Lab 5: Reinforcement Learning

本节实验共有两个**TODO**，请留意。

1 Markov Reward Process

1.1 Intro

Reward 在一个马尔可夫奖励过程中，从第t时刻状态s_t开始，直到终止状态时，所有奖励的衰减之和称为回报G_t. 公式为：

$$\begin{aligned} G_t :=& R_t + \gamma R_{t+1} + \gamma R_{t+2} + ... \\ =& \sum_{k=0}^{\infty} \gamma^k R_{t+k} \end{aligned}$$

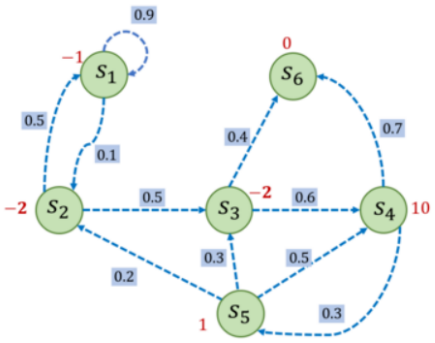

Figure 1: 马尔可夫奖励过程示例

例：从s_1开始，选取一条状态序列为$s_1 \to s_2 \to s_3 \to s_6, \gamma = 0.5, s_1$的回报$G_1$为：

$$\begin{aligned} G_1 =& -1 + 0.5 \times (-2) + 0.5^2 \times (-2) \\ =& -2.5 \end{aligned}$$

1.2 TODO-1

代码：实现回报公式。

报告：代码思路，运行结果截图，实验结果分析。

2 Markov Decision Process

2.1 Intro

相较于马尔可夫回报过程(MRP)，马尔可夫决策过程(MDP)还多了环境的刺激，我们将环境的刺激称为动作(action)，在马尔可夫回报过程中加入动作(action)就得到了马尔可夫决策过程，由五元组(S, A, P, r, γ)构成

- S 是状态的集合
- A 是动作的集合
- γ 是折扣因子
- $r(s,a)$ 是奖励函数，奖励可同时取决于状态s和动作a，也可只取决于状态s，当仅取决于状态s时奖励函数退化为$r(s)$
- $P(s',a)$ 是状态转移函数，表示状态s执行动作a后到达状态s'的概率

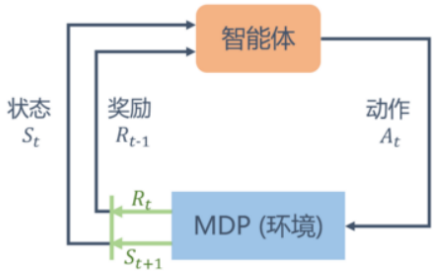

Figure 2: 智能体与环境MDP的交互示意图

policy (策略) ：智能体根据当前状态从动作集合A中选择一个动作的函数称为策略，通常用字母π表示，策略$\pi(a|s) = P(A_t = a|S_t = t)$是一个函数，表示在状态$s$时采取动作$a$的概率。策略在每个状态的输出是关于动作的概率分布：如果是确定性的动作，只有一个动作概率为1，其余为0

State value function (状态价值函数) 我们用$V^\pi(s)$表示MDP中基于策略π的状态价值函数，定义从状态s出发遵循策略π能获得的期望回报为

$$V^\pi(s) = \mathbb{E}_\pi[G_t|S_t = s]$$

State-action value function (动作价值函数) 由于动作的存在，马尔可夫决策过程的价值函数在马尔可夫回报过程的价值函数有些差异，定义:

$$Q^\pi(s,a) = \mathbb{E}_\pi[G_t|S_t = s, A_t = a]$$

表示马尔可夫决策过程遵循策略π时，对当前的状态s执行动作a得到的期望回报。状态s的价值等于在该状态下基于策略π采取所有动作的概率与相应的价值相乘再求和的结果：

$$V^\pi(s) = \sum_{a\in A}\pi(a|s)Q^\pi(s,a)$$

使用策略π时，状态s下采取动作a的价值等于即时奖励加上经过衰减后的所有可能的下一个状态的状态转移概率与相应的价值的乘积：

$$Q^\pi(s,a) = r(s,a) + \gamma\sum_{s'\in S}P(s'|s,a)V^\pi(s')$$

Bellman expectation equation (贝尔曼期望方程) 推导过程：

$$\begin{aligned}V^\pi(s) =& \mathbb{E}_\pi[G_t|S_t = s]\\ =& \mathbb{E}_\pi[R_t + \gamma V^\pi(s_{t+1})|S_t = s]\\ =& \sum_{a\in A}\pi(a|s)[r(s,a) + \gamma\sum_{s'\in S}P(s'|s,a)V^\pi(s')]\end{aligned}$$

$$\begin{aligned}Q^\pi(s,a) =& \mathbb{E}_\pi[R_t + \gamma Q^\pi(s_{t+1},a_{t+1})|S_t = s, A_t = a]\\ =& r(s,a) + \gamma\sum_{s'\in S}P(s'|s,a)\sum_{a'\in A}\pi(a'|s')Q^\pi(s',a')\end{aligned}$$

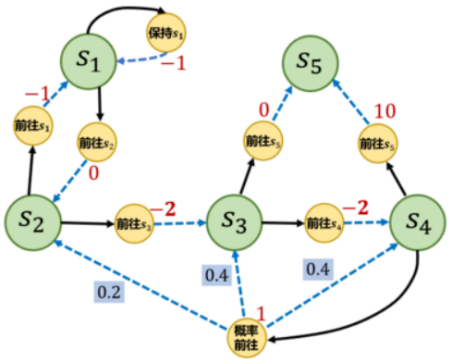

Figure 3: 马尔可夫决策过程一个简单例子

2.2 TODO-2

代码：根据实验代码，自行修改状态转移函数以及奖励函数等的参数，体会在马尔可夫决策过程中每个状态价值的变化。

报告：多次修改参数。每次修改时的运行结果截图和实验结果分析。

Submit

- 提交一个zip文件，注意命名，形如：2022101000+张三+实验5.zip
- 除非必要，zip中应当仅包含一个报告pdf和一个代码文件，命名不作要求
- https://k.ruc.edu.cn DDL2023.11.6 23:59