**Congratulations! You passed!**
To Pass 80% or higher

Keep Learning

Latest 100%

# Practice Assessment

Latest Submission Grade
100%

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

1 point

- [ ] 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.

- [ ] 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 the 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.

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

1 point

- [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

- [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

- [ ] Simulated experience can be used to improve the model

- [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

**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 stochastic 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 every state-action pair. This is because the model that is learned by the model step is in a some how would simply overwrite a state-action pair with a different next state and reward it perceives. So unless the update-model step is modified, we would be losing a lot of information about the environment. This may lead to 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

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

1 point

- [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] 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 step performed in the Q-learning algorithm, Dyna-Q performs additional steps of model-learning and planning.

- [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

- [ ] When compared with model-free methods, model-based methods are relatively more sample efficient. They always 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.

**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=50).**

1 point

Tabular Dyna-Q

Initialize $Q(s,a)$ and $Model(s,a)$ for all $s \in S$ and $a \in A(s)$
Loop forever:
(a) $S \leftarrow$ 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 $S$
$R, S' \leftarrow Model(S,A)$
$Q(S,A) \leftarrow Q(S,A) + \alpha[R + \gamma \max_a Q(S',a) - Q(S,A)]$

- [ ] Model learning (step e)
- [ ] Direct RL (step d)
- [ ] Action selection (step b)
- [x] Planning (indirect RL, step f)

Correct
Correct, the planning step performs search control (SC) with an environment dictionary implemented previously, generates a simulated experience (S,R), and updates the action-value function ($Q(s,a)$). This is repeated N times, for overall 100% of time complexity.

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

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.

- [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 and action, in order to estimate the conditional probabilities of next states and rewards, from a given state and action.

- [ ] There is too much exploration (e.g., epsilon in epsilon-greedy exploration is set to a high value of 0.5).

**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, but not intersecting goal states in Chapter 8 of the textbook? Recall that search control is the process that selects the starting states and actions in the simulated experience generated by the model. Assume that the Dyna-agent has already successful updates since 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.

- [ ] Sort 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 encountered in a fixed order (e.g., in a gridworld, starts top-left to bottom-right, and top-up, then, left, right).

- [x] All of these are equally good/bad.

Correct
Correct, such a heuristic allows us to focus the updates on relative action pairs which are expected to have non-zero updates. This speeds up the search for the optimal solutions, 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 environment. Which of the following statements about how the Dyna-Q+ agent differs from the Dyna-Q agent? (Select all that apply)**

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.

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

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

- [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 reached 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, left, right. Rewards are deterministic: the reward is +1 for transitioning to the goal state, G, and 0 on a transition of high reward, marked by the 1's. 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

**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 tabular planning Dyna-Q? (Select the best option)**

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 ensure non-zero values along the trajectory.

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

1 point

- [ ] 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 experience from the model in order to make updates to improve the policy.

- [ ] Expected Sarsa
- [x] Value iteration

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