中山大学计算机院本科生实验报告
（2024学年秋季学期）

课程名称：强化学习与博弈论　　批改人：

实验	Assignment1	专业（方向）	计算机科学与技术计科一班
学号	21307099	姓名	李英骏
Email	liyj323@mail2.sysu.edu.cn	完成日期	2024 年 11 月 19 日

目录

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

1 实验背景

Cliff Walk 是强化学习中的经典环境，用于测试算法在高风险场景下的表现。该环境由一个 4×12 的网格组成，起点 S 位于左下角，终点 G 位于右下角。起点和终点之间的悬崖区域会导致代理掉入后受到 -100 的惩罚，并返回到起点。每一步移动都会收到 -1 的奖励。

在该实验中，我们分别使用 SARSA 和 Q-learning 算法训练代理学习最优路径，并对比两种算法在路径选择和奖励表现上的差异。

2 算法核心思路

SARSA 和 Q-learning 都是基于 Q-learning 的值迭代算法。其核心在于通过更新 Q 值来指导代理选择动作。具体可见代码中的注释，如下：

2.1 SARSA 算法

SARSA 算法的更新公式如下：

$$Q(S, A) \leftarrow Q(S, A) + \alpha \cdot [R + \gamma \cdot Q(S', A') - Q(S, A)]$$

其中：

- S, A：当前状态和动作；
- R：即时奖励；
- S', A'：下一状态和实际选择的动作；
- α：学习率；
- γ：折扣因子。

2.2 Q-learning 算法

Q-learning 算法的更新公式如下：

$$Q(S, A) \leftarrow Q(S, A) + \alpha \cdot \left[R + \gamma \cdot \max_a Q(S', a) - Q(S, A)\right]$$

其中：

- S, A：当前状态和动作；
- R：即时奖励；
- S'：下一状态；
- $\max_a Q(S', a)$：下一状态下的最大 Q 值；

- α：学习率；
- γ：折扣因子。

3 实验设计

3.1 实验环境

实验环境为 Cliff Walk，网格大小为 4×12，起点和终点分别位于左下角和右下角。悬崖区域位于底部。

3.2 实验参数

实验参数设置如下：

- 学习率 $\alpha = 0.5$；
- 折扣因子 $\gamma = 0.9$；
- 探索率 $\epsilon = 0.1$；
- 最大训练回合数：5000。

3.3 实验目标

通过 SARSA 和 Q-learning 算法训练代理，比较以下两点：

1. 两种算法生成的行走路径；
2. 两种算法的奖励曲线变化趋势。

4 实验结果与分析

4.1 SARSA 算法路径

SARSA 算法的路径如图 1 所示。

SARSA 算法的路径明显避开了悬崖区域，选择了更保守的路线。agnet通过顶部区域向右移动，再从顶部到终点。路径表明 SARSA 倾向于学习更安全的策略。

4.2 Q-learning 算法路径

Q-learning 算法的路径如图 2 所示。

Q-learning 算法中agent直接沿着悬崖上方移动，选择了一条更短的路径到达终点。这表明 Q-learning 倾向于冒险以追求最优路径。

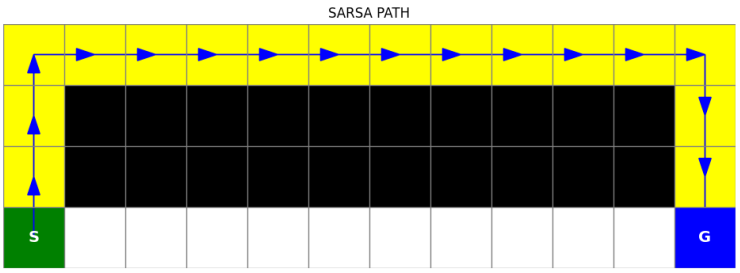

图 1: SARSA 算法路径

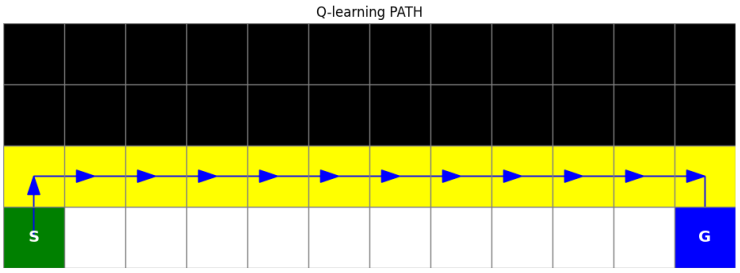

图 2: Q-learning 算法路径

4.3 奖励曲线分析

每回合的总奖励变化如图 3 所示，下面那张图 4 是平滑后的。

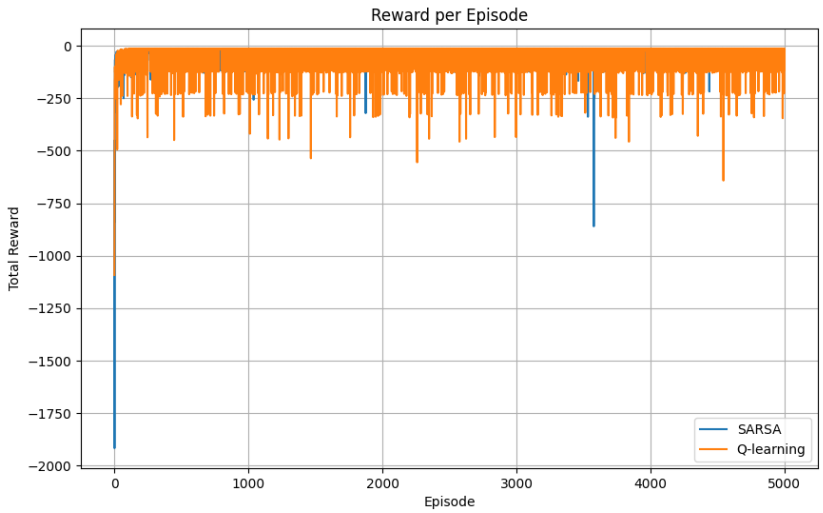

图 3: 每回合奖励曲线

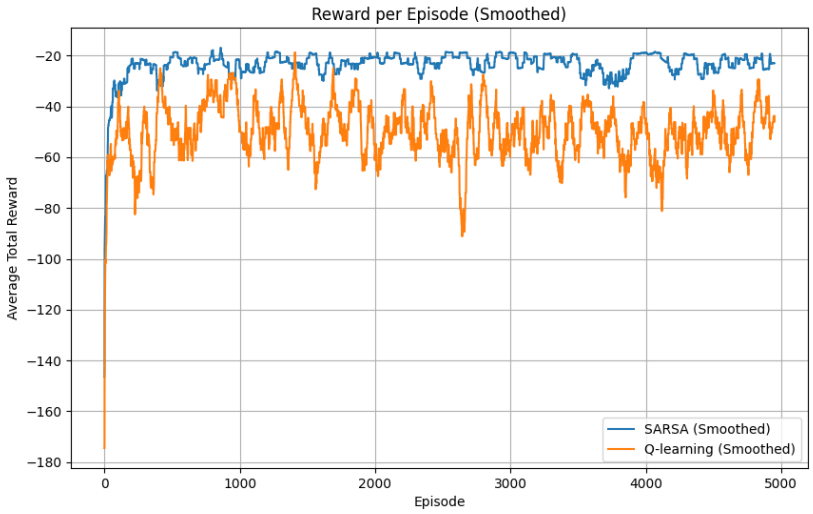

图 4: 每回合奖励曲线(smoothed)

从奖励曲线可以看出：

- SARSA 的奖励曲线收敛更快，波动较小。
- Q-learning 的奖励曲线初始阶段波动较大，最终逐渐收敛，但仍有较大的波动。

5 结论

通过实验结果分析，我们可以得出以下结论：

- SARSA 更适合高风险场景，其更新基于当前策略，倾向于学习保守路径。
- Q-learning 更适合追求全局最优解的场景，其更新基于最优动作值，但可能带来更高的风险。

参考

1. OpenAI Gym Documentation: `https://www.gymlibrary.dev/`