Reinforcement Learning

Homework 2 Solution

Samuel Simão Canada Gomes, 76415

January 30, 2019

Solution for "The Cliff Problem"

The total rewards of each episode averaged over 1000 runs for Q-learning and SARSA are plotted in Figure 1. Both algorithms start with total average reward sums close to -100 given that although the agent can walk freely, as soon as it falls of the cliff, the current episode ends. After 500 episodes, we can see that Q-Learning converges to a lower average total episode reward ($\approx$ -25) than SARSA ($\approx$ -50).

As displayed in Figure 2, SARSA and Q-Learning outputted final policies with different characteristics. Because SARSA considers the same stochastic policy to act and compute estimates, it returns the path which minimizes the punishment of exploring unknown cells. This is translated to a path which minimizes the chance of entering the cliff. On the other hand, Q-learning uses a greedy target policy which does not acknowledge the drawbacks of exploration. Therefore, unlike SARSA, it manages to compute the true optimal policy.

As observed in Figure 2, the policy that Q-learning achieves makes the agent walk close to the cliff. Consequently, as Q-Learning acts according to an ε-greedy strategy, the agent falls off frequently. We believe this fact justifies the lower average total episode rewards obtained by Q-Learning.

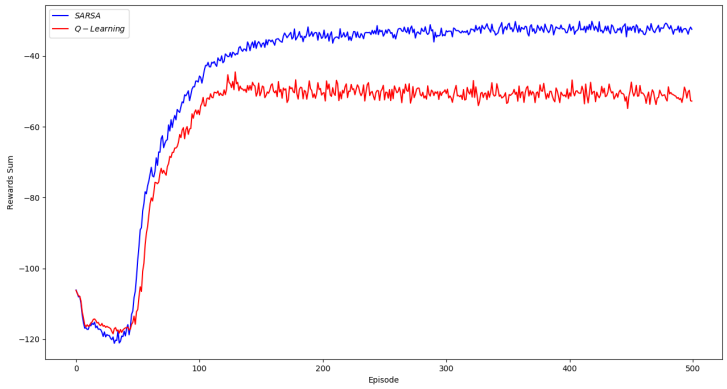

Figure 1: Chart comparing the average rewards sum obtained by SARSA and Q-Learning. The values were averaged over 1000 runs.

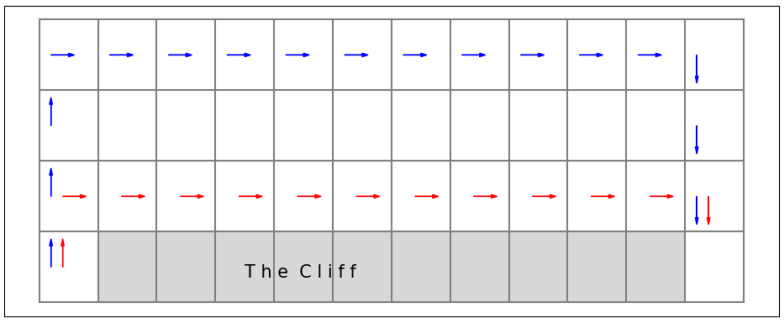

Figure 2: The final policies calculated by SARSA and Q-Learning. The SARSA trajectory is represented by the blue arrows and the Q-Learning trajectory by the red arrows. While SARSA prefers a safer path that does not punish exploration, Q-Learning finds the optimal path.

Solution for "TD learning with function approximation"

Figure 3 presents the norm of the weights vector for each step of Q-learning and SARSA averaged over 1000 runs. Q-learning uses a greedy target policy and so it tries to achieve an optimal policy. As the rewards of this problem are always the same (all of them are equal to 0), there is no policy better than any other. As such, Q-Learning increases the norm of the weight vector indefinitely and the estimates do not converge. SARSA, on the other hand, uses the same ε-greedy policy to update its weight vector and to act, and so it does not exhaustively look for an optimal policy. This makes the weight vector norms converge (to a value of ≈ 10). This convergence suggests that a stable policy to solve the problem was computed.

Although in tabular problems, there is the guarantee that both algorithms converge and find a policy (an optimal policy in the case of Q-Learning), this is not verified when using value approximation. The results observed for this exercise demonstrate this drawback. In fact, because SARSA is bound to find an ε-greedy optimal policy, it can still converge when applied to approximation problems such as this.

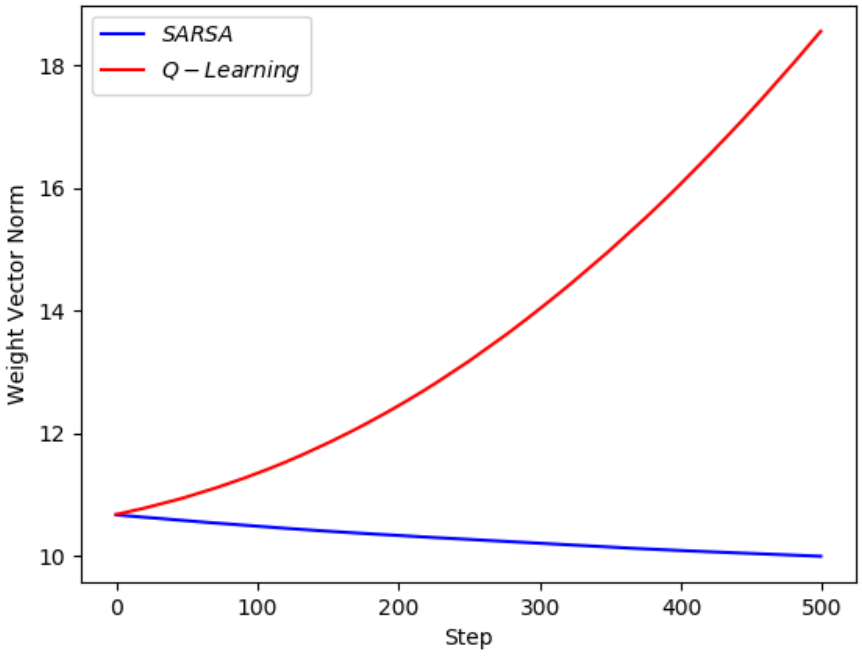

Figure 3: Chart comparing the weight vector norms of SARSA and Q-Learning for each step. The values were averaged over 1000 runs.

Solution for "The policy gradient theorem"

In order to show that $\nabla_\theta J(\theta) = \sum_{s\in S} \mu_\theta(s) \sum_{a\in A} \nabla_\theta \pi_\theta(a|s) q_{\pi_\theta}(s,a)$, we start by showing that for a fixed state $s \in S$: $\nabla_\theta v_{\pi_\theta}(s) = \sum_{s'\in S} \sum_{t=0}^{\infty} \gamma^t p_{\pi_\theta}^t(s'|s) \sum_{a\in A} \nabla_\theta \pi_\theta(a|s') q_{\pi_\theta}(s',a)$

Starting by the fact:
$v_{\pi_\theta}(s) = \sum_{a\in A} \pi_\theta(a|s) q_{\pi_\theta}(s,a)$
, we can deduct
$\nabla_\theta v_{\pi_\theta}(s) = \nabla_\theta[\sum_{a\in A}(\pi_\theta(a|s) q_{\pi_\theta}(s,a))] \leftrightarrow$
$\nabla_\theta v_{\pi_\theta}(s) = \sum_{a\in A}(\nabla_\theta \pi_\theta(a|s) q_{\pi_\theta}(s,a)) + \sum_{a\in A}(\pi_\theta(a|s) \nabla_\theta q_{\pi_\theta}(s,a)) \leftrightarrow$

(considering $\alpha(s,a) = \sum_{a\in A}(\nabla \pi_\theta(a|s) q_{\pi_\theta}(s,a))$ and $q_{\pi_\theta}(s,a) = r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a) v_{\pi_\theta}(s'))$

$\nabla_\theta v_{\pi_\theta}(s) = \alpha(s,a) + \sum_{a\in A}(\pi_\theta(a|s) \nabla_\theta q_{\pi_\theta}(s,a)) \leftrightarrow$
$\nabla_\theta v_{\pi_\theta}(s) = \alpha(s,a) + \sum_{a\in A}[\pi_\theta(a|s) \nabla_\theta (r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a) v_{\pi_\theta}(s'))] \leftrightarrow$
$\nabla_\theta v_{\pi_\theta}(s) = \alpha(s,a) + \sum_{a\in A}[\pi_\theta(a|s) \gamma \sum_{s'\in S} p(s'|s,a) \nabla_\theta v_{\pi_\theta}(s')] \leftrightarrow$

$\nabla_\theta v_{\pi_\theta}(s) = \alpha(s,a) + \sum_{a\in A}[\pi_\theta(a|s) \gamma \sum_{s'\in S} p(s'|s,a) [\alpha(s',a') + \sum_{a\in A}[\pi_\theta(a|s') \gamma \sum_{s''\in S} p(s''|s',a) \nabla_\theta v_{\pi_\theta}(s'')]] \leftrightarrow$

$\nabla_\theta v_{\pi_\theta}(s) = \alpha(s,a) + \sum_{s'\in S} \sum_{a\in A}[\gamma \pi_\theta(a|s) p(s'|s,a) [\alpha(s',a') + \sum_{s''\in S} \sum_{a\in A}[\gamma \pi_\theta(a|s') p(s''|s',a) \nabla_\theta v_{\pi_\theta}(s'')]] \leftrightarrow$

(considering $p_\pi^0(s'|s) = I[s = s']$ and $p_\pi^{t+1}(s'|s) = \sum_{s'\in A} \sum_{a\in A} p(s'|s'',a) \pi(a|s'') p_\pi^t(s'|s)$, we can rewrite $\nabla v_{\pi_\theta}(s)$)

$\nabla_\theta v_{\pi_\theta}(s) = \alpha(s,a) \gamma^0 p_\pi^0(s'|s) + \alpha(s',a') \gamma^1 p_\pi^1(s'|s) + \alpha(s'',a'') \gamma^2 p_\pi^2(s''|s) \nabla_\theta v_{\pi_\theta}(s'') \leftrightarrow$
$\nabla_\theta v_{\pi_\theta}(s) = \sum_{s'\in S} \sum_{t=0}^{\infty} [\gamma^t p_\pi^t(s'|s)] \alpha(s',a') \leftrightarrow$
$\nabla_\theta v_{\pi_\theta}(s) = \sum_{s'\in S} \sum_{t=0}^{\infty} [\gamma^t p_\pi^t(s'|s)] \sum_{a'\in A} \nabla_\theta \pi_\theta(a'|s') q_{\pi_\theta}(s',a')$

Considering the facts:
$J(\theta) = \sum_{s\in S} \mu_0(s) v_{\pi_\theta}(s)$
$\mu_\theta(s') = \sum_{s\in S} \mu_0(s) \sum_{t=0}^{\infty} [\gamma^t p_\pi^t(s'|s)]$
,

$\nabla_\theta J(\theta) = \nabla_\theta(\sum_{s\in S} \mu_0(s) v_{\pi_\theta}(s)) \leftrightarrow$
$\nabla_\theta J(\theta) = \sum_{s\in S} \mu_0(s) \nabla_\theta v_{\pi_\theta}(s) \leftrightarrow$
$\nabla_\theta J(\theta) = \sum_{s\in S} \mu_0(s) \sum_{s'\in S} \sum_{t=0}^{\infty} [\gamma^t p_\pi^t(s'|s)] \sum_{a'\in A} \nabla_\theta \pi_\theta(a'|s') q_{\pi_\theta}(s',a') \leftrightarrow$
$\nabla_\theta J(\theta) = \sum_{s'\in S} \sum_{s\in S} \mu_0(s) \sum_{t=0}^{\infty} [\gamma^t p_\pi^t(s'|s)] \sum_{a'\in A} \nabla_\theta \pi_\theta(a'|s') q_{\pi_\theta}(s',a') \leftrightarrow$
$\nabla_\theta J(\theta) = \sum_{s\in S} \sum_{s\in S} \mu_0(s) \sum_{t=0}^{\infty} [\gamma^t p_\pi^t(s'|s)] \sum_{a'\in A} \nabla_\theta \pi_\theta(a'|s') q_{\pi_\theta}(s',a') \leftrightarrow$
$\nabla_\theta J(\theta) = \sum_{s'\in S} \mu_\theta(s') \sum_{a'\in A} \nabla_\theta \pi_\theta(a'|s') q_{\pi_\theta}(s',a')$

Finally, substituting the names of the variables, we get the desired solution:

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

Solution for "The actor-critic architecture"

The following fact is presented in the description of the exercise:

$\nabla_\theta J(\theta) = \mathbb{E}_{S \approx \mu_\theta(.), A \approx \pi(.|S)}[\nabla_\theta log \pi_\theta(A|S) q_{\pi_\theta}(S, A)]$

As $\phi(s, a) = \nabla_\theta log \pi_\theta(a|s)$, we can rewrite the expression above:

$\nabla_\theta J(\theta) = \mathbb{E}_{S \approx \mu_\theta(.), A \approx \pi(.|S)}[\phi(S, A) q_{\pi_\theta}(S, A)]$

By multiplying and dividing by $\mathbf{\Phi}$, we get:

$\nabla_\theta J(\theta) = \mathbb{E}_{S \approx \mu_\theta(.), A \approx \pi(.|S)}[\phi(S, A) q_{\pi_\theta}(S, A) \mathbf{\Phi^{-1}} \mathbf{\Phi}]$

As $\mathbf{\Phi} = \mathbb{E}_{S \approx \mu_\theta(.), A \approx \pi(.|S)}[\phi(S, A) \phi^T(S, A)]$:

$\nabla_\theta J(\theta) = \mathbb{E}_{S \approx \mu_\theta(.), A \approx \pi(.|S)}[\phi(S, A) \phi^T(S, A) \mathbf{\Phi}^{-1} \mathbb{E}_{S \approx \mu_\theta(.), A \approx \pi(.|S)}[\phi(S, A) q_{\pi_\theta}(S, A)]]$

As $(\Gamma_\phi q_{\pi\theta})(S, A) = \phi^T(S, A) \mathbf{\Phi}^{-1} \mathbb{E}_{S \approx \mu_\theta(.), A \approx \pi(.|S)}[\phi(S, A) q_{\pi_\theta}(S, A)]$:

$\nabla_\theta J(\theta) = \mathbb{E}_{S \approx \mu_\theta(.), A \approx \pi(.|S)}[\phi(S, A)(\Gamma_\phi q_{\pi\theta})(S, A)]$,which is what we want to show.