

《人工智能导论》2026

技术科学试验班

第三次作业

2026 年 5 月 25 日

1. （贝尔曼方程）[20pt] 有一个容量为 $W = 6$ 的背包，现有 4 个物品，每个物品最多只能选择一次，物品的重量和价值如下：

物品编号	重量 w_i	价值 v_i
1	2	3
2	3	4
3	4	5
4	5	8

定义状态： $F(i, c)$ 表示当前考虑到第 i 个物品、背包剩余容量为 c ，动作 $a \in \{0, 1\}$ ，0 表示不选第 i 个物品，1 表示选。

请回答以下问题：

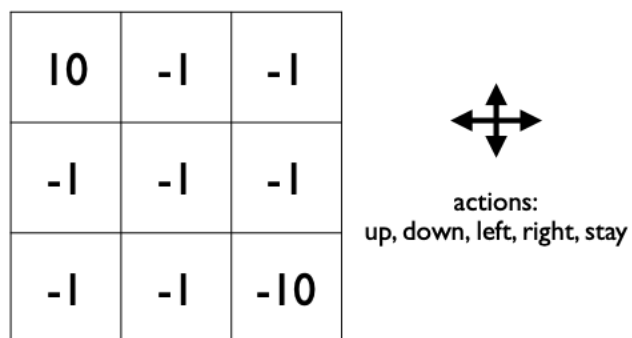
- 1) 写出该问题的贝尔曼最优方程，并解释其含义。
- 2) 请计算并填写相应的动态规划表

$F(i, c)$	$c = 0$	$c = 1$	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$
$i = 0$							

$i = 1$							
$i = 2$							
$i = 3$							
$i = 4$							

3) 根据动态规划表，给出能够背包能够装的最大价值，以及选择了哪些物品。

2. (价值迭代)[30pt] 在下图所示网格世界任务中，表格代表了 9 个 state，单元格内的数值表示达到这一状态能够得到的 reward，假定执行动作状态转移是确定的，折扣因子 $\gamma = 0.9$ 。请画出使用价值迭代算法前 5 轮每个 state 对应的状态值函数以及策略的变化情况。(提示：可参考课上例子，每一轮画出表格，在对应的单元格内填上值函数和当前的策略，初始状态值函数全部设置为 0)



3. （策略迭代）[30pt] 考虑如下任务，状态空间为 $\{A, B\}$ ，动作空间为{左，右}，折扣因子 $\gamma = 0.9$

状态	动作	下一状态	奖励
A	左	A	-1
A	右	B	+2
B	左	A	0
B	右	B	+1

初始策略 π_0 为：在 A 选“左”，在 B 选“左”

- （1）对 π_0 做策略评估（求解 V^{π_0} ）
- （2）对 π_0 做一次策略改进，得到 π_1
- （3）判断 π_1 是否为最优策略

4. （时序差分学习）[10pt] 智能体在一条轨迹中观测到：

$$S_1 \xrightarrow{r=1} S_2 \xrightarrow{r=2} S_3 \xrightarrow{r=3} \text{终止}$$

初始 $V(S_1) = V(S_2) = V(S_3) = 0$ ， $\alpha = 0.5$ ， $\gamma = 1$ （提示： α 为增量蒙特卡洛更新公式的更新率）

请分别用 TD 算法和蒙特卡洛强化学习算法更新一次，计算更新后各状态的 V 值。

5. （时序差分学习）[10pt] 智能体处于状态 S ，执行动作 a_1 ，转移到 S' ，获得奖励 $r = 1$ 。在 S' 有两个动作可选： $Q(S', a_1) = 3$ ， $Q(S', a_2) = 5$ 。当前 $Q(S, a_1) = 2$ ， $\alpha = 0.1$ ， $\gamma = 0.9$ ，智能体在 S' 实际选择了动作 a_1 ，请分别用 SARSA 和 Q-Learning 算法更新 $Q(S, a_1)$ 。