (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
  - $R, S' \leftarrow Model(S,A)$
  - $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$

Question 4. Consider an MDP with the state set $\mathcal{S} = \{1, 2\}$, action set $\mathcal{A} = \{\texttt{stay}, \texttt{switch}\}$, and the reward set $\mathcal{R} = \{0, 1\}$. Assume the state transitions to be deterministic: the state stays the same if the action is "`stay`" and the state switches to the other state if the action is "`switch`". The rewards are random and are described by the distribution

$$\mathbb{P}(R_{t+1} = r | S_t = 1, A_t = \texttt{stay}) = \begin{cases} 0.4 & \text{if } r = 0 \\ 0.6 & \text{if } r = 1 \end{cases},$$

$$\mathbb{P}(R_{t+1} = r | S_t = 1, A_t = \texttt{switch}) = \begin{cases} 0.5 & \text{if } r = 0 \\ 0.5 & \text{if } r = 1 \end{cases},$$

$$\mathbb{P}(R_{t+1} = r | S_t = 2, A_t = \texttt{stay}) = \begin{cases} 0.6 & \text{if } r = 0 \\ 0.4 & \text{if } r = 1 \end{cases},$$

$$\mathbb{P}(R_{t+1} = r | S_t = 2, A_t = \texttt{switch}) = \begin{cases} 0.5 & \text{if } r = 0 \\ 0.5 & \text{if } r = 1 \end{cases}.$$

1. How might you learn the reward model?

   (Hint: Think about how you can estimate probabilities. For example, what if you were to estimate the probability of a coin landing on heads? If you observed 10 coin flips with 8 heads and 2 tails, then you could estimate the probabilities by counting: $p(\texttt{heads}) \approx \frac{8}{10} = 0.8$ and $p(\texttt{tails}) \approx \frac{2}{10} = 0.2$.)

2. Modify the tabular Dyna-Q algorithm to handle this MDP with stochastic rewards.

Question 5. Challenge Question: Consider an MDP with three states (and let $\mathcal{S} = \{1, 2, 3\}$), where each state has two possible actions (and let $\mathcal{A} = \{1, 2\}$). Set the discount factor $\gamma = 0.5$. Suppose the estimates of $Q(S, A)$ are initialized to 0 and you observed the following episode according to an unknown behaviour policy:

$$S_0 = 1, A_0 = 1, R_1 = -7, S_1 = 2, A_1 = 2, R_2 = 5, S_2 = 1, A_2 = 1, R_3 = 10, S_3 = \texttt{<terminal>}.$$

where `<terminal>` represents the terminal state. Then answer the following questions:

1. Suppose you used Q-learning (with a stepsize of $\alpha = 0.1$) and the above trajectory to estimate $Q(S, A)$. Then, what would your new estimates be for $Q(1, 1)$?
2. What is one possible model for this environment? Is the model stochastic or deterministic?
3. Suppose in the planning loop, after search control, we would like to update $Q(1, 1)$ with Q-planning. What are the possible outputs of Model$(1, 1)$?
4. Suppose the result of the query Model$(1, 1)$ happens to be $(10, \texttt{<terminal>})$. Then using this simulated experience, compute $Q(1, 1)$ after one Q-planning update. Use the estimates of $Q(S, A)$ from before.

Question 6. (*Exercise 8.2 S&B*) Why did the Dyna agent with exploration bonus, i.e. Dyna-Q+, perform better both in the first phase as well as in the second phase of the blocking experiment (as shown in Figure 8.4 and reproduced below)?

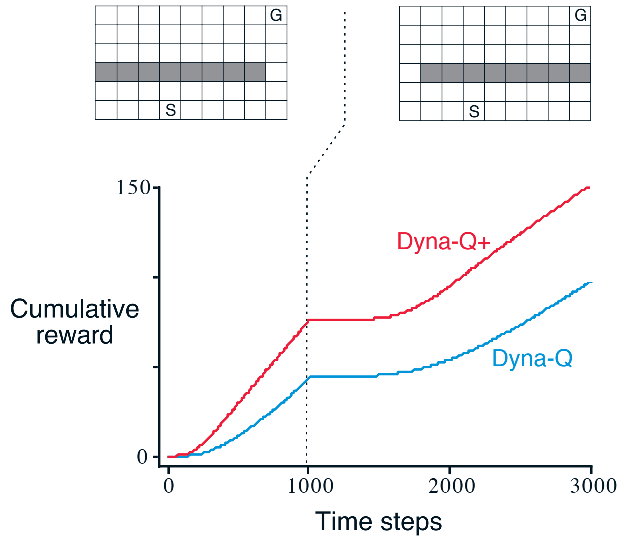

Figure 8.4: Average performance of Dyna agents on a blocking task. The left environment was used for the first 1000 steps, the right environment for the rest. Dyna-Q+ is Dyna-Q with an exploration bonus that encourages exploration. ■

Question 7. (*Exercise 8.3 S&B*) **Challenge Question.** Careful inspection of Figure 8.5 from the textbook (also reproduced below) reveals that the difference between Dyna-Q+ and Dyna-Q narrowed slightly over the first part of the experiment. What is the reason for this?

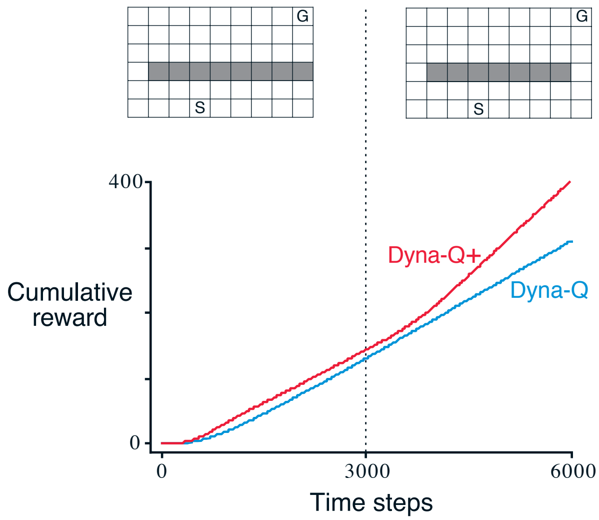

Figure 8.5: Average performance of Dyna agents on a shortcut task. The left environment was used for the first 3000 steps, the right environment for the rest.

■