RL_exp1 实验报告

1. 实验目标

实现 MC 算法、Sarsa 算法和 Q-learning 算法

2. 实验方法

MC 算法：

状态 S_t 是的 Action 使用 ε-贪婪的策略，更新目标为：

$$
\begin{aligned}
q_\pi(s,\pi'(s)) &= \sum_a \pi'(s|a) q_\pi(s,a) \\
&= \frac{\epsilon}{\mathcal{A}(s)} + (1-\epsilon)\max_a q_\pi(s,a) \\
&\geq \frac{\epsilon}{\mathcal{A}(s)} + (1-\epsilon)\sum_a \frac{\pi(a|s) - \frac{\epsilon}{\mathcal{A}(s)}}{1-\epsilon} q_\pi(s,a) \\
&= \frac{\epsilon}{\mathcal{A}(s)} - \frac{\epsilon}{\mathcal{A}(s)} + \sum_a \pi(s|a) q_\pi(s,a) \\
&= v_\pi(s)
\end{aligned}
$$

伪代码：

初始化，$\forall s \in \mathcal{S}, a \in \mathcal{A}(s)$

$Q(s,a) \leftarrow$ 任意值

$\pi(s) \leftarrow$ 任意值

$Returns(s,a) \leftarrow$ 空 list

Repeate Forever:

(a) 使用策略π来生成episode

(b) For each (s,a) in episode:

$G \leftarrow (s,a)$第一次出现的 Return

把G加到Returns(s,a)里

$Q(s,a) \leftarrow average(Returns(s,a))$

(c) For s in episode:

$A = arg\max_a Q(s,a)$

For all $a \in \mathcal{A}(s)$:

$$
\pi(a|s) \leftarrow \begin{cases} 1-\epsilon+\frac{\epsilon}{|\mathcal{A}(s)|} & if\ a = A \\ \frac{\epsilon}{|\mathcal{A}(s)|} & if\ a \neq A \end{cases}
$$

Sarsa 算法：

状态 S_t 是的 Action 使用 ε-贪婪的策略，更新目标为：

$$
\begin{aligned}
&R_{t+1}+\gamma V(S_{t+1},A')-Q(S_t,A_t)\\
&=R_{t+1}+\gamma V(S_{t+1},arg\max_{a'}Q(S_{t+1},a'))-Q(S_t,A_t)\\
&=R_{t+1}+\gamma\max_{a}V(S_{t+1},a)-Q(S_t,A_t)
\end{aligned}
$$

伪代码：

$\forall s\in\mathcal{S},a\in\mathcal{A}(s)$，随机初始化$Q(s,a)$，初始化 Q(终止状态,) 为 0

Repeat

初始化S

使用Q得到的 -贪婪的策略，并根据它选择A

Repeat

采取行为A，得到R和S'

使用策略采取行为A'

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

$A\leftarrow A',S\leftarrow S'$

Until S是终止状态

Q-learning 算法：

状态 S_t 是的 Action 使用 ε-贪婪的策略，更新目标为：

$$
\begin{aligned}
&R_{t+1}+\gamma V(S_{t+1},A')-Q(S_t,A_t)\\
&=R_{t+1}+\gamma V(S_{t+1},arg\max_{a'}Q(S_{t+1},a'))-Q(S_t,A_t)\\
&=R_{t+1}+\gamma\max_{a}V(S_{t+1},a)-Q(S_t,A_t)
\end{aligned}
$$

伪代码：

$\forall s\in\mathcal{S},a\in\mathcal{A}(s)$，随机初始化$Q(s,a)$，初始化 Q(终止状态,) 为 0

Repeat

初始化S

Repeat

使用Q得到的 -贪婪的策略，并根据它选择A

采取行为A，得到R和S'

使用策略采取行为A'

$Q(S,A)\leftarrow Q(S,A)+\alpha(R+\gamma max_aQ(S',a)-Q(S,A))$

$S\leftarrow S'$

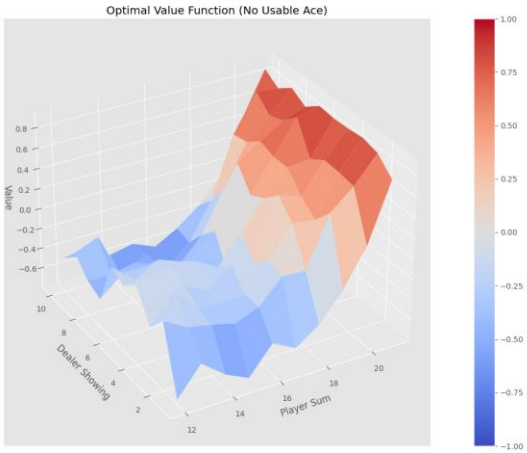
Until S是终止状态

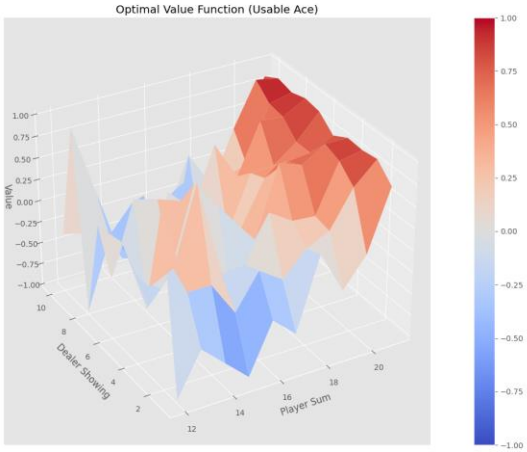
---

3. 实验结果

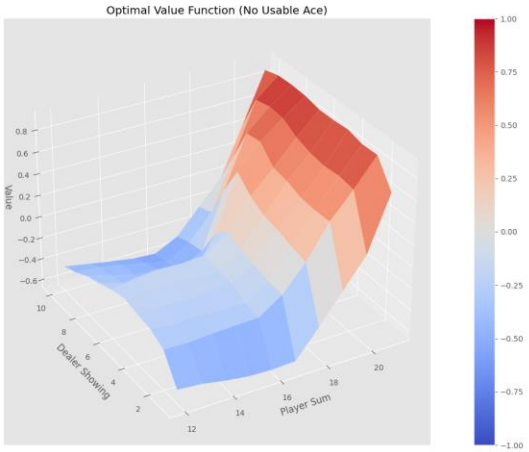
MC 算法

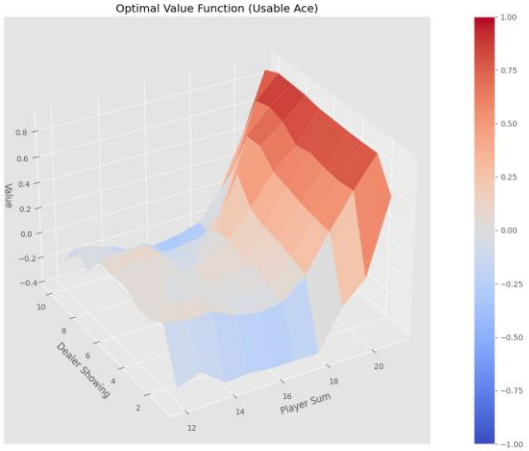
参数：num_episodes=10000, epsilon=0.1



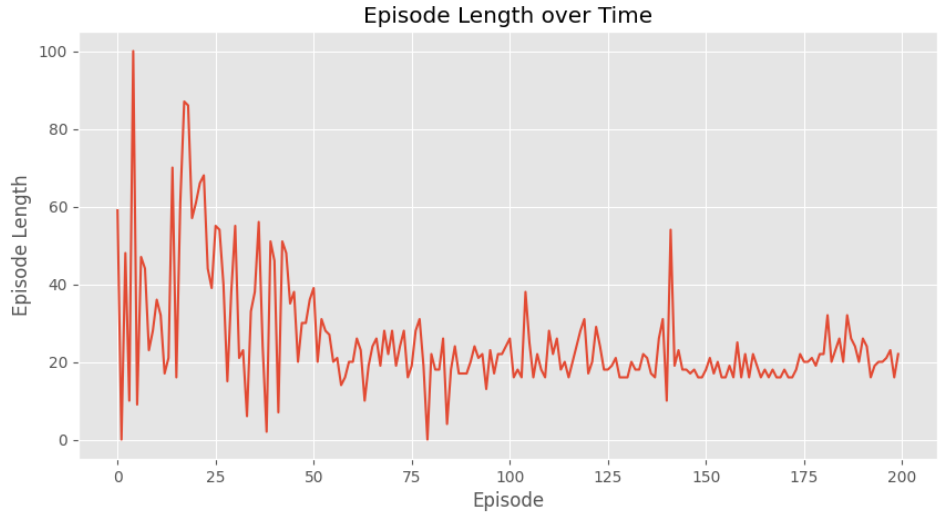


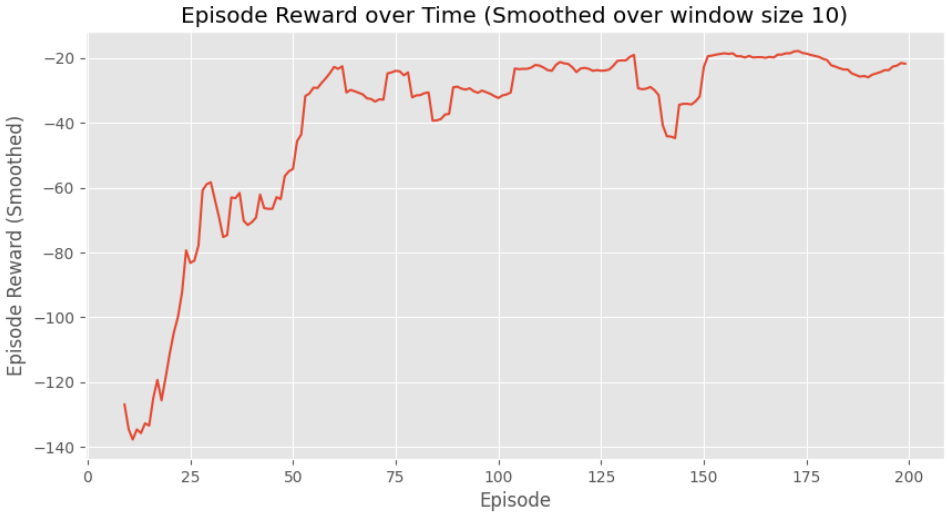
参数：num_episodes=500000, epsilon=0.1

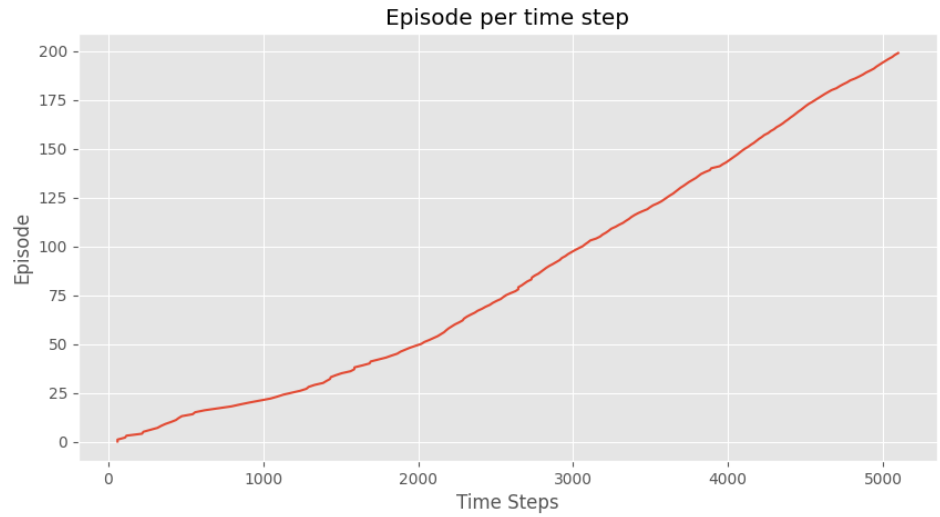




Sarsa 算法


Episode Length over Time
Episode Length
Episode
0
20
40
60
80
100
0
25
50
75
100
125
150
175
200


Episode Reward over Time (Smoothed over window size 10)
Episode Reward (Smoothed)
Episode
−20
−40
−60
−80
−100
−120
−140
0
25
50
75
100
125
150
175
200


Episode per time step
Episode
Time Steps
0
25
50
75
100
125
150
175
200
0
1000
2000
3000
4000
5000

Q_learning 算法

参数：num_episodes=500, discount_factor=1.0, alpha=0.5, epsilon=0.1