机器学习第四次作业

汪隽立 2021012957

2024 年 1 月 13 日

解答 2.1.

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{\tau=0}^{\infty}\gamma^{\tau}R_{t+\tau+1}\middle|S_t = s\right]$$

解答 2.2.

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[R_{t+1} + \gamma V^{\pi}(S_{t+1})\middle|S_t = s\right]$$

解答 2.3.

$$\begin{aligned}
V_1^{\pi_0}(A) &= -4 + \gamma V_0^{\pi_0}(B) = -4 \\
V_1^{\pi_0}(B) &= \frac{1}{2}\times(1 + \gamma V_0^{\pi_0}(A)) + \frac{1}{2}\times(2 + \gamma V_0^{\pi_0}(C)) = 1.5 \\
V_1^{\pi_0}(C) &= \frac{1}{2}\times(8 + \gamma(\frac{1}{4}V_0^{\pi_0}(C) + \frac{3}{4}V_0^{\pi_0}(A))) + \frac{1}{2}\times(0 + \gamma V_0^{\pi_0}(B)) = 4
\end{aligned}$$

解答 2.4.

$$\begin{aligned}
q_{\pi_0}(B, ba) &= 1 + \gamma V_1^{\pi_0}(A) = -1 \\
q_{\pi_0}(B, bc) &= 2 + \gamma V_1^{\pi_0}(C) = 4 \\
q_{\pi_0}(C, ca) &= 8 + \gamma(\frac{1}{4}V_1^{\pi_0}(C) + \frac{3}{4}V_1^{\pi_0}(A)) = 7 \\
q_{\pi_0}(C, cb) &= 0 + \gamma V_1^{\pi_0}(B) = 0.75
\end{aligned}$$

故更新后，$\pi_1(A) = ab$, $\pi_1(B) = bc$, $\pi_1(C) = ca$.

解答 3.1.

$$V(A) = \frac{1}{2}(0+2) = 1$$
$$V(B) = \frac{1}{2}(-2-3) = -\frac{5}{2}$$
$$Q(A,a) = \frac{1}{2}(0+2) = 1$$
$$Q(B,b) = \frac{1}{2}(-2-3) = -\frac{5}{2}$$

解答 3.2.

$$V(A) = \frac{1}{4}(0+2-3+1) = 0$$
$$V(B) = \frac{1}{4}(-2-3-3-3) = -\frac{11}{4}$$
$$Q(A,a) = \frac{1}{4}(0+2-3+1) = 0$$
$$Q(B,b) = \frac{1}{4}(-2-3-3-3) = -\frac{11}{4}$$

解答 3.3.

在初始时，$V(A) = V(B) = 0$, $Q(A,a) = Q(B,b) = 0$。V 的迭代过程如下：

$$V(B) = 0.1 \times (-2) = -0.2$$
$$V(A) = 0.1 \times (3 + V(B)) = 0.28$$
$$V(B) = 0.9 \times (-0.2) + 0.1 \times (-3+0) = -0.48$$
$$V(A) = 0.9 \times (0.28) + 0.1 \times (3 + V(A)) = 0.58$$
$$V(A) = 0.9 \times (0.58) + 0.1 \times (2 + V(B)) = 0.674$$
$$V(B) = 0.9 \times (-0.48) + 0.1 \times (-4 + V(A)) = -0.7646$$
$$V(A) = 0.9 \times (0.674) + 0.1 \times (4 + V(B)) = 0.93014$$
$$V(B) = 0.9 \times (-0.7646) + 0.1 \times (-3+0) = -0.98814$$

于是我们得到 $V(A) = 0.93014$, $V(B) = -0.98814$。因为整个过程中在状态 A, B 分别只采取了策略 a, b，故 $Q(A,a) = V(A) = 0.93014$，$Q(B,b) = V(B) = -0.98814$.

解答 4.3.

实验结果见表 1和表 2。可以看出，对 Q_Learning 和 Sarsa 而言，最佳的学习率都为 0.1，过高和过低的学习率会导致模型没有学习能力/损失函数震荡。

Learning Rate	Average Reward	Average Moves
0.01	−177.645	179.85
0.1	8.03	12.97
1	7.565	13.435
2	−200.0	200.0

表 1: 不同学习率下 Q_Learning 的表现

Learning Rate	Average Reward	Average Moves
0.01	−179.78	181.775
0.1	7.955	13.045
1	7.625	13.435
2	−200.0	200.0

表 2: 不同学习率下 SARSA 的表现

另外，可以看出 Q_Learning 和 Sarsa 都对学习率呈现了一定的鲁棒性，当学习率在 0.1 至 1 之间时，模型的表现都比较稳定。

解答 4.4.

实验结果如 1和 2所示。可以看出 Sarsa 相较于 Q_Learning 而言更稳定。

除此以外，观察实验框架给出的设定可知，初始时 ϵ-Greedy 算法的 $\epsilon = 1$，而随着学习的进行，ϵ 会慢慢降低。这可能导致的结果是，在 Sarsa 算法中，由于初始的 ϵ 过高，在更新 $Q(s, a)$ 时完全基于随机策略。故 Sarsa 的收敛性很可能受到前几次随机采样的影响。一个可以改进的地方是，可以将初始的 ϵ 适当调小，以平衡探索与利用在算法中的作用。

解答 5.3.

实验结果如图所示。AC 在 CartPole-v0 和 CartPole-v1 上均取得了最佳的平均奖励，而 REINFORCE 则有一些抖动。

可以明显地看出 TDActorCritic 相较于 REINFORCE 有更好的稳定性与收敛速率。这是因为，TDActorCritic 是基于 TD 的一步采样，拥有更小的方差。

除此以外，TDActorCritic 还展现了更好的鲁棒性。在 CartPole-v0 实验中，即使某一时刻偏离了最佳策略，TDActorCritic 也能迅速恢复到最佳策略上去。

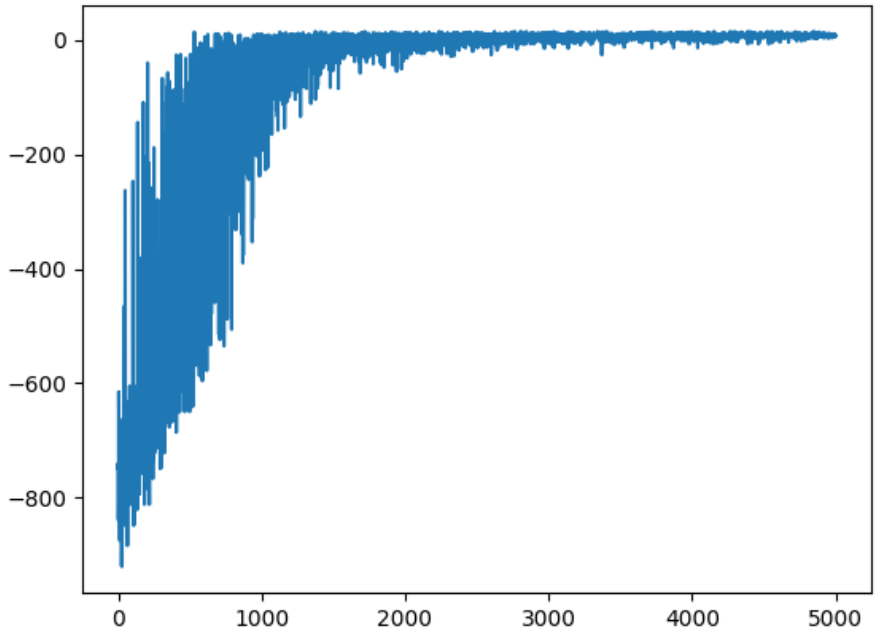

图 1: Q_Learning with lr=0.1

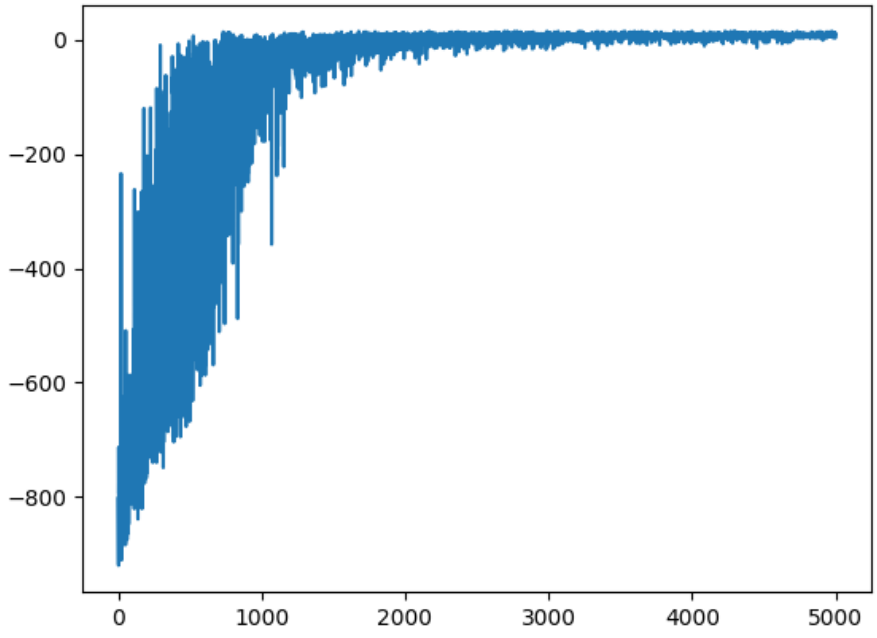

图 2: Sarsa with lr=0.1

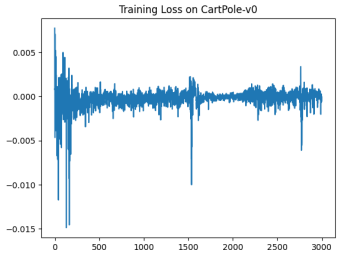

图 3: TDActorCritic 在 CartPole-v0 上的训练损失

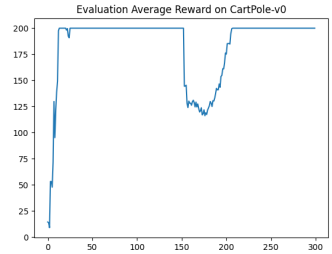

图 4: TDActorCritic 在 CartPole-v0 上的平均奖励

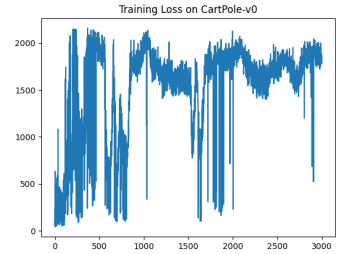

图 5: REINFORCE 在 CartPole-v0 上的训练损失

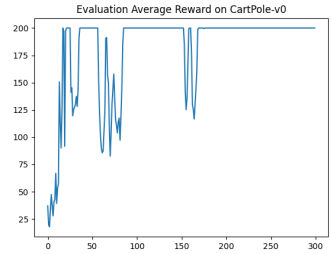

图 6: REINFORCE 在 CartPole-v0 上的平均奖励

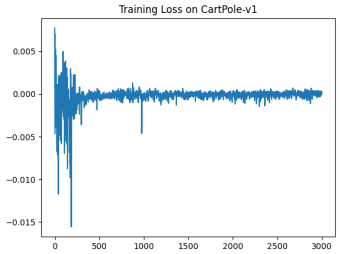

图 7: TDActorCritic 在 CartPole-v1 上的训练损失

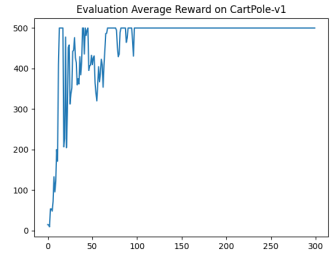

图 8: TDActorCritic 在 CartPole-v1 上的平均奖励

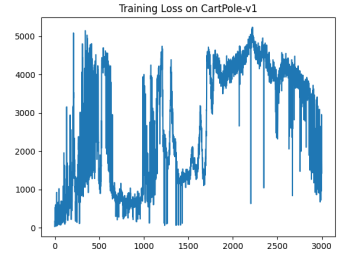

图 9: REINFORCE 在 CartPole-v1 上的训练损失

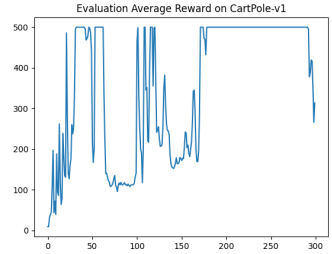

图 10: REINFORCE 在 CartPole-v1 上的平均奖励