Reinforcement Learning

Assignment 3 (Theoretical Questions)

Group Members

Group ID - 31
Matheus da Silva Araujo – 261218407
Miguel Ángel Carrillo – 261205372

February 27, 2026

1. Maximization Bias in Q-Learning. Given an MDP with a state space S and an action space A. Assume the learned action-value estimate follows a simple additive noise model:

$$Q(s,a) = Q^*(s,a) + \epsilon_{s,a}, \text{ where } \epsilon_{s,a} \overset{i.i.d.}{\sim} \mathcal{N}(0,1).$$

Here, $Q^*(s,a)$ denotes the true optimal action-value function, and the noise terms $\{\epsilon_{s,a}\}$ represent independent estimation errors added to each state-action value.

(a) [6 points] Show that even if each $Q(s,a)$ is an unbiased estimator of $Q^*(s,a)$, the maximized estimate is biased upward, i.e.,

$$\forall s,\ \mathbb{E}\left[\max_{a\in A} Q(s,a)\right] \geq \max_{a\in A} Q^*(s,a).$$

Solution to 1(a): Since $Q(s,a) = Q^*(s,a) + \epsilon_{s,a}$ and $\mathbb{E}[\epsilon_{s,a}] = 0$, it follows that

$$\mathbb{E}[Q(s,a)] = Q^*(s,a),$$

meaning that each estimator is unbiased.

The function max() is convex. Applying Jensen's inequality, it follows that

$$\mathbb{E}\left[\max_{a\in A} Q(s,a)\right] \geq \max_{a\in A} \mathbb{E}[Q(s,a)] = \max_{a\in A} Q^*(s,a).$$

If $|A| \geq 2$, the inequality is strict due to independent noise. Hence, the maximized estimate is biased upward.

(b) [5 points] Recall that Q-learning uses the bootstrapped target

$$Y_{QL} = r + \gamma \max_{a'} Q(s',a').$$

Using part (a), explain why Y_{QL} introduces overestimation bias. Consider repeated updates where each new target uses the previously learned Q-values. Explain how the bias evolves through bootstrapping over time.

Solution to 1(b): From part (a),

$$\mathbb{E}\left[\max_{a'} Q(s', a')\right] \geq \max_{a'} Q^*(s', a').$$

Therefore, it followss that

$$\mathbb{E}[Y_{QL}] = r + \gamma \mathbb{E}\left[\max_{a'} Q(s', a')\right] \geq r + \gamma \max_{a'} Q^*(s', a').$$

Thus, the target is positively biased.

Since Q-learning is bootstrapped, each update uses previously biased estimates. The overestimation propagates through future targets and compounds over iterations, which causes overestimation bias.

(c) [5 points] The Expected SARSA update uses the target

$$Y_{ES} = r + \gamma \sum_{a'} \pi(a'|s') Q(s', a').$$

Show that if each $Q(s', a')$ is unbiased, then Y_{ES} is also unbiased.

Solution to 1(c): Assume that for every action a',

$$Q(s', a') = Q^*(s', a') + \epsilon_{s',a'}, \quad \text{with } \mathbb{E}[\epsilon_{s',a'}] = 0.$$

Hence,

$$\mathbb{E}[Q(s', a')] = Q^*(s', a').$$

The expected SARSA target is

$$Y_{ES} = r + \gamma \sum_{a'} \pi(a'|s') Q(s', a').$$

Taking expectation of the expression above gives

$$\mathbb{E}[Y_{ES}] = r + \gamma \mathbb{E}\left[\sum_{a'} \pi(a'|s') Q(s', a')\right].$$

Since expectation is linear and $\pi(a'|s')$ is fixed given s', and also using the unbiasedness asumption from the problem statement in the second expression below, it follows that

$$= r + \gamma \sum_{a'} \pi(a'|s') \mathbb{E}[Q(s', a')] = r + \gamma \sum_{a'} \pi(a'|s') Q^*(s', a').$$

The right-hand side is the corresponding target constructed using the true action-value function. Because no maximization operator is involved and expectation is distributed through the summation, no additional bias term appears. It follos that Y_{ES} is an unbiased estimator of its corresponding true target.

(d) [8 points] Double learning. Now, consider another estimator,

$$Q'(s,a) = Q^*(s,a) + \epsilon'_{s,a}, \text{ where } \epsilon'_{s,a} \overset{i.i.d.}{\sim} \mathcal{N}(0,1).$$

Double Q-learning uses the target

$$Y_{DQ} = r + \gamma Q'\big(s', \arg\max_{a'} Q(s',a')\big).$$

Assume the independence between ϵ and ϵ'. Compare

$$\mathbb{E}\Big[Q'\big(s', \arg\max_{a'} Q(s',a')\big)\Big] \quad \text{and} \quad \mathbb{E}\Big[\max_{a'} Q(s',a')\Big],$$

and explain why double learning alleviates the maximization bias in standard Q-learning. Hint: You can let $a^* \in \arg\max_{a'} Q(s',a')$. Note that a^* is also a random variable!

Solution to 1(d): Let

$$a^* = \arg\max_{a'} Q(s',a'),$$

where

$$Q(s',a') = Q^*(s',a') + \epsilon_{s',a'}.$$

The action a^* is a random variable since it depends on the noise ϵ.

In standard Q-learning,

$$\max_{a'} Q(s',a') = Q(s',a^*) = Q^*(s',a^*) + \epsilon_{s',a^*}.$$

Taking expectation,

$$\mathbb{E}\left[\max_{a'} Q(s',a')\right] = \mathbb{E}[Q^*(s',a^*)] + \mathbb{E}[\epsilon_{s',a^*}].$$

It remains to show that $\mathbb{E}[\epsilon_{s',a^*}] > 0$. Rearranging the expression above,

$$\mathbb{E}[\epsilon_{s',a^*}] = \mathbb{E}\left[\max_{a'} Q(s',a')\right] - \mathbb{E}[Q^*(s',a^*)].$$

Since a^* maximizes Q but not necessarily Q^*, it holds for every realization that $Q^*(s',a^*) \leq \max_{a'} Q^*(s',a')$, and therefore

$$\mathbb{E}[Q^*(s',a^*)] \leq \max_{a'} Q^*(s',a').$$

From part (a),

$$\mathbb{E}\left[\max_{a'} Q(s',a')\right] \geq \max_{a'} Q^*(s',a').$$

Combining the two inequalities,

$$\mathbb{E}[\epsilon_{s',a^*}] = \mathbb{E}\left[\max_{a'} Q(s',a')\right] - \mathbb{E}[Q^*(s',a^*)] \geq \max_{a'} Q^*(s',a') - \max_{a'} Q^*(s',a') = 0.$$

For $|A| \geq 2$, the inequality in part (a) is strict, so $\mathbb{E}[\epsilon_{s',a^*}] > 0$, meaning that this is an upward bias.

For double Q-learning, considering an independent estimator

$$Q'(s', a) = Q^*(s', a) + \epsilon'_{s',a},$$

where ϵ' is independent of ϵ and satisfies $\mathbb{E}[\epsilon'_{s',a}] = 0$.

Double Q-learning evaluates $Q'(s', a^*)$. Taking expectation,

$$\mathbb{E}[Q'(s', a^*)] = \mathbb{E}[Q^*(s', a^*)] + \mathbb{E}[\epsilon'_{s',a^*}].$$

Because ϵ' is independent of ϵ, it is also independent of the selection of a^* (which depends only on ϵ). Therefore,

$$\mathbb{E}[\epsilon'_{s',a^*}] = \mathbb{E}[\epsilon'_{s',a}]\big|_{a=a^*} = 0,$$

which gives

$$\mathbb{E}[Q'(s', a^*)] = \mathbb{E}[Q^*(s', a^*)].$$

Comparing the two, it is possible to see that

$$\mathbb{E}\left[\max_{a'} Q(s', a')\right] = \mathbb{E}[Q^*(s', a^*)] + \underbrace{\mathbb{E}[\epsilon_{s',a^*}]}_{>0},$$

whereas

$$\mathbb{E}[Q'(s', a^*)] = \mathbb{E}[Q^*(s', a^*)].$$

Double Q-learning removes the positive bias term $\mathbb{E}[\epsilon_{s',a^*}]$ by decoupling action selection and evaluation. Although $\mathbb{E}[Q^*(s', a^*)]$ may differ from $\max_{a'} Q^*(s', a')$ due to noisy selection, the overestimation caused by maximization of noise estimates is eliminated.

2. Policy Improvement Theorem and Policy Gradient.

(a) [8 points] Let π be a policy with action-value function $q_\pi(s, a)$ and state-value function $v_\pi(s)$. Define the greedy (argmax) policy with respect to q_π as:

$$\pi_{greedy}(a|s) = \begin{cases} 1, & a \in \arg\max_{a'} q_\pi(s, a'), \\ 0, & \textit{otherwise.} \end{cases}$$

Consider the stochastic mixture policy

$$\pi'(a|s) = (1 - \alpha)\pi(a|s) + \alpha\pi_{greedy}(a|s),$$

where $\alpha \in (0, 1]$ is a scalar constant. Prove that the mixture policy satisfies the policy improvement condition

$$\sum_a \pi'(a|s) q_\pi(s, a) \geq v_\pi(s).$$

Conclude that π' is an improved policy.

Solution to 2(a): Analyzing the definition of π':

$$\sum_a \pi'(a|s) q_\pi(s, a) = \sum_a \big[(1 - \alpha)\pi(a|s) + \alpha\pi_{\text{greedy}}(a|s)\big] q_\pi(s, a).$$

Distributing the sum gives

$$= (1-\alpha)\sum_a \pi(a|s)q_\pi(s,a) + \alpha\sum_a \pi_{\text{greedy}}(a|s)q_\pi(s,a).$$

By definition of the state-value function,

$$\sum_a \pi(a|s)q_\pi(s,a) = v_\pi(s).$$

Since π_{greedy} selects an action in $\arg\max_{a'} q_\pi(s,a')$,

$$\sum_a \pi_{\text{greedy}}(a|s)q_\pi(s,a) = \max_a q_\pi(s,a).$$

Therefore,

$$\sum_a \pi'(a|s)q_\pi(s,a) = (1-\alpha)v_\pi(s) + \alpha\max_a q_\pi(s,a).$$

Because a maximum is always greater than or equal to an average,

$$\max_a q_\pi(s,a) \geq v_\pi(s),$$

which implies

$$(1-\alpha)v_\pi(s) + \alpha\max_a q_\pi(s,a) \geq v_\pi(s).$$

Thus,

$$\sum_a \pi'(a|s)q_\pi(s,a) \geq v_\pi(s).$$

By the Policy Improvement Theorem, π' is therefore an improved policy.

(b) [8 points] The policy gradient theorem states

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\big[\nabla_\theta \log \pi_\theta(a|s) Q_{\pi_\theta}(s,a)\big].$$

Consider learning the critic by minimizing the objective

$$L(\phi) = \mathbb{E}_{\pi_\theta}\left[\big(Q_\phi(s,a) - Q_{\pi_\theta}(s,a)\big)^2\right],$$

where $Q_\phi(s,a) \approx Q_{\pi_\theta}(s,a)$. This means

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{\pi_\theta}\big[\nabla_\theta \log \pi_\theta(a|s) Q_\phi(s,a)\big].$$

(i) [2 points] Derive the gradient $\nabla_\phi L(\phi)$.

Solution to 2(b)(i): The objective is

$$L(\phi) = \mathbb{E}_{\pi_\theta}\Big[\big(Q_\phi(s,a) - Q_{\pi_\theta}(s,a)\big)^2\Big].$$

Differentiating with respect to ϕ and applying the chain rule, it follows the gradient:

$$\nabla_\phi L(\phi) = \mathbb{E}_{\pi_\theta}\Big[2\big(Q_\phi(s,a) - Q_{\pi_\theta}(s,a)\big)\nabla_\phi Q_\phi(s,a)\Big].$$

(ii) [6 points] Then, let $\phi^* = \arg\min_\phi L(\phi)$. *Assume*

$$\nabla_\phi Q_\phi(s,a) = \nabla_\theta \log \pi_\theta(a|s).$$

Show that under the given conditions, although Q_{ϕ^*} *need not equal* Q_{π_θ} *pointwise, the policy gradient using* Q_{ϕ^*} *is exact,*

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\big[\nabla_\theta \log \pi_\theta(a|s) Q_{\phi^*}(s,a)\big].$$

Solution to 2(b)(ii): Let $\phi^* = \arg\min_\phi L(\phi)$. At the minimum,

$$\nabla_\phi L(\phi^*) = 0.$$

From part (i), this implies

$$\mathbb{E}_{\pi_\theta}\Big[\big(Q_{\phi^*}(s,a) - Q_{\pi_\theta}(s,a)\big)\nabla_\phi Q_{\phi^*}(s,a)\Big] = 0.$$

Using the compatibility condition

$$\nabla_\phi Q_\phi(s,a) = \nabla_\theta \log \pi_\theta(a|s),$$

it follows that

$$\mathbb{E}_{\pi_\theta}\Big[\big(Q_{\phi^*}(s,a) - Q_{\pi_\theta}(s,a)\big)\nabla_\theta \log \pi_\theta(a|s)\Big] = 0.$$

Rearranging,

$$\mathbb{E}_{\pi_\theta}\big[\nabla_\theta \log \pi_\theta(a|s) Q_{\phi^*}(s,a)\big] = \mathbb{E}_{\pi_\theta}\big[\nabla_\theta \log \pi_\theta(a|s) Q_{\pi_\theta}(s,a)\big].$$

By the policy gradient theorem, the right-hand side equals $\nabla_\theta J(\theta)$. Therefore,

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\big[\nabla_\theta \log \pi_\theta(a|s) Q_{\phi^*}(s,a)\big].$$

Thus, even Q_{ϕ^*} need not to be to equal Q_{π_θ}, its policy gradient is the exactly the same.

1 Author Contribution

All the team members contributed equally to the assignment. Miguel led the coding questions, while Matheus led the theory questions. Subsequently, Miguel contributed to verify correctness and fix errors in the theory part, while Matheus did the same for the coding exploration questions.

2 LLM Usage Statement

For the theory questions, ChatGPT was used to fix LaTeX compilation errors, e.g., *The following block of equations is not compiling on TeXstudio, how to fix it?*. No LLM usage for the coding part.