Planning, Learning and Acting

Rupam Mahmood

March 11, 2020

Dyna-Q+

1. Adds a bonus $\kappa\sqrt{\tau(s,a)}$ to reward in planning

$\tau(s,a)$ **denotes the number of time steps** (s,a) **has not been tried**

2. Actions that have not been tried from a previously visited state are allowed to be considered in planning

Where would you put these steps in Dyna-Q to get Dyna-Q+?

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
$R, S' \leftarrow Model(S,A)$
$Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$

Dyna-Q+: calculating visitation counts

Consider an MDP with one actions (L) and two states (x, y) with the following episode

S_0	A_0	S_1	A_1	S_2	A_2	S_3	A_3	S_4	A_4
y	**L**	**x**	**L**	**x**	**L**	**y**	**L**	**x**	**L**

Calculate $\tau(s, a)$ for all state-action pairs at each step

Worksheet question

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

$$S_0 = 0, A_0 = 1, R_1 = 1, S_1 = 1, A_1 = 1, R_2 = 1$$

$$S_0 = 0, A_0 = 0, R_1 = 0, S_1 = 0, A_1 = 1, R_2 = 1, S_2 = 1, A_2 = 1, R_3 = 1$$

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

(a) $\text{Model}(S = 0, A = 0)$:

(b) $\text{Model}(S = 0, A = 1)$:

(c) $\text{Model}(S = 1, A = 0)$:

(d) $\text{Model}(S = 1, A = 1)$:

Worksheet 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.

(a) 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**?

(b) Which of the following are possible (or not possible) simulated transitions $\{S, A, R, S'\}$ 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. $\{S = 2, A = 2, R = -1, S' = 2\}$

Worksheet 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
  - $R, S' \leftarrow Model(S,A)$
  - $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$

Worksheet question

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

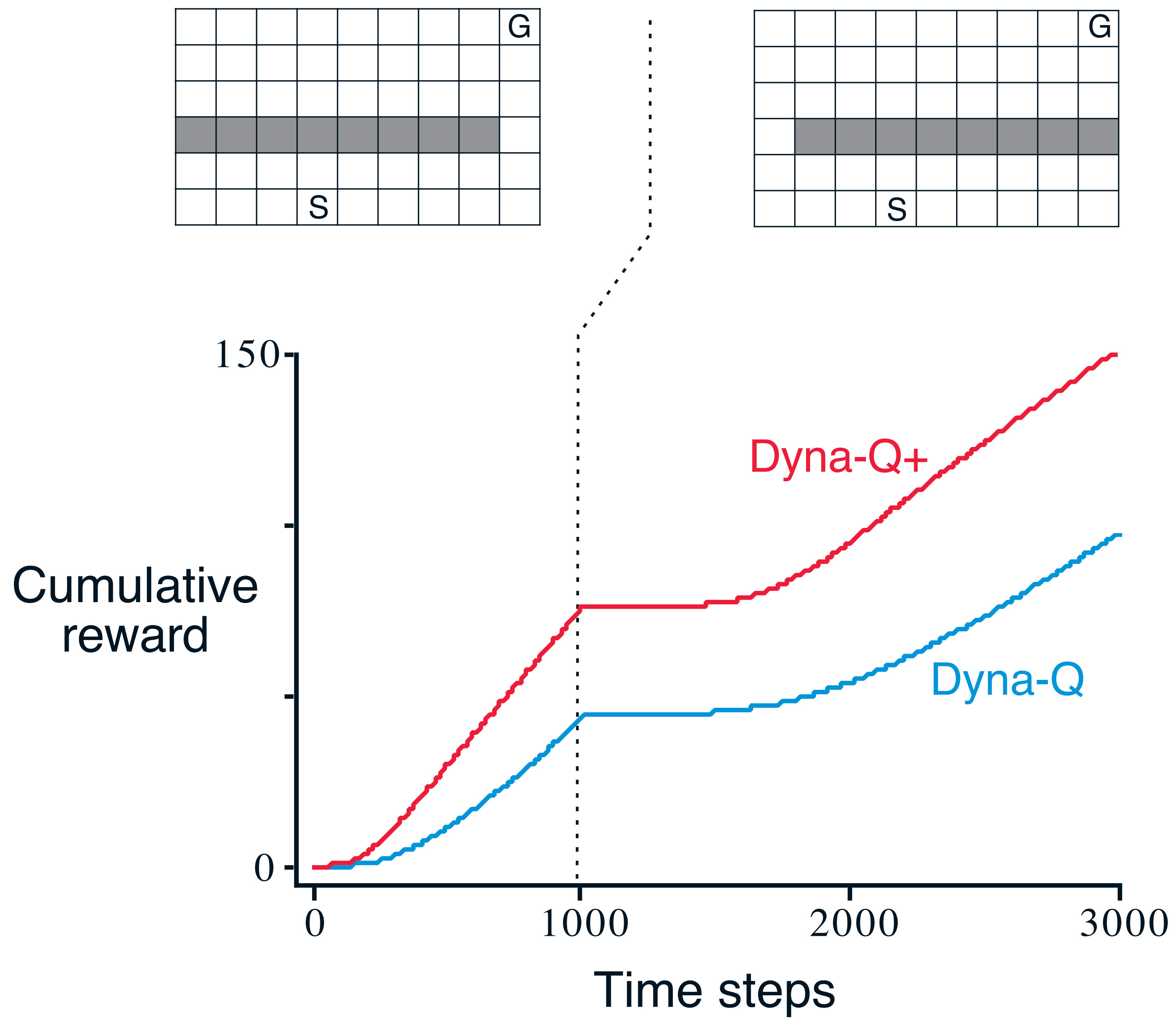

Figure 8.4: Average performance of Dyna agents on a blocking task. The left environment was used for the first 1000 steps, the right environment for the rest. Dyna-Q+ is Dyna-Q with an exploration bonus that encourages exploration. ■

Worksheet question

7. (*Exercise 8.3 S&B*) **Challenge Question:** Careful inspection of Figure 8.5 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?

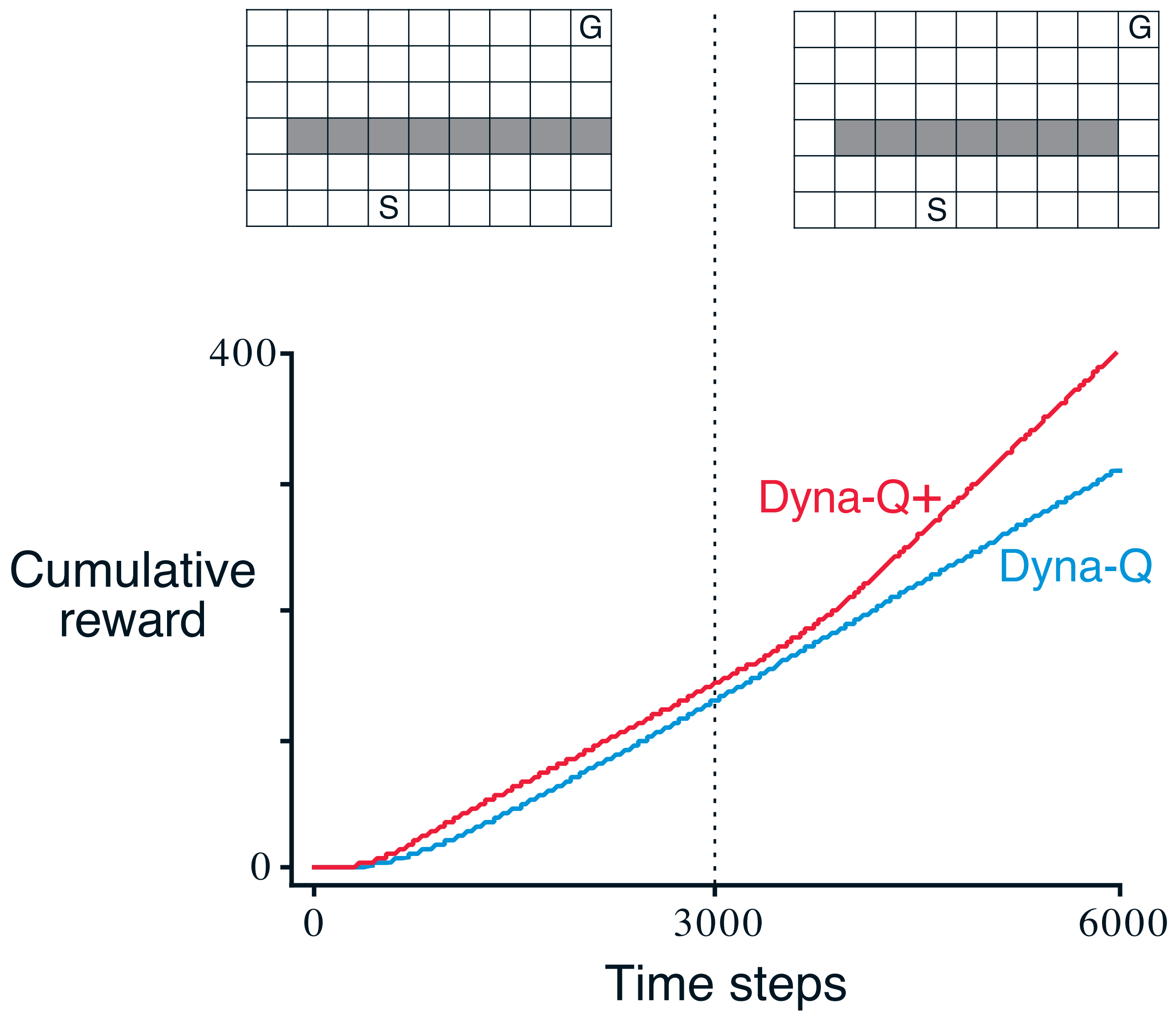

Figure 8.5: Average performance of Dyna agents on a shortcut task. The left environment was used for the first 3000 steps, the right environment for the rest.

■