

Roll No :

08/05/2026

Problem 1: Q Learning

A.. Consider a single state state MDP with two actions. That is, $\mathcal{S} = \{s\}$ and $\mathcal{A} = \{a_1, a_2\}$. Assume the discount factor of the MDP is γ and the horizon length to be 1. Both actions yield random rewards, with expected reward for each action being a constant $c \geq 0$. That is,

$$\mathbb{E}(r|a_1) = c \text{ and } \mathbb{E}(r|a_2) = c$$

where $r \sim \mathcal{R}^{a_i}, i \in \{1, 2\}$.

(a) What are the true values of $Q(s, a_1)$, $Q(s, a_2)$ and $V^*(s)$? (1 Point)

(b) Let us now consider that the first action a_1 always gives a constant reward of c whereas the second action a_2 gives a reward $c + \mathcal{N}(-0.1, 1)$ (normal distribution with mean -0.1 and unit variance). Which is the better action to take in expectation ? Would the TD control algorithms like Q-learning or SARSA control, trained using finite samples, always favor the action that is best in expectation ? Explain. (3 Points)

B.. A RL agent collects experiences of the form $(s_t, a_t, r_{t+1}, s_{t+1}, a_{t+1})$ to update Q values. At each time step, to choose an action, the agent follows a fixed policy π with probability 0.5 or chooses an action in uniform random fashion. Assume the updates are applied infinitely often, state-action pairs are visited infinitely often, the discount factor $\gamma < 1$ and the learning rate scheduling is appropriate.

(a) The Q learning agent performs following update

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_t [r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)]$$

Will this update converge to the optimal Q function ? Why or Why not ? If not, will it converge to anything at all ? (3 Points)

(b) Another reinforcement learning called SARSA agent, performs the following update

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha_t [r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

Will this update converge to the optimal Q function ? Why or Why not ? If not, will it converge to anything at all ? (3 Points)

Problem 2 : Deep Q Learning

A. In standard supervised learning, optimization algorithms like Stochastic Gradient Descent assume that data samples are Independent and Identically Distributed (i.i.d.). In Reinforcement Learning, an agent typically collects transitions (s_t, a_t, r_t, s_{t+1}) sequentially.

- (a) Explain why training a neural network on a raw sequential stream of experience violates the i.i.d. assumption and how this temporal correlation can lead to instability or divergence in the parameters ϕ . (1.5 Points)
- (b) Describe how the use of an Experience Replay buffer and uniform random sampling from that buffer recovers the i.i.d. property required for stable optimization. (1.5 Points)

B. The Deep Q-Network (DQN) algorithm utilizes a Target Network with parameters ϕ^- that are kept frozen for a fixed interval, while the Online Network parameters ϕ are updated at every step.

- (a) Describe the "chasing its own tail or moving target " phenomenon that occurs if a single network (where $\phi = \phi^-$ at all times) is used to calculate both the current estimate $Q(s, a; \phi)$ and the target value $r + \gamma \max_{a'} Q(s', a'; \phi)$. (1.5 Points)
- (b) How does maintaining a separate, slowly-updating ϕ^- create a more stationary optimization objective? Justify your answer in the context of the loss function landscape. (1.5 Points)

C. Standard DQN is prone to "Overestimation Bias" because it uses the same parameters ϕ to both select the action that maximizes the Q-value and to estimate the value of that action.

- (a) Given that Q -value estimates are inherently noisy, explain how the $\max$ operator in the target calculation $y = r + \gamma \max_{a'} Q(s', a'; \phi)$ leads to a systematic positive bias in value estimation. (1.5 Points)

- (b) In Double DQN, the target calculation is modified such that the Online Network (ϕ) selects the action $a^* = \arg \max_a Q(s', a; \phi)$, but the Target Network (ϕ^-) provides the value estimate $Q(s', a^*; \phi^-)$. Provide a reasoning for why this decoupling reduces the likelihood of noise-driven overestimation. (1.5 Points)

Problem 3 : Policy Gradient Methods

A. The policy gradient estimate is generally expressed as $g = \mathbb{E}[\nabla_{\theta} \log \pi_{\theta}(a|s) \Psi_t]$. One common form for Ψ_t is the return $G_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$, which represents the total discounted reward from time t until the end of the horizon.

- (a) Explain the conceptual role of the term $\nabla_{\theta} \log \pi_{\theta}(a|s)$ in the gradient update. If a specific action a_t results in a high positive value of Ψ_t , how does this term facilitate the "reinforcement" of that action within the policy distribution? (1.5 Points)
- (b) REINFORCE is known to suffer from high variance because $\Psi_t = G_t$ is calculated from a single stochastic trajectory. Describe how subtracting a state-dependent baseline $b(s)$ from Ψ_t (i.e., using $\Psi_t = G_t - b(s)$) reduces variance without introducing bias into the gradient estimate. (1.5 Points)

B. Actor-Critic architectures attempt to improve upon vanilla policy gradients by utilizing two function approximators: an Actor (π_{θ}) and a Critic (V_{ϕ}). In this framework, Ψ_t is typically replaced by the Temporal Difference (TD) error: $\delta_t = r_t + \gamma V_{\phi}(s_{t+1}) - V_{\phi}(s_t)$.

- (a) Define the distinct roles of the parameters θ and ϕ . Specifically, explain how the Critic network (ϕ) assists the Actor (θ) in reducing the variance of the gradient updates compared to the REINFORCE algorithm. (1.5 Points)
- (b) By using the Critic's estimate V_{ϕ} to compute the update, the algorithm is said to "bootstrap." Explain the fundamental trade-off involved here: why does this transition from the empirical return G_t to the learned estimate V_{ϕ} reduce variance but simultaneously introduce bias into the policy gradient? (1.5 Points)

Problem 4 : Multi Arm Bandits

Suppose you are a data scientist at a company X and are trying to develop a new feature for a product based on data analytics. We need to determine whether or not making this feature will have a positive (statistically significant) impact for the company based on a core metric, such as efficient usage translating into more revenue. Assume that the impacts are known as soon as the feature is exposed to a subject. To this end,

the company considers the standard A/B testing where we compare the same metric across the A group (current system, the control group) vs the B group (new system, the experimental group). If the B group has a significant improvement in this metric over the A group, we should replace the current system with the new one. Typically A/B testing involves putting a large percent of the population in the A group (say 99 percent), and a small percent of the population in the B group (remaining 1 percent).

- (a) What consequences can the company face when testing is done with such heavily imbalanced distribution of subjects via A/B testing ? (2 Points)
- (b) Suppose that the company chooses to ditch the A/B testing route, and adopt MAB framework with 2 arms corresponding to the current system and the new system. How would the MAB framework address the challenges arising from A/B wherein there is heavy imbalanced distribution of subjects ? (2 Points)
- (c) Assume that the reward metric for the new feature is significantly higher than the old feature. Would a plain vanilla UCB algorithm assign subjects to the old feature after the initialization phase ? Explain (1 Point)

ALL THE BEST