RL Assignment2 Report

520030910342 Jiyu Liu

1. Convergence of temporal difference value learning

To prove that $\{V_n\}$ is a Cauchy sequence, we need to show that for any given $\epsilon > 0$, there exists a positive integer N, such that for any $m > n > N$, we have $|V_m - V_n| < \epsilon$.

Fisrt of all, consider the value of $|V_n - V_{n-1}|$

$$|V_n - V_{n-1}| = |\alpha_n(x_n - V_{n-1})| \leq |\alpha_n| \cdot (C_1 + C_2)$$

Denoting $C_1 + C_2$ as C_3, we can infer that:

$$\begin{aligned} |V_m - V_n| &= |\sum_{i=n+1}^{m} (V_i - V_{i-1})| \\ &\leq \sum_{i=n+1}^{m} |V_i - V_{i-1}| \\ &\leq \sum_{i=n+1}^{m} \frac{C_3}{i^2} \\ &\leq \sum_{i=N+2}^{\infty} \frac{C_3}{i^2} \end{aligned}$$

Because that $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$, as the value of N increases, $\sum_{\infty}^{i=N+2}$ will converge to 0. Thus, we can indicate that for any $\epsilon > 0$, there exists a positive integer N, such that for any $m > n > N$, we have $|V_m - V_n| < \epsilon$.

2. Implementation of the SARSA and Q-learning algorithms

2.1. coding

The code and its corresponding results, including the performance of SARSA algorithm and Q-learning algorithm using or not using the target policy with different valus of ϵ in $\epsilon - greedy$, can be found in the folder "code".

2.2. discussion

(a)When the value of ϵ is very small(e.g. 0.01, 0.05), the algorithms mainly exploit rather than explore, the performance of the algorithms are steady and converge relatively well.

As the value of ϵ becomes larger, the agent begins to do more exploration, and the performance of SARSA and Q-learning using the behavor policy becomes more and more unsteady. The cumulative reward of Q-learning with the behavor policy no longer converges as ϵ reaches 0.1 and SARSA cannot converge as ϵ reaches 0.5. Nonetheless, Q-learning algorithm with the target policy keeps convergence even though ϵ is raised to 0.9.

(b)The performance of the target policy is usually more optical than the behavor policy, especially when the value of ϵ gets very large, the reason of which might be that the target policy is greedy and always picks the action with highest Q-value.