Reinforcement Learning Homework 02

Qiu Yihang

March 2023

1 Convergence of Temporal Difference Value Learning

Proof. First we prove that $\{V_n\}$, where $V_n = \left(1 - \frac{1}{n^2}\right) V_{n-1} + \frac{1}{n^2} x_n$ is a Cauchy sequence.

For any $\varepsilon > 0$, consider $N = \left\lceil \frac{|C_1 - C_2|}{\varepsilon} \right\rceil$.

Without loss of generality, we assume $m \geq n$. Then we have

$$\begin{aligned}
V_m &= \left(1 - \frac{1}{m^2}\right) V_{m-1} + \frac{1}{m^2} x_m \\
&= \left(1 - \frac{1}{m^2}\right)\left(1 - \frac{1}{(m-1)^2}\right) V_{m-2} + \left(1 - \frac{1}{m^2}\right) \frac{1}{(m-1)^2} x_{m-1} + \frac{1}{m^2} x_m \\
&= \dots \\
&= \prod_{i=n+1}^{m} \left(1 - \frac{1}{i^2}\right) V_n + \frac{1}{m^2} x_m + \sum_{k=n+1}^{m-1} \left[\prod_{i=k+1}^{m} \left(1 - \frac{1}{i^2}\right)\right] \frac{1}{k^2} x_k \\
|V_m - V_n| &= \left| \left[\prod_{i=n+1}^{m} \left(1 - \frac{1}{i^2}\right) - 1\right] V_n + \frac{1}{m^2} x_m + \sum_{k=n+1}^{m-1} \left[\prod_{i=k+1}^{m} \left(1 - \frac{1}{i^2}\right)\right] \frac{1}{k^2} x_k \right| \\
&\leq \left| \left[\prod_{i=n+1}^{m} \left(1 - \frac{1}{i^2}\right) - 1\right] |V_n| + \frac{1}{m^2} |x_m| + \sum_{k=n+1}^{m-1} \left[\prod_{i=k+1}^{m} \left(1 - \frac{1}{i^2}\right)\right] \frac{1}{k^2} |x_k| \right| \\
&\leq \left| \left[\prod_{i=n+1}^{m} \left(1 - \frac{1}{i^2}\right) - 1\right] C_2 + \left\{ \frac{1}{m^2} + \sum_{k=n+1}^{m-1} \left[\prod_{i=k+1}^{m} \left(1 - \frac{1}{i^2}\right)\right] \frac{1}{k^2} \right\} C_1 \right| \\
&= \left| \left[\prod_{i=n+1}^{m} \left(1 - \frac{1}{i^2}\right) - 1\right] C_2 + \left\{ 1 - \left(1 - \frac{1}{m^2}\right) + \sum_{k=n+1}^{m-1} \left[\prod_{i=k+1}^{m} \left(1 - \frac{1}{i^2}\right)\right] \frac{1}{k^2} \right\} C_1 \right| \\
&= \dots \\
&= \left| \left[\prod_{i=n+1}^{m} \left(1 - \frac{1}{i^2}\right) - 1\right] C_2 + \left[1 - \prod_{i=n+1}^{m} \left(1 - \frac{1}{i^2}\right)\right] C_1 \right| \\
&= \left| \left[1 - \prod_{i=n+1}^{m} \left(1 - \frac{1}{i^2}\right)\right] (C_1 - C_2) \right| = \left| \left(1 - \frac{(m+1)n}{m(n+1)}\right) (C_1 - C_2) \right|
\end{aligned}$$

$$
\begin{aligned}
&< \left(1 - \frac{n}{n+1}\right) |C_1 - C_2| = \frac{1}{n+1} |C_1 - C_2| \\
&< \frac{1}{N} |C_1 - C_2| \leq \frac{\varepsilon}{|C_1 - C_2|} |C_1 - C_2| = \varepsilon
\end{aligned}
$$

i.e. for any $\varepsilon > 0$, exists $N = \left\lceil \frac{|C_1 - C_2|}{\varepsilon} \right\rceil$ such that $|V_n - V_m| < \varepsilon$ for any $n, m \geq N$. Thus, $\{V_n\}$ is a Cauchy sequence.

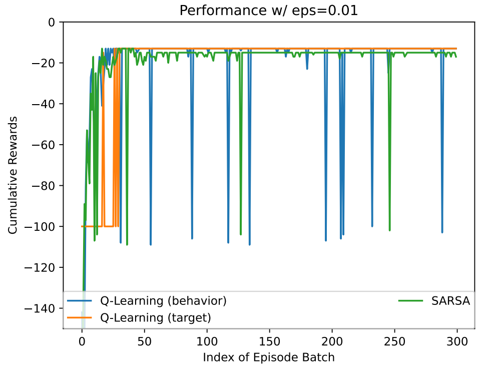
Now we prove that TD value learning with $\alpha_n = \frac{1}{n}$ will converge.

Let $\varepsilon \to 0$. Since $\{V_n\}$ is a Cauchy sequence, $\lim_{k\to\infty} V_k = V^*$.

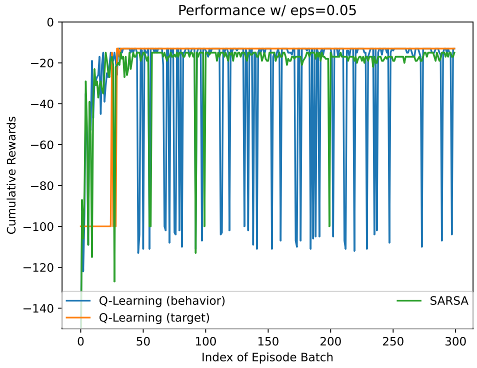
Therefore, TD value learning with $\alpha_n = \frac{1}{n}$ will converge. ■

2 Implementation of the SARSA and Q-learning algorithms

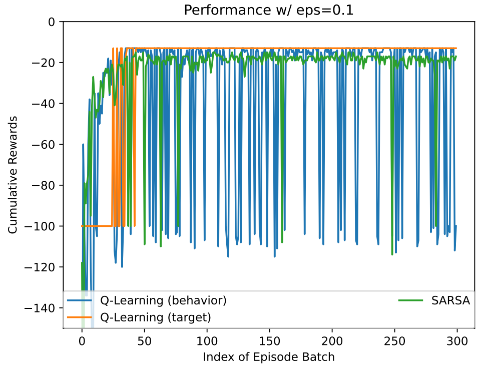
Solution. Under different values of ε in ε-greedy, the performances of SARSA, Q-learning algorithm are plotted in green and blue respectively as follows. Also, the performances of target policy in Q-learning are depicted in orange.



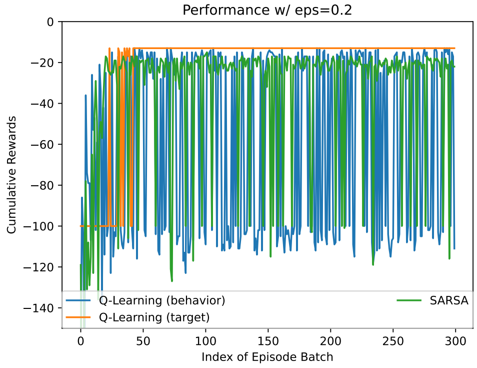
(a) $\varepsilon = 0.01$



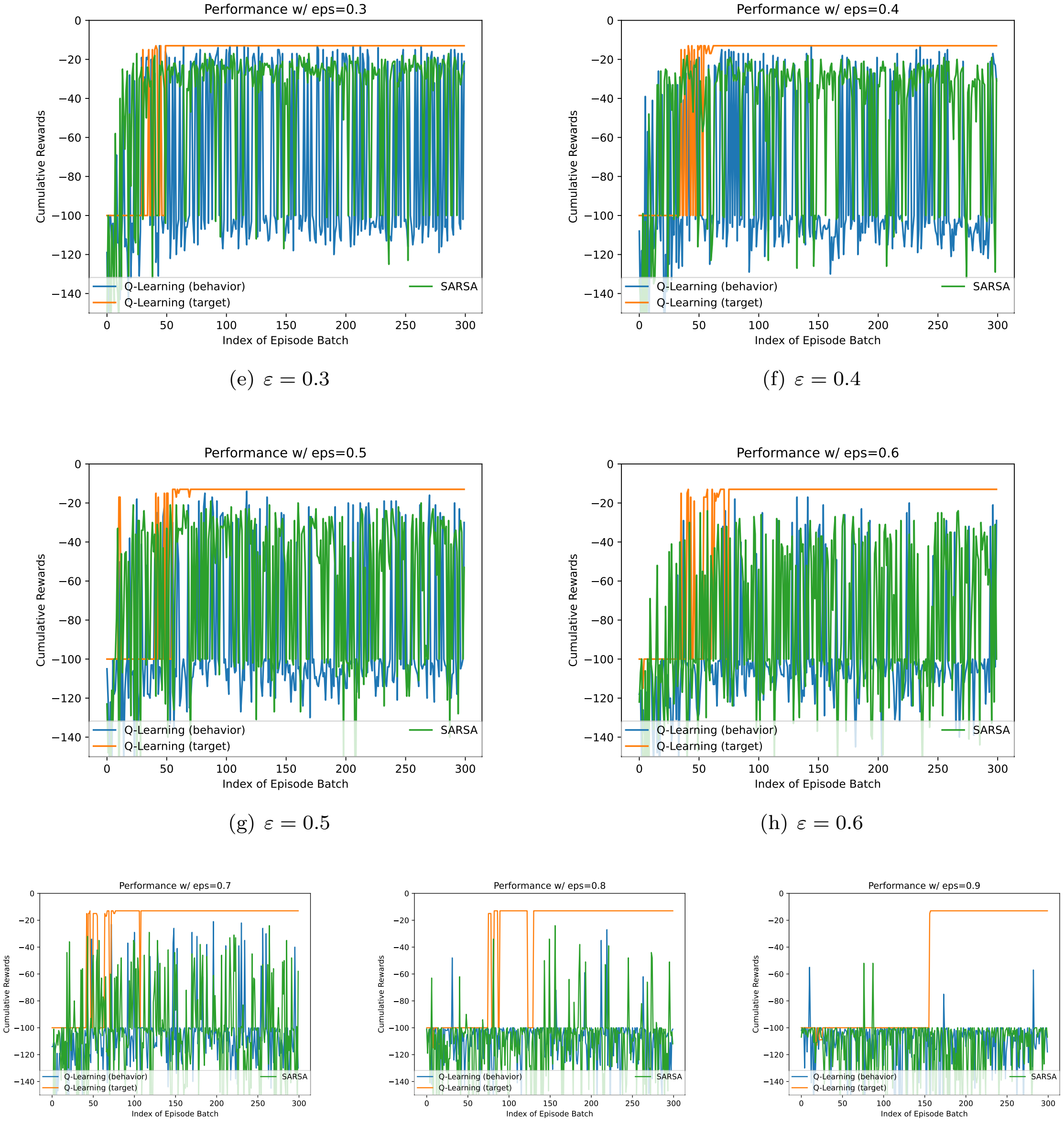
(b) $\varepsilon = 0.05$



(c) $\varepsilon = 0.1$



(d) $\varepsilon = 0.2$



(e) $\varepsilon = 0.3$

(f) $\varepsilon = 0.4$

(g) $\varepsilon = 0.5$

(h) $\varepsilon = 0.6$

(i) $\varepsilon = 0.7$

(j) $\varepsilon = 0.8$

(k) $\varepsilon = 0.9$

Figure 1: Performance of SARSA and Q-learning algorithms (behaviour policy and target policy) under different values of ε in ε-greedy.

2.1 Comparisons of Different Policies

We know that in theory, the best cumulative reward is -12.

For **SARSA**, the best performance under all ε is lower than that of **Q-learning**, especially in the case when $\varepsilon = 0.01, 0.05, 0.1, 0.2, 0.3$.

Meanwhile, when $\varepsilon < 0.2$, **Q-learning** can maintain the best performance in most time.

2.2 Impacts of Different Values of ε

We know when ε is smaller, the ε-greedy tends to be conservative and maintains the optimal policy at the time. When ε is larger, the ε-greedy tends to be more bold and explores more actions.

This is also reflected in the results. When ε is smaller, the cumulative rewards of both SARSA and Q-learning behaviour policy appear more stable. When ε is larger, the cumulative rewards of both SARSA and Q-learning behaviour policy become more fluctuated.

Moreover, when ε is too large, $\varepsilon = 0.7, 0.8, 0.9$ for example, the cumulative rewards of both SARSA and Q-learning behaviour policy tend to be much lower than that when ε is smaller. In other words, when ε is too large, the ε-greedy tends to be too bold and explores too much actions instead of optimizing a stable policy.

Meanwhile, the target policy of Q-learning converges to an optimal policy slower as ε becomes larger.

2.3 Differences Between Behaviour Policy and Target Policy of Q-Learning

Compared with target policy, behaviour policy explores more and tends to be fluctuated. It is plain to see from the results that the cumulative reward of target policy is more stable than that of behaviour policy.

■