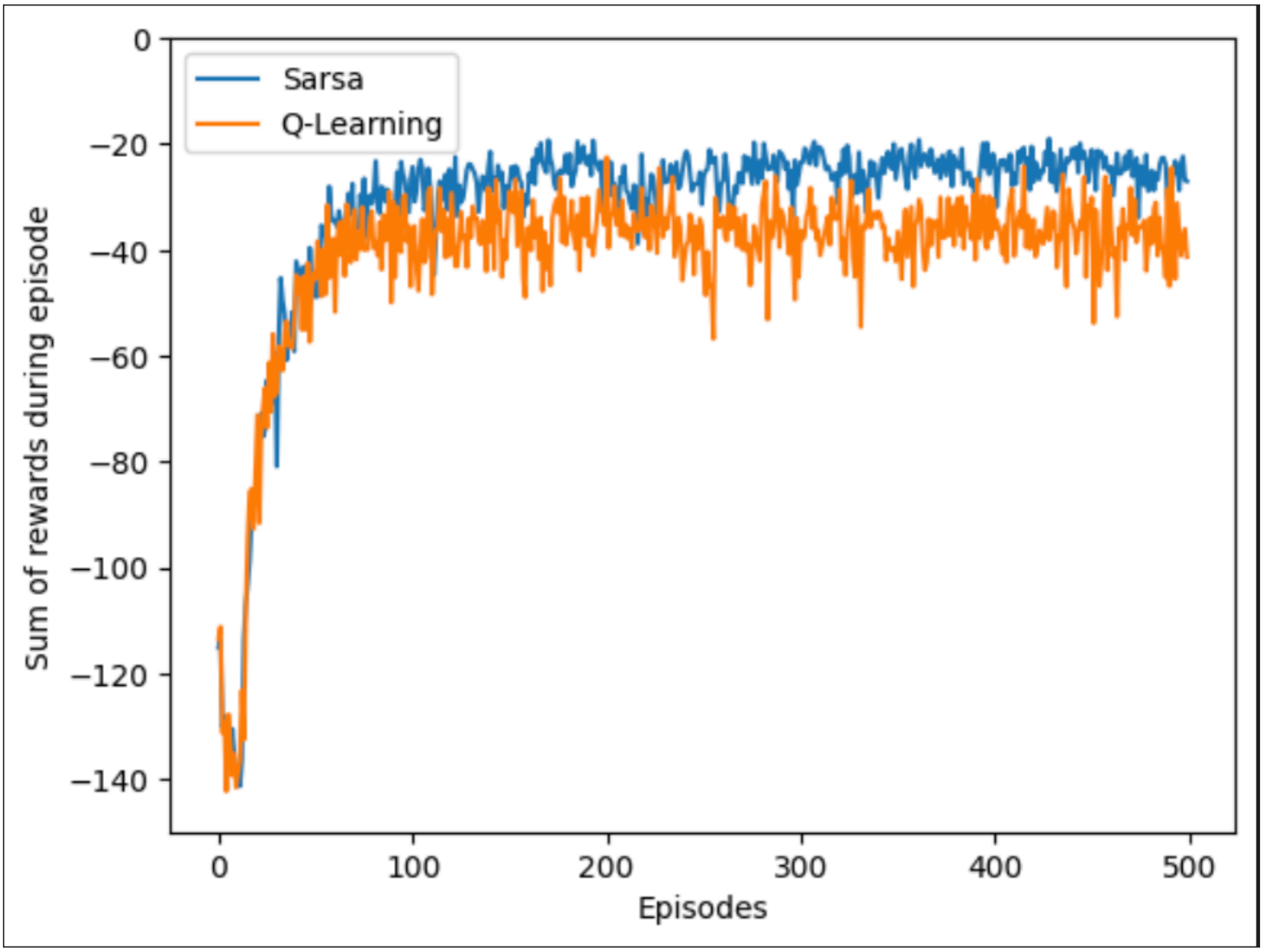
Question 1

- **The total reward in each episode for Q-learning and SARSA:**



- **SARSA's resulting policy after 500 episodes:**

```
State: (0, 2, up, -1)
State: (0, 1, up, -1)
State: (1, 1, right, -1)
State: (1, 0, up, -1)
State: (2, 0, right, -1)
State: (3, 0, right, -1)
State: (4, 0, right, -1)
State: (5, 0, right, -1)
State: (6, 0, right, -1)
State: (7, 0, right, -1)
State: (8, 0, right, -1)
State: (9, 0, right, -1)
State: (10, 0, right, -1)
State: (11, 0, right, -1)
State: (11, 1, down, -1)
State: (11, 2, down, -1)
Goal!
State: (11, 3, down, 0)
```

```
 O X X X X X X X X X X X
X X O O O O O O O O O X
X O O O O O O O O O O X
X C C C C C C C C C C X
```

- **Q-learning’s resulting policy after 500 episodes:**

```
State: (0, 2, up, -1)
State: (1, 2, right, -1)
State: (2, 2, right, -1)
State: (3, 2, right, -1)
State: (4, 2, right, -1)
State: (5, 2, right, -1)
State: (6, 2, right, -1)
State: (7, 2, right, -1)
State: (8, 2, right, -1)
State: (9, 2, right, -1)
State: (10, 2, right, -1)
State: (11, 2, right, -1)
Goal!
State: (11, 3, down, 0)
```

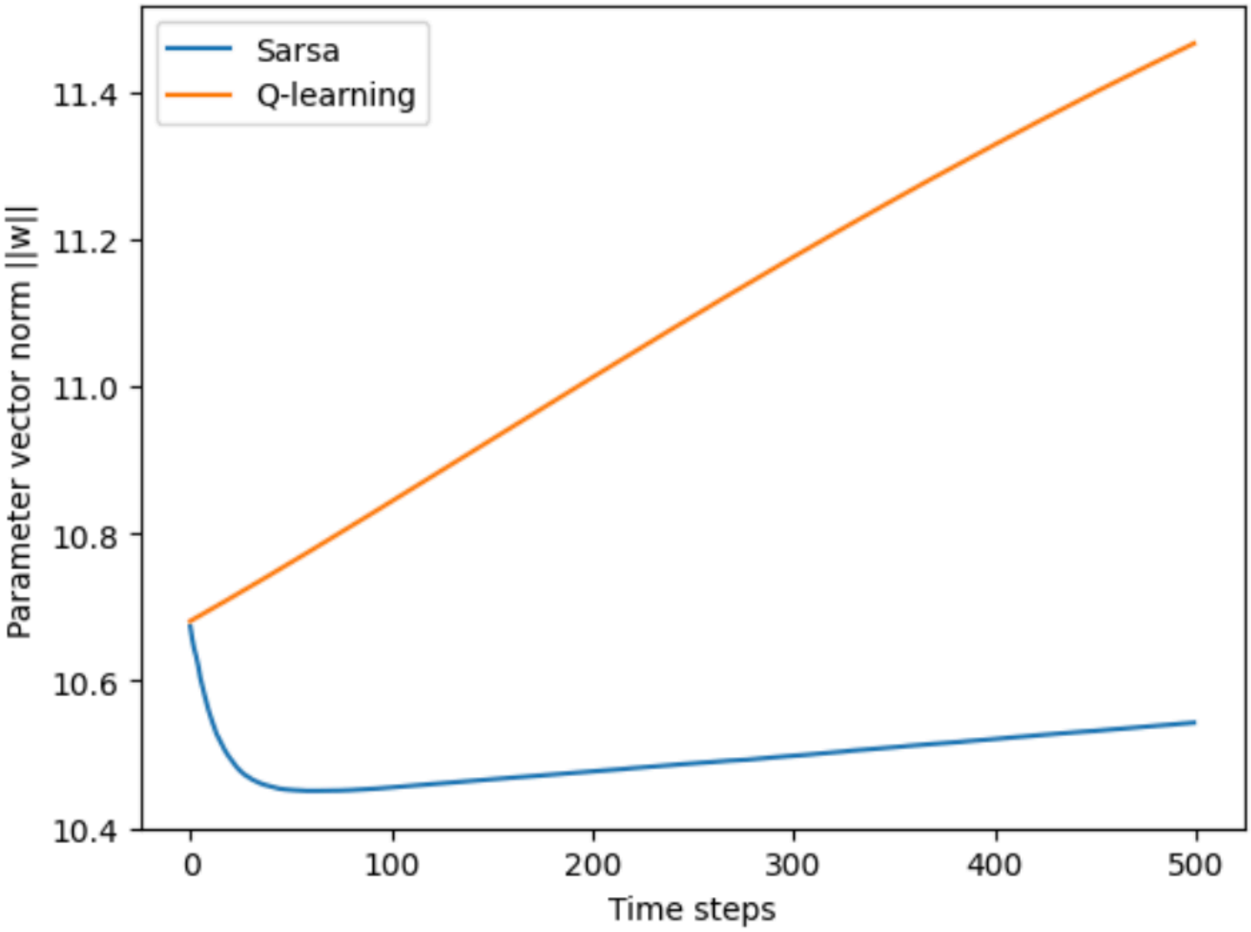
```
O O O O O O O O O O O O
O O O O O O O O O O O O
X X X X X X X X X X X X
X C C C C C C C C C C X
```

- **Interpretation:**

To obtain this plot, I averaged the sum of rewards across 50 different runs. In the first episodes, both SARSA and Q-learning are making several suboptimal decisions, obtaining very negative sums of rewards between -110 and -140. As the number of episodes increases, the sums of rewards for both algorithms increase very fast, stabilising around after 100 episodes. After these 100 episodes, we can observe the trend that most distinguishes both algorithms for this scenario: SARSA receives slightly higher sums of rewards and has less variability than Q-learning. This is due to SARSA, despite not achieving the optimal policy, being an on-policy algorithm, it uses the next action it takes to update its q-values. This way, it learns that going close to the edge of the cliff can be costly (even after updating the q-values for several episodes, it still performs exploratory actions due to the epsilon greedy policy) and stays away from the cliff. Q-learning, on the other hand, is an off-policy method, meaning that it does not use the policy used to choose the next action to update the q-values. It uses a slightly different policy, in which it chooses the action with highest q-value and does not take into account the epsilon-exploration, so it does not think that going to the edge is costly. This allows it to learn the optimal policy, but it ends up performing worse in terms of sums of rewards because the agent will end up falling in the cliff more easily, since all it takes is an exploratory action to fall into the cliff.

Question 2

- **Compare the norm of the parameter vector at each time step for Q-learning and SARSA:**



- **Interpretation:**

SARSA: The norm of the parameter vector for SARSA seems to decrease initially and then stabilises relatively quickly. This indicates that the SARSA algorithm is converging to a solution where the updates to the weights become smaller over time, suggesting that the algorithm is stabilising its policy. The values of the vector norm decreasing also match SARSA's conservative updates, but may also indicate a poor policy.

Q-learning: In contrast, the parameter vector norm for Q-learning is consistently increasing over time. This trend suggests that the weights are growing without stabilisation, which could be a sign of divergence or that the algorithm is still actively learning and adjusting its policy. Most likely it did not converge.

Comparison: SARSA seems to be more stable in this scenario compared to Q-learning. This could be due to SARSA being an on-policy algorithm, meaning it learns the value of the policy it follows, possibly leading to more conservative updates. The increasing trend in the Q-learning norm could be a result of its off-policy nature, where it learns the value of the best possible policy while following another policy. This might cause larger updates if the environment has many states and actions to explore. The increasing norm in Q-learning might also indicate potential overfitting or overshooting of the value function approximation. It's possible that Q-learning's greedy nature in updating its policy could be leading to over-estimations of the action values. This plot suggests that SARSA might be more appropriate for this particular MDP due tp being more stable.

These results match with my expectations regarding function approximation, which may lead to divergence and poor policies.

Question 3

$$\nabla_\theta v_{\pi_\theta}(s) = \sum_{s' \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p^t_{\pi_\theta}(s'|s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s') q_{\pi_\theta}(s', a) \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta v_{\pi_\theta}(s) = \nabla_\theta \sum_{a \in \mathcal{A}} \pi_\theta(a|s) q_{\pi_\theta}(s,a) \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta v_{\pi_\theta}(s) = \boxed{\sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s) q_{\pi_\theta}(s,a)}$$

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} \mu_\theta(s) \boxed{\sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s) q_{\pi_\theta}(s,a)} \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} \mu_\theta(s) \nabla_\theta v_{\pi_\theta}(s) \Leftrightarrow$$

$$\Leftrightarrow J(\theta) = \sum_{s \in \mathcal{S}} \mu_\theta(s) v_{\pi_\theta}(s)$$

$$q^{\pi_\theta}(s,a) = \sum_{s' \in \mathcal{S}} P^a(s'|s) \left[R^a(s'|s) + \gamma v_{\pi_\theta}(s') \right]$$

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} \mu_\theta(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s) \sum_{s' \in \mathcal{S}} P^a(s'|s) \left[R^a(s'|s) + \gamma v_{\pi_\theta}(s') \right] \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} \mu_\theta(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s) \sum_{s' \in \mathcal{S}} P^a(s'|s) \left[R^a(s'|s) + \gamma \sum_{a' \in \mathcal{A}} \pi_\theta(a'|s') q_{\pi_\theta}(s', a') \right] \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta J(\theta) = \sum_{s' \in \mathcal{S}} \mu(s') \sum_{a \in \mathcal{A}} \nabla_\theta \pi_\theta(a|s) \sum_{s' \in \mathcal{S}} P^a(s'|s) \left[R^a(s'|s) + \gamma \sum_{a' \in \mathcal{A}} \pi_\theta(a'|s') q_{\pi_\theta}(s', a') \right] \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta J(\theta) = \sum_{s'\in\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \mu_t(s') \sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a|s) \sum_{s'\in\mathcal{S}} p^a(s'|s)\left[R^a(s'|s) + \gamma \sum_{a'\in\mathcal{A}} \pi_\theta(a'|s') q_{\pi_\theta}(s',a')\right] \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta J(\theta) = \sum_{s'\in\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t \left[\sum_{s\in\mathcal{S}} \mu_t(s) \sum_{a\in\mathcal{A}} p^a(s'|s) \pi_\theta(a|s)\right] \sum_{a'\in\mathcal{A}} \nabla_\theta \pi_\theta(a'|s') q_\pi(s',a')$$

Question 4

$$\nabla_\theta J(\theta) = \sum_{s\in\mathcal{S}} \mu_\theta(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(a|s)\, q_{\pi_\theta}(s,a) \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta J(\theta) = \sum_{s\in\mathcal{S}} \mu_\theta(s) \sum_{a\in\mathcal{A}} \nabla_\theta \frac{\pi_\theta(a|s)\, q_{\pi_\theta}(s,a)\, \pi_\theta(a|s)}{\pi_\theta(a|s)} \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta J(\theta) = \sum_{s\in\mathcal{S}} \mu_\theta(s) \sum_{a\in\mathcal{A}} \nabla_\theta \log \pi_\theta(a|s)\, q_{\pi_\theta}(s,a)\, \pi_\theta(a|s) \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta J(\theta) = \mathbb{E}_{S\sim\mu_\theta(\cdot), A\sim\pi_\theta(\cdot|S)}\left[\nabla_\theta \log \pi_\theta(A|S)\, q_{\pi_\theta}(S,A)\right) \Leftrightarrow$$

$$\Leftrightarrow \nabla_\theta J(\theta) = \mathbb{E}_{S\sim\mu_\theta(\cdot), A\sim\pi_\theta(\cdot|\cancel{X})}\left[\phi(\cancel{S}|A)\, q_{\pi_\theta}(S,A)\right) \Leftrightarrow$$

Given an arbitrary $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$:

$$(\Gamma_f)(s,a) = \phi^T(s,a)\, \mathbb{E}_{S\sim\mu_\theta(\cdot), A\sim\pi(\cdot|S)}\left[\phi(S,A) f(S,A)\right]^{-1}$$

and

$$\Phi = \mathbb{E}_{S\sim\mu_\theta(\cdot), A\sim\pi(\cdot|S)}\left[\phi(S,A)\,\phi^T(X,A)\right)$$

using $f(s,a) = q_{\pi_\theta}(s,a)$

$$\nabla_\theta J(\theta) = \mathbb{E}_{S\sim\mu(\cdot), A\sim\pi_\theta(\cdot|S)}\left[\phi(S,A)\, \Gamma_\phi q_{\pi_\theta}(S,A)\right)$$