Sample Quiz on Reinforcement Learning

Instructor: Prof. XYZ

Instructions

Answer the following questions. For computational questions, make sure to show all steps of your work.

Problems

Q1. **Conceptual: Markov Decision Process (MDP)**
Describe the four key elements of a Markov Decision Process (MDP). How does an agent's goal in MDP relate to maximizing cumulative reward?

Solution:
The four key elements of an MDP are:

(a) **States (S):** The set of all possible states in the environment.

(b) **Actions (A):** The set of all actions available to the agent.

(c) **Transition function (P):** The probability of transitioning from one state to another given an action.

(d) **Reward function (R):** The immediate reward received after transitioning from state s to state s' via action a.

An agent's goal is to find a policy that maximizes the expected cumulative reward over time, often referred to as the return.

Q2. **Computational: Bellman Equation for Value Function**
Given an MDP with states $S = \{s_1, s_2\}$, actions $A = \{a_1, a_2\}$, transition probabilities $P(s'|s, a)$, and rewards $R(s, a)$ as follows:

$$P(s_2|s_1, a_1) = 0.8, \quad P(s_2|s_1, a_2) = 0.2$$

$$R(s_1, a_1) = 3, \quad R(s_1, a_2) = 5$$

Compute the value function $V(s_1)$ for discount factor $\gamma = 0.9$.

Solution:
Using the Bellman equation for the value function:

$$V(s_1) = \max_a \left[R(s_1, a) + \gamma \sum_{s'} P(s'|s_1, a)V(s') \right]$$

Assume $V(s_2) = 0$ (for simplicity):

$$V(s_1) = \max\left[3 + 0.9 \cdot (0.8 \cdot 0), 5 + 0.9 \cdot (0.2 \cdot 0)\right] = 5$$

Q3. **Conceptual: Bias and Variance in TD and MC methods**
Explain the trade-off between bias and variance in Monte Carlo (MC) methods and Temporal Difference (TD) learning.

Solution:
Monte Carlo (MC) methods have zero bias since they update based on actual returns observed at the end of episodes. However, they have high variance because outcomes can vary significantly across episodes. Temporal Difference (TD) learning, on the other hand, updates after each step, which introduces bias (due to bootstrapping) but reduces variance as it does not rely on complete episodes.

Q4. **Computational: Q-learning Update**
Consider an agent using Q-learning with the following Q-values:

$$Q(s_1, a_1) = 2, \quad Q(s_1, a_2) = 3$$

The agent takes action a_2, receives a reward of 4, and transitions to state s_2 where the optimal future action has a Q-value of $Q(s_2, a_*) = 5$. Update the Q-value $Q(s_1, a_2)$ using a learning rate $\alpha = 0.1$ and discount factor $\gamma = 0.9$.

Solution:
The Q-learning update rule is:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a)\right]$$

Substituting the given values:

$$Q(s_1, a_2) = 3 + 0.1\left[4 + 0.9 \cdot 5 - 3\right] = 3 + 0.1 \cdot 5.5 = 3.55$$

Q5. **Conceptual: SARSA vs. Q-learning**
Compare and contrast SARSA and Q-learning in terms of their on-policy and off-policy learning strategies.

Solution:
SARSA is an on-policy method, meaning it updates the Q-values based on the action actually taken by the agent according to its current policy. Q-learning is off-policy, meaning it updates the Q-values assuming the agent always chooses the optimal action in the future, regardless of its current policy.

Q6. **Computational: TD(0) Update**
Given the following sequence of state visits and rewards:

$$s_1 \xrightarrow{r=2} s_2 \xrightarrow{r=1} s_3 \quad \text{with} \quad V(s_2) = 0.5, V(s_3) = 0.8$$

Update $V(s_2)$ using TD(0) with learning rate $\alpha = 0.1$ and discount factor $\gamma = 0.9$.

Solution:
The TD(0) update rule is:

$$V(s) \leftarrow V(s) + \alpha \left[r + \gamma V(s') - V(s)\right]$$

Substituting the values:

$$V(s_2) = 0.5 + 0.1 \left[1 + 0.9 \cdot 0.8 - 0.5\right] = 0.5 + 0.1 \cdot 0.72 = 0.572$$

Q7. **Conceptual: Exploration vs. Exploitation**
Explain the exploration vs. exploitation trade-off in reinforcement learning and describe one method for balancing it.

Solution:
The exploration vs. exploitation trade-off involves choosing between exploring new actions to gather more information (exploration) or exploiting known actions to maximize reward (exploitation). One method to balance this trade-off is the ϵ-greedy method, where the agent mostly takes the action with the highest estimated reward but occasionally (with probability ϵ) chooses a random action.

Q8. **Computational: Policy Evaluation**
Given a policy $\pi(a|s)$ and the following rewards and transition probabilities for an MDP:

$$R(s_1, a_1) = 4, \quad P(s_2|s_1, a_1) = 0.7, \quad P(s_3|s_1, a_1) = 0.3$$

Evaluate the value of state s_1 under the policy using discount factor $\gamma = 0.9$.

Solution:
The value function for policy evaluation is:

$$V^\pi(s_1) = \sum_a \pi(a|s_1) \left[R(s_1, a) + \gamma \sum_{s'} P(s'|s_1, a) V(s')\right]$$

Assuming $V(s_2) = 1$ and $V(s_3) = 2$, and $\pi(a_1|s_1) = 1$:

$$V^\pi(s_1) = 4 + 0.9 \left[0.7 \cdot 1 + 0.3 \cdot 2\right] = 4 + 0.9 \cdot 1.3 = 5.17$$

Q9. **Conceptual: Discount Factor γ**
Why is the discount factor γ used in reinforcement learning, and how does it affect the agent's decision-making?

Solution:
The discount factor γ determines the weight given to future rewards in reinforcement learning. A higher γ means the agent values future rewards more, leading to long-term planning, while a lower γ makes the agent focus on immediate rewards.

Q10. **Computational: Monte Carlo Update**

An agent observes the following sequence of rewards in an episode: $r_1 = 1$, $r_2 = 2$, $r_3 = 3$. Compute the return G_t for $t = 1$ with a discount factor $\gamma = 0.9$.

Solution:

The return G_t is computed as:

$$G_1 = r_1 + \gamma r_2 + \gamma^2 r_3 = 1 + 0.9 \cdot 2 + 0.9^2 \cdot 3 = 1 + 1.8 + 2.43 = 5.23$$