2025-2026-1 学期强化学习课程 - 第一次作业

Chen Fang, Linkang Dong

September, 29 2025

1 马尔可夫决策过程

你来到一家高端电竞馆！你手上有 20 元，会一直玩下去，直到输光所有钱，或者本金翻倍（即持有金额至少达到 40 元）。你可以从两台老虎机中选择一台进行游戏：

1. 老虎机 A：每次下注 10 元，有 0.05 的概率获得 20 元（即净赚 10 元），否则获得 0 元（即输掉 10 元）。

2. 老虎机 B：每次下注 20 元，有 0.01 的概率获得 30 元（即净赚 10 元），否则获得 0 元（即输掉 20 元）。

在结束之前，你每轮都要选择玩老虎机 A 还是老虎机 B。请在下方给出一个能够刻画上述情景的 MDP，描述其状态空间、动作空间、奖励函数与转移概率。假设折扣因子 $\gamma = 1$。你可以用方程、表格或图示来表达解答。

解答：请在此处填写解答。

2 Gridworld 小游戏

考虑如下的网格环境：

- 从任意非阴影的格子出发，你可以向上、向下、向左或向右移动。动作是是确定性的，意味着动作执行后一定会从一个状态到达另一个状态（例如从状态 13 向上走，可以直接到状态 9）。
- 较粗的边表示墙壁，尝试向墙壁方向移动将会导致智能体原地不动（例如从状态 13 向右走，无法到达状态 14，仍会停留在原状态 13）。
- 在绿色目标格子（编号 3）采取任何动作将获得奖励 r_g（因此 $r(3,a)=r_g$ 对所有动作 a 成立），并结束回合。在红色死亡格子（编号 14）采取任何动作将获得奖励 r_r（因此 $r(14,a)=r_r$ 对所有动作 a 成立），并结束回合。
- 在其他所有格子中，采取任何动作都与奖励 $r_s \in \{-1,0,+1\}$ 相关（即使该动作导致智能体保持在原地）。

除了特殊规定外，以下假设折扣因子 $\gamma=1$，$r_g=+3$，且 $r_r=-3$。

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

(a) [30 分] 最短路径策略

定义 r_s 的值，使得最优策略会返回绿色目标格子（编号 3）的最短路径。使用这个 r_s，找到每个格子的最优价值。

解答：请在此处填写解答。

(b) [20 分] 奖励变化的影响

让我们将 (a) 中导出的价值函数称为 $V_{\text{old}}^{\pi_g}$ 和策略称为 π_g。假设我们现在在一个新的网格世界中，特殊状态的奖励（r_r、r_g）都乘以 2，每步的奖励 r_s 都加 2，考虑仍然遵循原始网格世界的 π_g，在这个第二个网格世界中新的值 $V_{\text{new}}^{\pi_g}$ 将是什么?

解答：请在此处填写解答。

(c) [30 分] 奖励变化的一般表达式

考虑一个一般的马尔可夫决策过程（MDP），其中包含奖励和转移。考虑折扣因子 γ。在这种情况下，假设该马尔可夫链是无限的（即没有终止）。在这个 MDP 中，策略 π 引发了一个价值函数 V^{π}（我们将其称为 V_{old}^{π}）。现在假设我们有一个新的 MDP，唯一的不同是所有的奖励都乘上了一个常数 c。

1. 你能给出一个表达式，用于表示策略 π 在这个第二个 MDP 中引发的新价值函数 V_{new}^{π} 与 V_{old}^{π}、c 和 γ 之间的关系吗?
2. 是否存在特定的 c 使得最优策略发生变化? 如果存在，请给出 c 使得策略变化的取值范围，并说明变化理由，反之则给出不存在的理由。

解答：请在此处填写解答。

(d) [20 分] 正奖励的影响

让我们回到 (a) 中的网格世界，使用默认值 r_g、r_r、γ 和你指定的 r_s 值。假设我们现在通过对所有奖励（r_s、r_g 和 r_r）增加一个常数 c 来导出第二个网格世界，使得 $r_s = +2$。最优策略将如何变化（只需给出一两句话的描述)? 非阴影格子的价值将变为多少?

解答：请在此处填写解答。