Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 80%. We keep your highest score.

Next item →

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

1 / 1 point

- [ ] 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.

- [ ] 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.

- [x] Both sample models and distribution models can be used to obtain a possible next state and reward, given the current state and action.

Correct
Correct; given any state and action, you can sample the next state and reward using a sample model or distribution model.

- [x] A distribution model can be used as a sample model.

Correct
Correct; a distribution model contains all the information about the transition dynamics of the system, which can be used to 'sample' new states and rewards given the current state and action – just like a sample model.

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

1 / 1 point

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

Correct
Correct; we do this in the planning step of the tabular Dyna-Q algorithm

- [x] Real experience can be used to improve the model

Correct
Correct; we do this in the model-learning step of the tabular Dyna-Q algorithm

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

Correct
Correct; we do this in the direct-RL step of the tabular Dyna-Q algorithm

- [ ] 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 / 1 point

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

- [x] The environment is assumed to be deterministic.

Correct
Correct; the algorithm assumes that the environment deterministically transitions to a single next state and reward for a given state-action pair. If the environment is stochastic, the update-model step

state and reward for a given state-action pair. If the environment is stochastic, the update-model step in its current form would simply overwrite a state-action pair with a different next state and reward transition. So unless the update-model step is modified, we would be losing a lot of useful information. This may lead to a poor performance even though we are using a planning-based method.

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

> **Correct**
> Correct; this is because in the tabular Dyna-Q algorithm, the model stores the next state and action for every state-action pair that is encountered

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

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

- [x] 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.

> **Correct**
> Correct; we have seen examples of this in the lectures and [Chapter 8] of Sutton and Barto's RL textbook

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

> **Correct**
> Correct; additional memory is required to store the model.

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

> **Correct**
> Correct; the performance of model-based methods depends heavily on the model.

- [x] 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.

> **Correct**
> Correct; apart from the direct RL steps performed in the Q-learning algorithm, Dyna-Q performs additional steps of model-learning and planning.

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 / 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) $A \leftarrow \varepsilon\text{-greedy}(S, Q)$
  (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
    $R, S' \leftarrow Model(S,A)$
    $Q(S,A) \leftarrow Q(S,A) + \alpha\big[R + \gamma \max_a Q(S',a) - Q(S,A)\big]$

- ( ) Model learning (step e)

Model learning (step e)

Direct RL (step d)

Action selection (step b)

Planning (Indirect RL; step f)

Correct

Correct; the planning step performs search control (O(1) with an appropriate dictionary implementation), generates a simulated experience (O(1)), and updates the action-value function (O(|A|)). This is repeated N times, for overall O(N*|A|) time complexity.

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

1 / 1 point

- [x] The environment has changed.

Correct

Correct; if the environment has changed (e.g., a new wall has come up in the gridworld, changing the transition probabilities), then the learned model is no longer accurate

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

Correct

Correct; if there are stochastic transitions from certain states and actions, you might require many samples to form reliable estimates in the model. For a stochastic environment, we can keep counts of the number of times each next state and reward is experienced from each state-action pair. We can use this to estimate probabilities of next states and rewards, from a given state and action.

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 / 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)

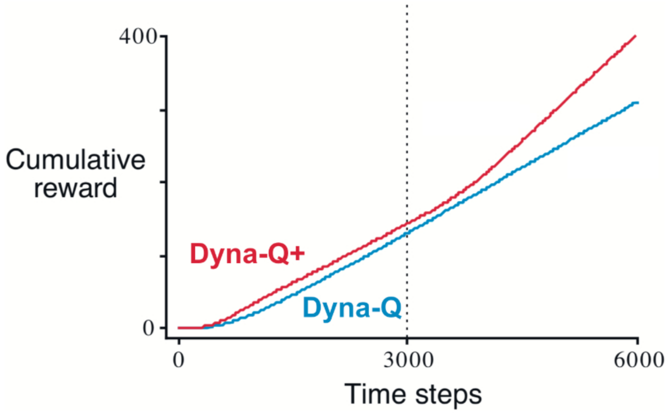
All of these are equally good/bad.

Correct

Correct; such a heuristic allows us to focus the updates on station-action pairs which are expected to have non-zero updates. This speeds up the search for the optimal solution, and is the intuition behind backward focusing and prioritized sweeping (check out Section 8.4 of Sutton and Barto's RL textbook).

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 / 1 point



- [x] The Dyna-Q+ agent performs better than the Dyna-Q agent even in the first half of the experiment because of the increased exploration.

Correct
Correct; the increased exploration due to the reward bonus helps the agent discover the path to the goal relatively faster.

- [ ] The Dyna-Q agent can never discover shortcuts (i.e., when the environment changes to become better than it was before).

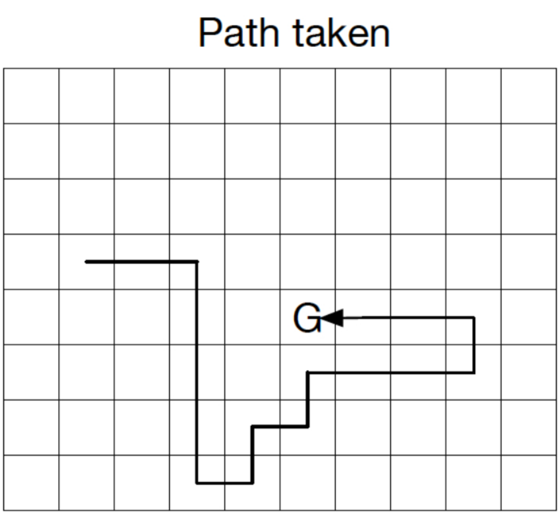
- [x] The difference between Dyna-Q+ and Dyna-Q narrowed slightly over the first part of the experiment. This is because the Dyna-Q+ agent keeps exploring even when the environment isn't changing.

Correct
Correct; such exploration can lead to a slightly suboptimal behaviour even if the optimal policy has been learned for a stationary environment.

- [ ] None of the above are true.

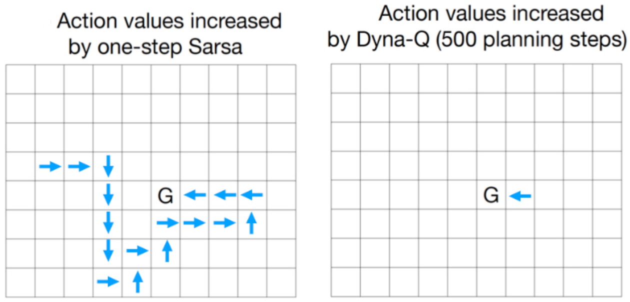
9. Consider the gridworld depicted in the diagram below. There are four actions corresponding to up, down, right, and left movements. Marked is the path taken by an agent in a single episode, ending at a location of high reward, marked by the G. In this example the values were all zero at the start of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

1 / 1 point

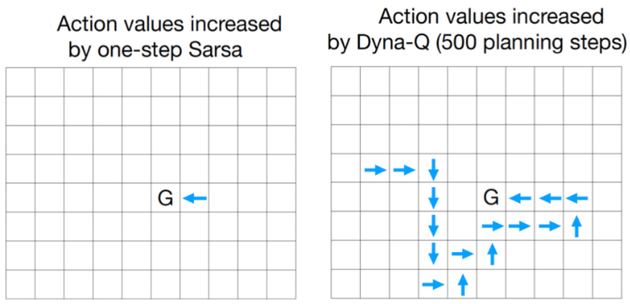


Now which of the following figures best depicts the action values that would've increased by the end of the episode using ***one*-step Sarsa** and ***500*-step-planning Dyna-Q**? (Select the best option)

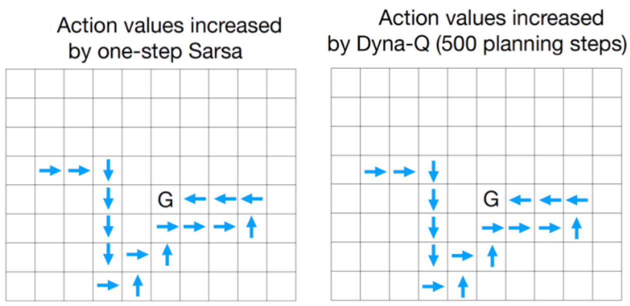
○


Action values increased
by one-step Sarsa
G
Action values increased
by Dyna-Q (500 planning steps)
G

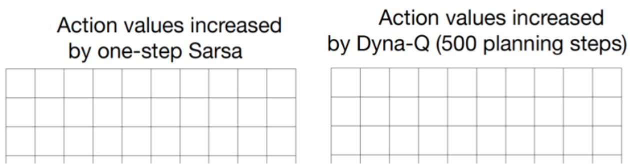
◉

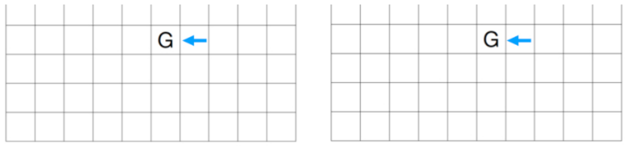

Action values increased
by one-step Sarsa
G
Action values increased
by Dyna-Q (500 planning steps)
G

○


Action values increased
by one-step Sarsa
G
Action values increased
by Dyna-Q (500 planning steps)
G

○


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



Correct
Correct; one-step Sarsa would make a single non-zero update for the state-action pair leading to the goal state, but 500 planning steps would lead to more non-zero steps along this trajectory.

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

1 / 1 point

- [x] Value Iteration

Correct
Correct; Value Iteration is a Dynamic Programming method that uses a model to improve the policy.

- [ ] Expected Sarsa
- [ ] Q-learning
- [x] Dyna-Q

Correct
Correct; Dyna-Q combines model-free Q-learning with planning. It uses both the experience from the environment as well as simulated experiment from the model in order to make updates to improve the policy.