# Practice Assessment

**TOTAL POINTS 10**

1. Which of the following are the most accurate characterizations of sample models and distribution models? (Select all that apply) 1 point

- [ ] A sample model can be used to compute the probability of all possible trajectories in an episodic task based on the current state and action.
- [ ] A distribution model can be used as a sample model.
- [ ] Both sample models and distribution models can be used to obtain a possible next state and reward, given the current state and action.
- [ ] A sample model can be used to obtain a possible next state and reward given the current state and action, whereas a distribution model can only be used to compute the probability of this next state and reward given the current state and action.

2. Which of the following statements are TRUE for Dyna architecture? (Select all that apply) 1 point

- [ ] Real experience can be used to improve the value function and policy
- [ ] Real experience can be used to improve the model
- [ ] Simulated experience can be used to improve the value function and policy
- [ ] Simulated experience can be used to improve the model

3. Mark all the statements that are TRUE for the tabular Dyna-Q algorithm. (Select all that apply) 1 point

- [ ] The algorithm **cannot** be extended to stochastic environments.

- [ ] The memory requirements for the model in case of a deterministic environment are quadratic in the number of states

- [ ] The environment is assumed to be deterministic.

- [ ] For a given state-action pair, the model predicts the next state and reward

4. Which of the following statements are TRUE? (Select all the apply) 1 point

- [ ] Model-based methods often suffer more from bias than model-free methods, because of inaccuracies in the model.

- [ ] Model-based methods like Dyna typically require more memory than model-free methods like Q-learning.

- [ ] The amount of computation per interaction with the environment is larger in the Dyna-Q algorithm (with non-zero planning steps) as compared to the Q-learning algorithm.

- [ ] When compared with model-free methods, model-based methods are relatively more sample efficient. They can achieve a comparable performance with comparatively fewer environmental interactions.

5. Which of the following is generally the most computationally expensive step of the Dyna-Q algorithm? Assume N>1 planning steps are being performed (e.g., N=20). 1 point

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

- ○ Model learning (step e)
- ○ Direct RL (step d)
- ○ Action selection (step b)
- ○ Planning (Indirect RL; step f)

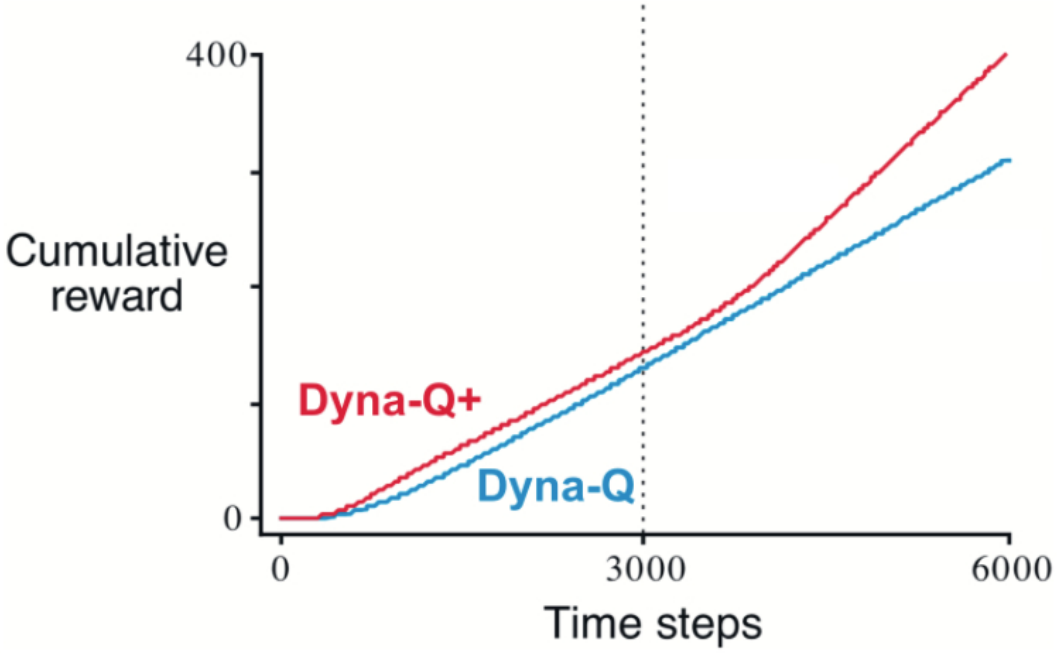
6. What are some possible reasons for a learned model to be inaccurate? (Select all that apply) 1 point

- ☐ The environment has changed.
- ☐ The transition dynamics of the environment are stochastic, and only a few transitions have been experienced.
- ☐ There is too much exploration (e.g., epsilon is epsilon-greedy exploration is set to a high value of 0.5)
- ☐ The agent's policy has changed significantly from the beginning of training.

7. In search control, which of the following methods is likely to make a Dyna agent perform better in problems with a large number of states (like the rod maneuvering problem in Chapter 8 of the textbook)? Recall that search control is the process that selects the starting states and actions in planning. Also recall the navigation example in the video lectures in which a large number of wasteful updates were being made because of the basic search control procedure in the Dyna-Q algorithm. (Select the best option) 1 point

- ○ Select state-action pairs uniformly at random from all previously experienced pairs.
- ○ Start backwards from state-action pairs that have had a non-zero update (e.g., from the state right beside a goal state). This avoids the otherwise wasteful computations from state-action pairs which have had no updates.
- ○ Start with state-action pairs enumerated in a fixed order (e.g., in a gridworld, states top-left to bottom-right, actions up, down, left, right)
- ○ All of these are equally good/bad.

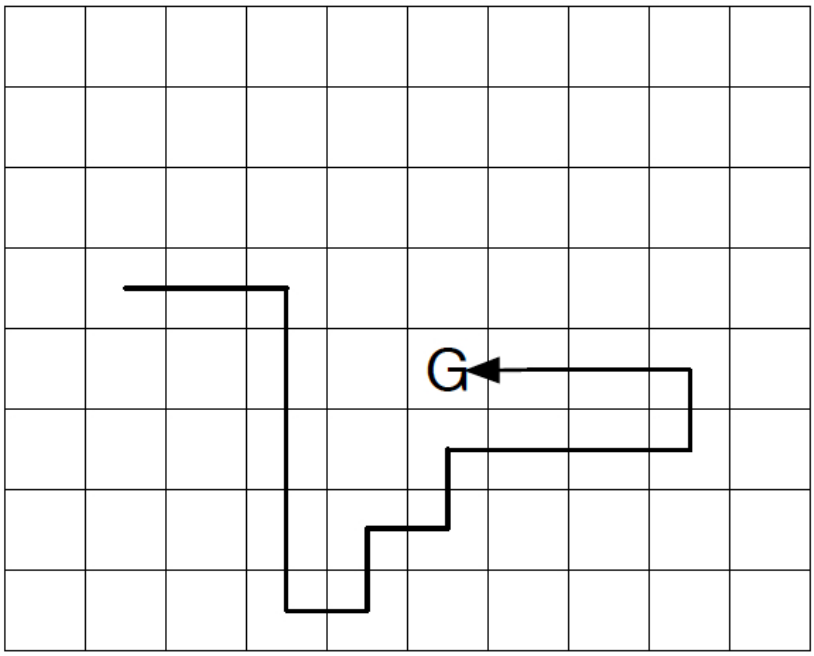
8. In the lectures, we saw how the Dyna-Q+ agent found the newly-opened shortcut in the shortcut maze, whereas the Dyna-Q agent didn't. Which of the following implications drawn from the figure are TRUE? (Select all that apply) 1 point




- [ ] The Dyna-Q+ agent performs better than the Dyna-Q agent even in the first half of the experiment because of the increased exploration.
- [ ] The Dyna-Q agent can never discover shortcuts (i.e., when the environment changes to become better than it was before).
- [ ] The difference between Dyna-Q+ and Dyna-Q narrowed slightly over the first part of the experiment. This is because the Dyna-Q+ agent keeps exploring even when the environment isn't changing.
- [ ] None of the above are true.

9. Consider the gridworld depicted in the diagram below. There are four actions corresponding to up, down, right, and left movements. Marked is the path taken by an agent in a single episode, ending at a location of high reward, marked by the G. In this example the values were all zero at the start of the episode, and all rewards were zero during the episode except for a positive reward at G. 1 point

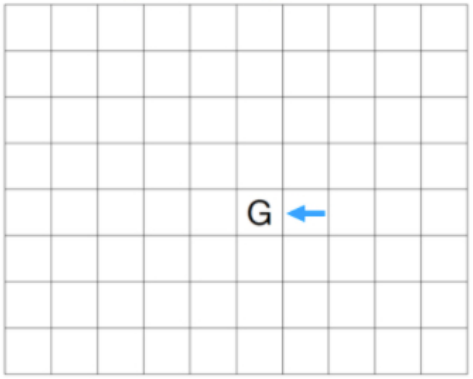
Path taken



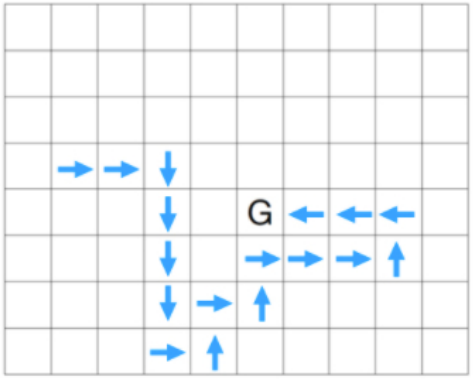

Now which of the following figures best depicts the action values that would've increased by the end of the episode using ***one*-step Sarsa** and ***500*-step-planning Dyna-Q**? (Select the best option)

○

Action values increased by one-step Sarsa

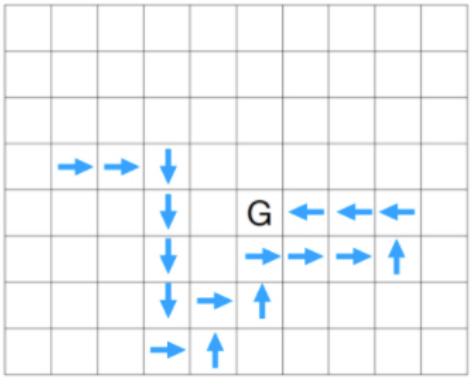



Action values increased by Dyna-Q (500 planning steps)

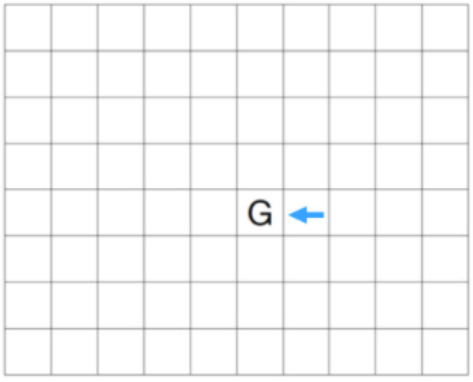


○

Action values increased by one-step Sarsa

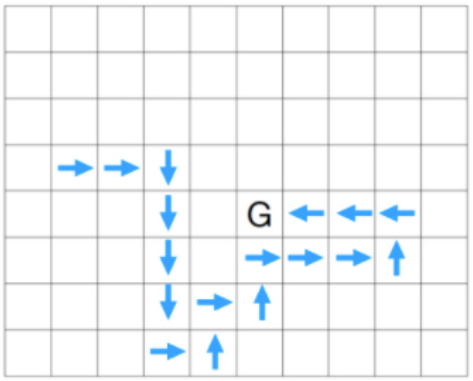



Action values increased by Dyna-Q (500 planning steps)

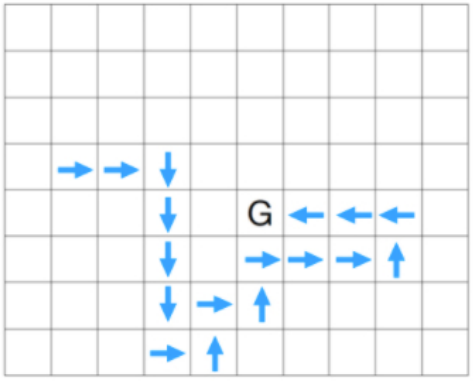



○

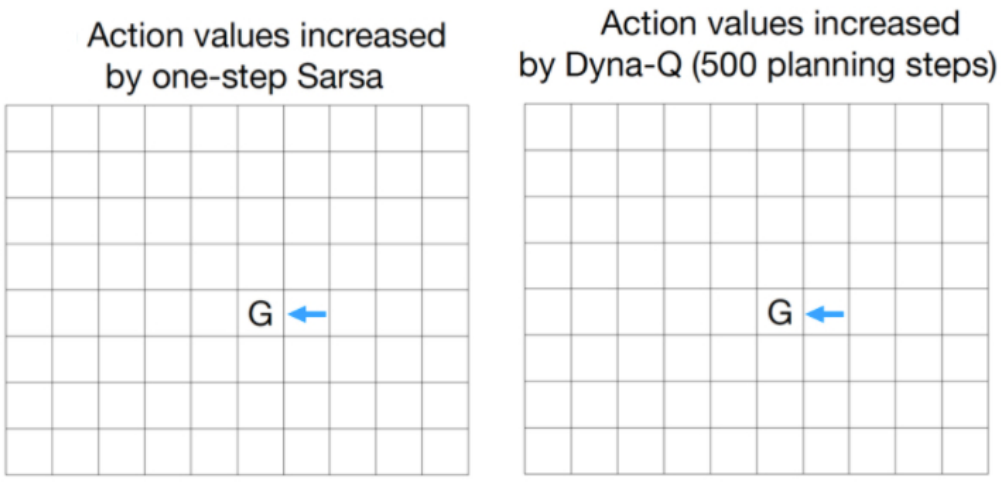
Action values increased by one-step Sarsa




Action values increased by Dyna-Q (500 planning steps)




○




10. Which of the following are planning methods? (Select all that apply) 1 point

- [ ] Dyna-Q
- [ ] Q-learning
- [ ] Expected Sarsa
- [ ] Value Iteration

I, **Dhawal Gupta**, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.

Learn more about Coursera's Honor Code

Save Submit