Windy Gridworld

AAIS in PKU　陈伟杰 1901111420

January 7, 2020

1 Problem Setting

设置一个 10×7 的有终点行走的网格空间 $\mathbb{S}$, 行动空间 $\mathbb{A} = \{\uparrow, \downarrow, \leftarrow, \rightarrow\}$，终点为 $s_g = (7, 3)$，起点为 $s_0 = (0, 3)$ 定义行走规则和边界条件如下：

$$\begin{aligned} (s, a) &\to (s', r = -1) \qquad \text{if } s' \neq s_g \\ (s, a) &\to (s', r = 0) \qquad \text{if } s' = s_g \\ (s, a) &\to (s, r = -1) \qquad \text{if } s' \notin \mathbb{S} \end{aligned} \tag{1}$$

其中设置有风区 Ω_1 和 Ω_2（图中淡蓝色和深蓝色区域），分别吹风力为 $\omega_1 = 1$ 和 $\omega_2 = 2$ 的向上的风，定义如下：

$$\begin{aligned} (s = (x, y) \in \Omega_1, a = (a_x, a_y)) &\to (s' = (x + a_x, y + a_y + \omega_1), r = -1) \\ (s = (x, y) \in \Omega_2, a = (a_x, a_y)) &\to (s' = (x + a_x, y + a_y + \omega_2), r = -1) \end{aligned} \tag{2}$$

此外，附加设置了随机风，在有风区 Ω_1 和 Ω_2 中，各有 1/3 概率使风力偏离 ± 1。即 $\omega_1 \in \{0, 1, 2\}$，$\omega_2 \in \{1, 2, 3\}$。

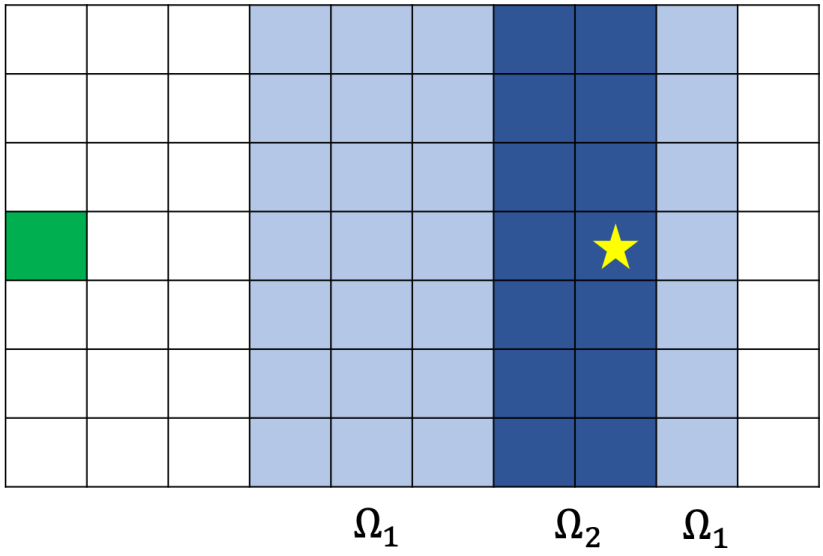

Figure 1: 带风格子世界示意图，绿色格子为起点，星标为终点，Ω_1 和 Ω_2 为有风区。

2 Original Windy Gridworld

首先解决无随机风的网格世界问题，在实验中分别采用 SARSA 和 Q-learning 算法，其中参数设置 $\gamma = 1$，$\alpha = 0.2$，$\epsilon = 0.1$。

2.1 SARSA

SARSA 采用 ϵ-贪婪法进行动作决策，在值函数迭代时值函数 $Q(s', a')$ 的 a' 也由 ϵ-贪婪法进行选择，且 a' 为下一步的实际行动。

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a') - Q(s, a)] \tag{3}$$

对应的最优策略表格 [3]，其中标出了到达终点的最优路径。在实际数值实验中，偏离最优路径的区域的动作选

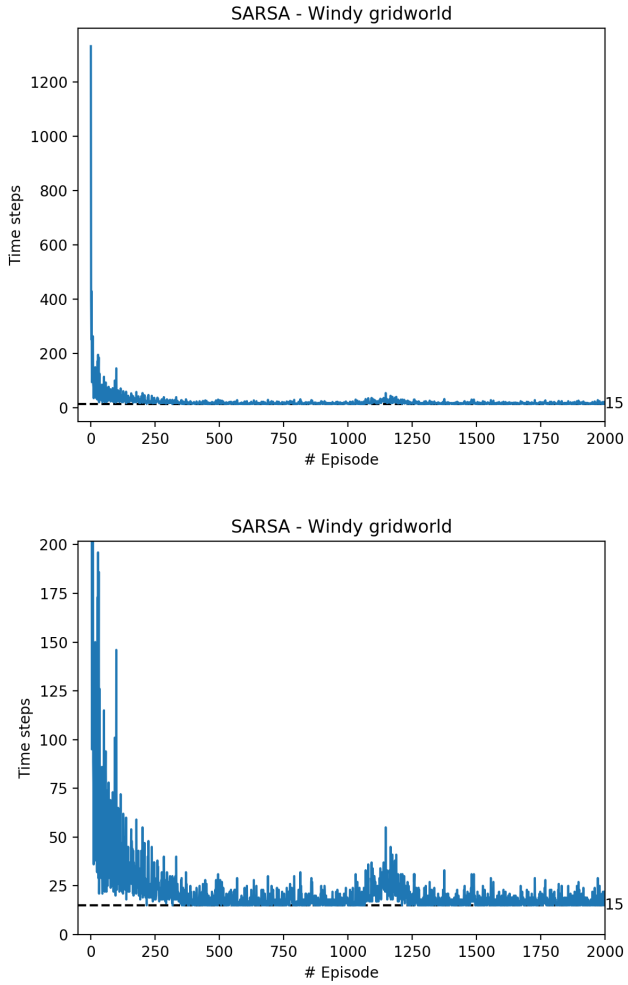

Figure 2: Time steps-Episode 关系曲线（无随机风情况下 SARSA 算法）

择具有任意性，因此某些格子的最优策略在重复实验时可能会不同。

2.2 Q-learning

Q-learning 采用 ϵ-贪婪法进行动作决策，在值函数迭代时对于 s' 选择最优的值函数，即 $\max_a Q(s', a)$。

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_A Q(s', A) - Q(s, a)] \tag{4}$$

对应的最优策略表格为 [5], 大体与 SARSA 的结果一致，少数格子由于随机性有不同。

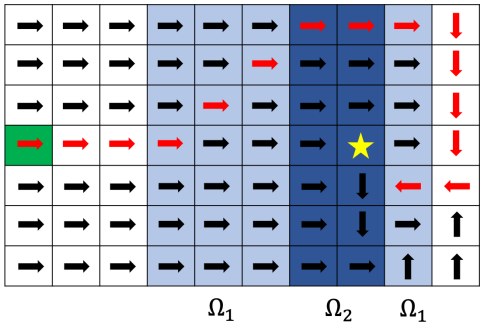

Figure 3: 无随机风 SARSA 算法对应的最优策略，其中红色箭头表示最优路径 (Episodes = 5e5)

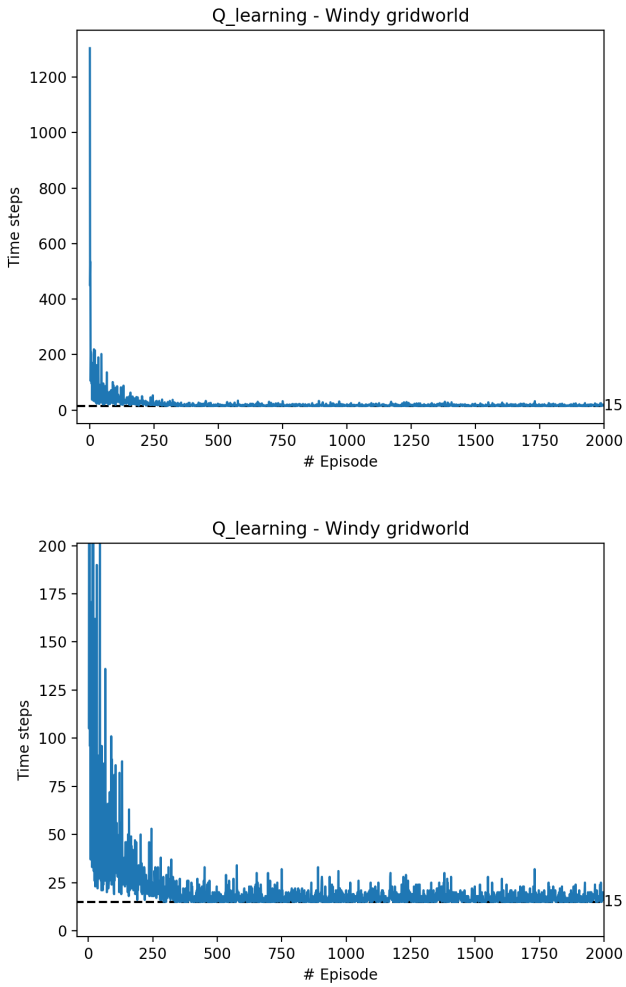

Figure 4: Time steps-Episode 关系曲线（无随机风情况下 Q-learning 算法）

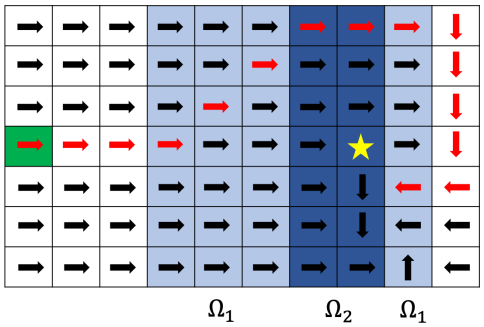

Figure 5: 无随机风 Q-learning 算法对应的最优策略，其中红色箭头表示最优路径 (Episodes = 5e5)

3 Stochastic Windy Gridworld

将原本的恒定有风区改为随机风力的有风区，重新使用 SARSA 和 Q-learning 算法解决该问题。参数设定与前述相同。

3.1 SARSA

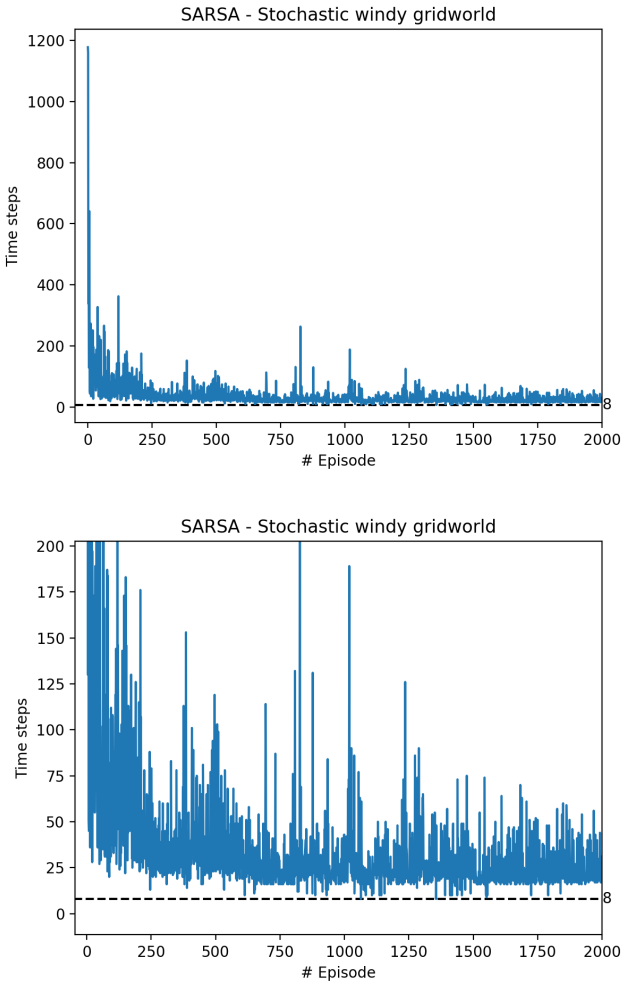

Figure 6: Time steps-Episode 关系曲线（随机风情况下 SARSA 算法）

相比于无随机风，可以明显看出到达终点所需的步数的 variation 变大，并且最短路径从 15 步降到 8 步。同时可以看出最短路径相比于无随机风更难达到，同样 2000 个 episode 内出现频率明显下降。

3.2 Q-learning

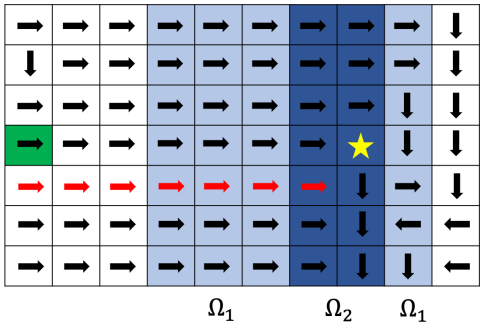

Figure 7: 有随机风 SARSA 算法对应的最优策略，其中红色箭头表示可能的最优路径 (Episodes = 5e5)

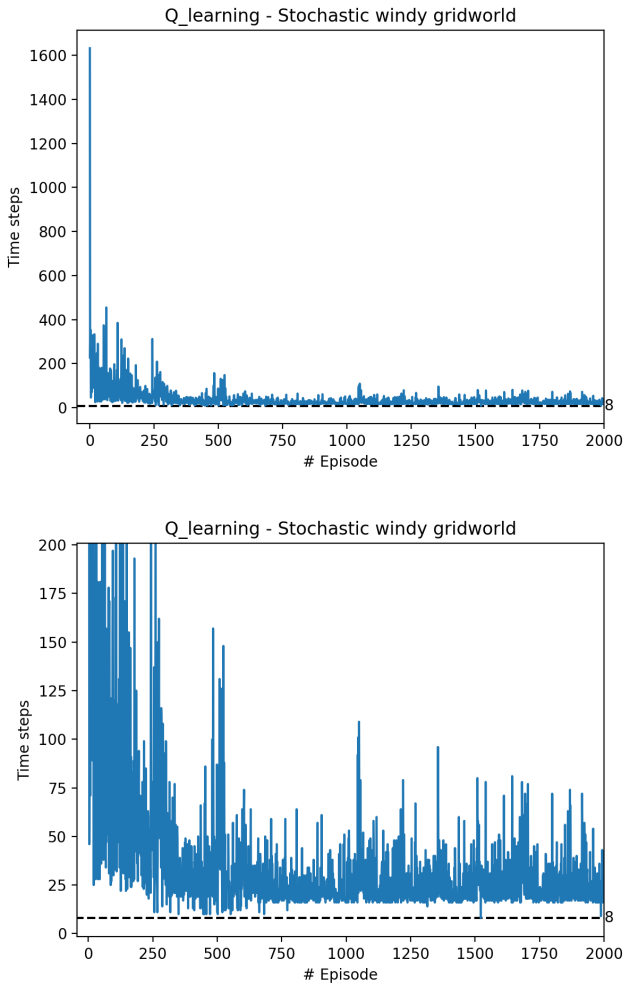

Figure 8: Time steps-Episode 关系曲线（随机风情况下 Q-learning 算法）

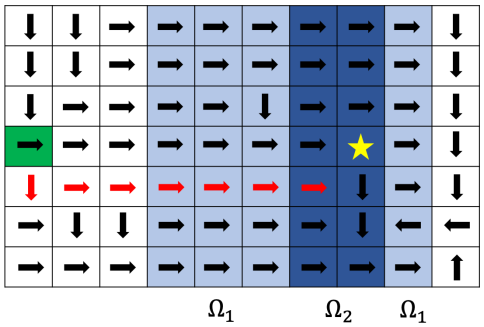

Figure 9: 有随机风 Q-learning 算法对应的最优策略，其中红色箭头表示可能的最优路径 (Episodes = 5e5)

4 Discussion

在有风网格世界问题中，SARSA 和 Q-learning 的表现相差不大，这是因为该问题对于非最优路径的罚项不足，与课程中展示的悬崖问题不同。由于在实验中，Q-learning 和 SARSA 采用相同的 ϵ 贪婪采样策略，仅在目标更新有概率为 0.1 的不同。若非最优路径罚项极大，例如掉落悬崖，则 0.1 的采样更新对目标函数的值会影响很大，使得整体策略趋于保守。若非最优路径罚项较小甚至与最优路径区别不大，则 0.1 的采样更新对目标函数的值影响不显著，使得两个算法结果趋同。

对于加入随机风的情况，由于环境的随机性增加了，给值函数更新带来噪声，使得学习曲线的 variation 变大。另一方面，最优路径的达成率除了依赖于 ϵ 的大小，还依赖于环境的随机性，使得最优路径的达成率明显下降。

5 README

代码包括 WindyGrid.py 和 test.py 两个程序，需要 python3 环境。直接运行 test.py 可以得到上述结果。