强化学习 HW2

— 饶翔云 520030910366

Problem 1

Reformalize: $V_n = (1-\alpha_n)V_{n-1} + \alpha_n x_n$, where x_n is a sample of value, and $|x_n| \le C_1, |V_n| \le C_2$. Please prove $\{V_n\}$ coverges.

I suppose to use Cachy coverges rule:
An array coverges $\iff \forall \epsilon > 0, \exists N \in \mathbb{N}, \forall m > N, \forall n > N, |x_m - x_n| \le \epsilon$

Proof: $\forall \epsilon$, when n closed to ∞

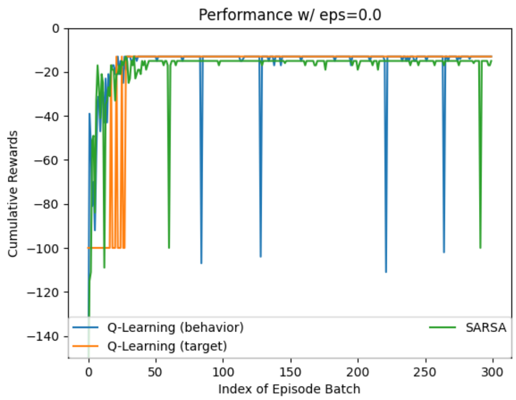
$\because |V_m| = |\prod_{i=n+1}^{m}(1-\alpha_i)V_n + \sum_{i=n+1}^{m}\prod_{i=n+1}^{m}(1-\alpha_i)\alpha_i x_i|$

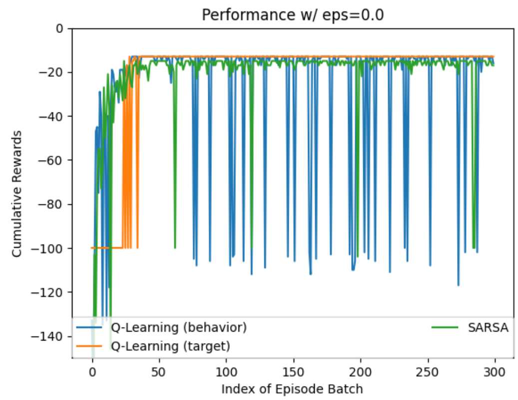
$\therefore |V_m - V_n| = |\prod_{i=n+1}^{m}(1-\alpha_i)V_n + \sum_{i=n+1}^{m}\prod_{i=n+1}^{m}(1-\alpha_i)\alpha_i x_i - V_n|$
$= |(\prod_{i=n+1}^{m}(1-\alpha_i) - 1)V_n + \sum_{i=n+1}^{m}(\frac{n}{m(n+1)})\alpha_i x_i|$
$\le |(\prod_{i=n+1}^{m}(1-\alpha_i) - 1)V_n| + |\sum_{i=n+1}^{m}(\frac{n}{m(n+1)})\alpha_i x_i|$
$\le |((1-\frac{1}{N^2}) - 1)V_n| + |(\frac{n(m-n)}{m(n+1)N^2})x_i|$
$\le |\frac{C_2}{N^2}| + |\frac{C_1}{N^2}|(\sum_{i=1}^{\infty}\alpha_i \text{ coverges to } \frac{\pi^2}{6})$
$\because |x_m - x_n| \le \epsilon, \forall m > N, \forall n > N$
$\therefore \forall \epsilon, \exists N \ge \sqrt{\frac{C_1+C_2}{\epsilon}}, \forall m > N, \forall n > N, |x_m - x_n| \le \epsilon.$

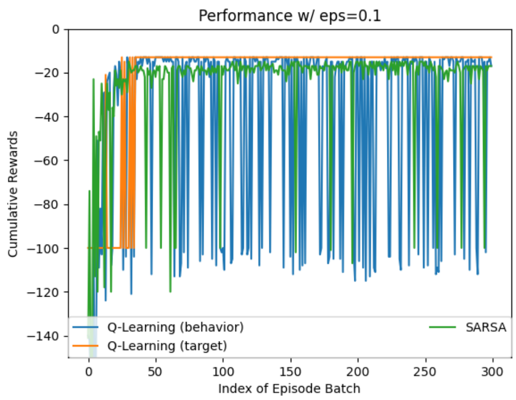
Therefore, TD-learning coverges

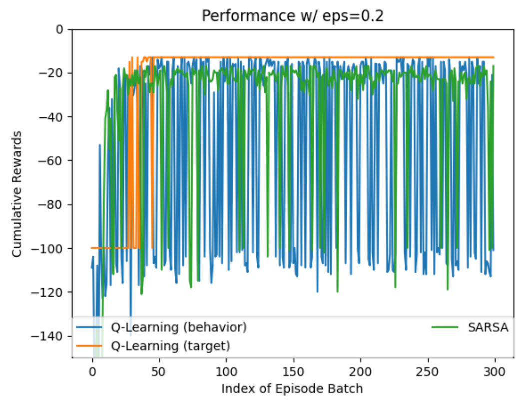
Problem 2

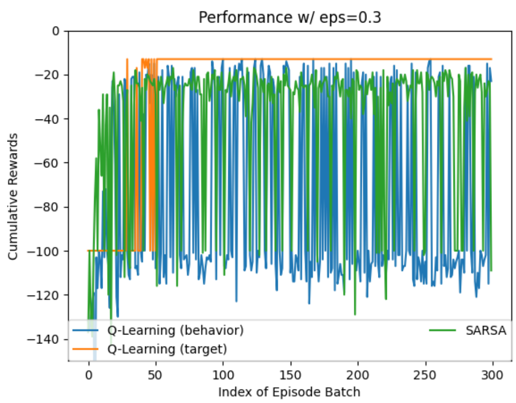
以下是从0.01到0.9的ϵ得到的累计价值随批次的变化曲线：

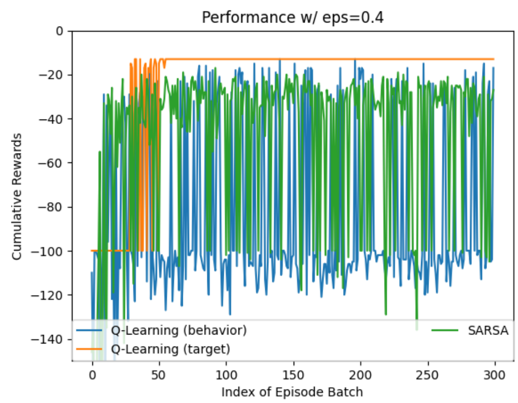

Performance w/ eps=0.0
Cumulative Rewards
Index of Episode Batch
Q-Learning (behavior)
Q-Learning (target)
SARSA

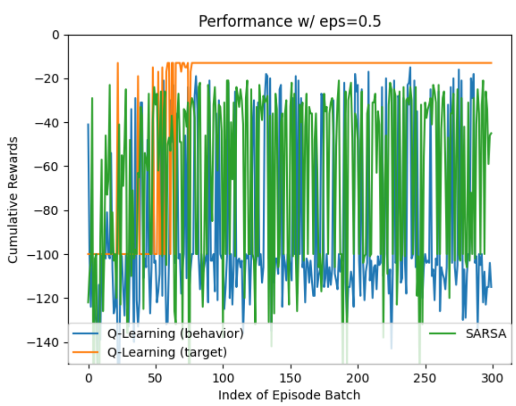

Performance w/ eps=0.0
Cumulative Rewards
Index of Episode Batch
Q-Learning (behavior)
Q-Learning (target)
SARSA

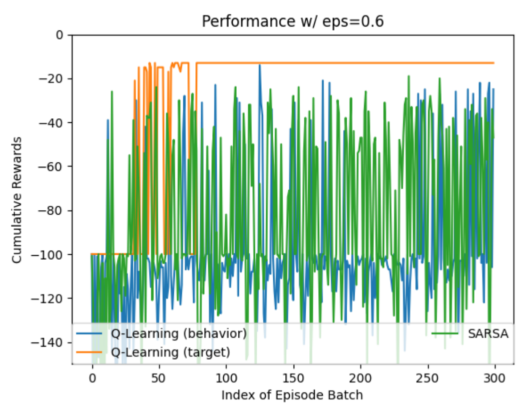

Performance w/ eps=0.1
Cumulative Rewards
Index of Episode Batch
Q-Learning (behavior)
Q-Learning (target)
SARSA

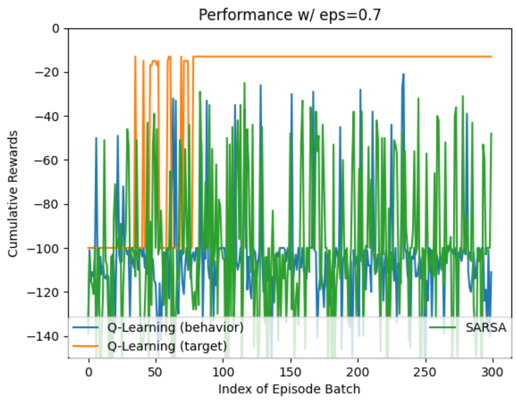

Performance w/ eps=0.2
Cumulative Rewards
Index of Episode Batch
Q-Learning (behavior)
Q-Learning (target)
SARSA

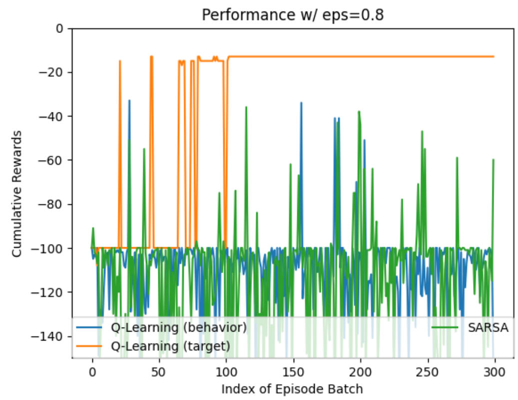

Performance w/ eps=0.3
Cumulative Rewards
Index of Episode Batch
Q-Learning (behavior)
Q-Learning (target)
SARSA

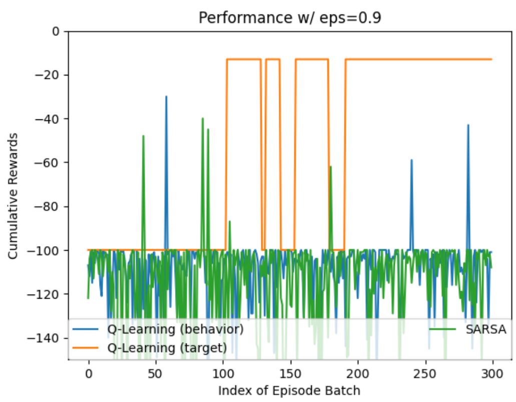

Performance w/ eps=0.4
Cumulative Rewards
Index of Episode Batch
Q-Learning (behavior)
Q-Learning (target)
SARSA


Performance w/ eps=0.5
Cumulative Rewards
Index of Episode Batch
Q-Learning (behavior)
Q-Learning (target)
SARSA


Performance w/ eps=0.6
Cumulative Rewards
Index of Episode Batch
Q-Learning (behavior)
Q-Learning (target)
SARSA







(a) what are the impacts of different values of ϵ on the performance of the above three algorithms?

> With low ϵ, here is less uncertainty and thus we have a stable curve. And with the ϵ increasingm, there is more uncertainty and curve become more unstable, that is, more and more closed to random walk in this cliff-walking environment.

(b) what is the difference between the performance of the behavior policy of Q-learning algorithm and the performance of the target policy of Q-learning algorithm?

ϵ	**Q-learning(behaviour)**	**Q-learning(target)**	**Sarsa**
low	Nearly the same because it is closed to full greedy and that is a certain policy	Nearly the same because it is closed to full greedy and that is a certain policy	Nearly the same because it is closed to full greedy and that is a certain policy
high	Performs poor for extraordinary unstable	Performs poor but better than others because it use target policy with the same Q-table as behaviour policy but execute full greedy.	Performs poor for extraordinary unstable