2025-2026-1 学期强化学习课程 - 第二次作业

1120231863　左逸龙

October, 26 2025

1 通勤北理工

(i) 既然我们已经知晓了最优的 Q^* 表，那么每一状态下的最优策略满足：

$$\pi^*(a|s) = \begin{cases} 1 \text{ if } a = \text{argmax}_{a \in A(s)} Q^*(s,a) \\ 0 \text{ otherwise} \end{cases}$$

因此最优策略为：

- 在状态 S_1 下，乘坐班车
- 在状态 S_2 下，乘坐班车
- 在状态 S_3 下，乘坐地铁

(ii) 此时最优策略为：

- 在状态 S_{12} 下，乘坐班车
- 在状态 S_3 下，乘坐班车

显然，此时得到的最优策略与使用真实三状态表示时得到的最优策略不同，关键区别在于原来状态 S_3 的最优策略为**乘坐地铁**，而现在状态 S_3 的最优策略为**乘坐班车**。

之所以会出现这样的变化，是因为**状态聚合**导致智能体 Agent 无法区分S_1与S_2的**未来价值**，使得决策依据退化为**即时奖励**。分析如下：

- Q-learning 算法的更新公式为：

$$Q(s,a) = Q(s,a) + \alpha \left[R + \gamma \max_{a \in A(s')} Q(s',a) \right]$$

其中α为学习率，R为即时奖励，γ为折扣因子。由公式可以看出，决定 Q 值的不仅仅只有**即时奖励**，还有**未来价值的预期**。

- 在原始的三状态模型中，Agent 可以精确地知道每个动作会导向哪个具体的状态。从 S_3 出发，乘坐地铁会到达 S_2，而 S_2 的长期价值 ($V^*(S_2) = 1.95$) 远高于乘坐班车所到达的 S_1 的长期价值 ($V^*(S_1) = 1.65$)。尽管坐地铁的即时奖励更低，但为了追求 S_2 带来的更高未来收益，最优策略是选择乘坐地铁。

- 在聚合后的二状态模型中，S_1 和 S_2 被合并为宏状态 S_{12}。此时，无论从 S_3 出发选择乘坐班车（到达 S_1）还是乘坐地铁（到达 S_2），在 Agent 看来，下一个状态都是**同一个** S_{12}。因此，这两个动作所带来的未来价值预期是完全相同的（都等于 $\gamma V^*(S_{12})$）。
- 当两个动作的**未来价值预期相同时**，决策的优劣就完全取决于**即时奖励**。根据表 1，$R(S_3, \text{班车}) = -0.5$，而 $R(S_3, \text{地铁}) = -0.7$。由于 $-0.5 > -0.7$，选择乘坐班车能获得更好的即时奖励。因此，在这种信息受限的情况下，最优策略从乘坐地铁转变为乘坐班车。

综上，导致这种策略上变化的原因是状态表示的粒度变粗后，Q-learning 算法泛化（或平均化）了不同状态的价值，导致决策依据从**长远未来价值**退化为**即时奖励**。

2 Frozenlake 小游戏

(i)