人工智能原理

作业 7

注意：

1） 请在网络学堂提交**电子版**；

2） 请在 **2025 年 6 月 4 日晚 23:59:59** 前提交作业，**不接受补交**；

3） **3 个选做题中任选 2 道题作答，多做不加分**，多做则按照题目的解答顺序，只计算前 2 道题目的分数，例如提交作业中题目解答顺序是 1、3、2，则只对 1、3 计分

4） 如有疑问，请联系助教：

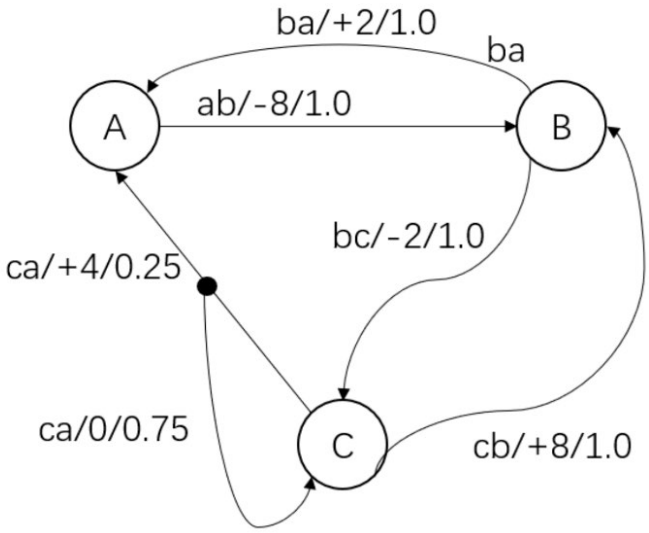
李 震：lizhen22@mails.tsinghua.edu.cn

李可伊：lky23@mails.tsinghua.edu.cn

王子安：wangza24@mails.tsinghua.edu.cn

1. 价值迭代

考虑如下图所示的马尔可夫决策过程，折现因子$\gamma = 0.5$。图中大写字母表示状态；状态之间的有向边表示转移；边上的三元组“actions/rewards/probability”给出了动作、回报及转移概率。



现有均匀随机策略$\pi_1(a|s)$，即从一个状态s出发，等概率地选择下一个动作。假设有初始状态值$V_1(a) = V_1(b) = V_1(c) = 4$，请完成如下任务：

(1) 计算经过一轮同步价值迭代后的状态价值，并根据确定性贪心策略给出策略$\pi_2(a|s)$。

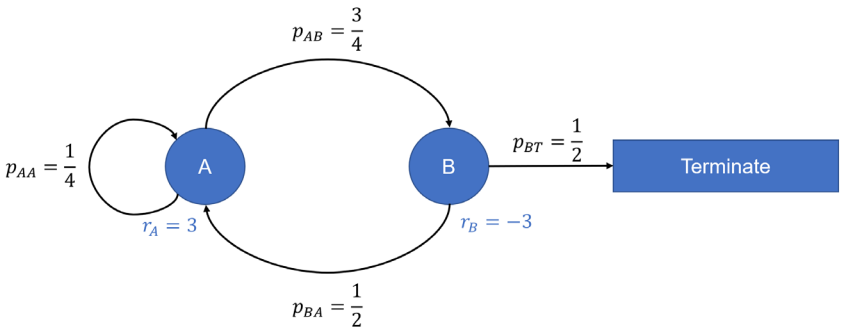
(2) 计算经过一轮异步价值迭代后的状态价值，并根据确定性贪心策略给出策略$\pi_2'(a|s)$，约定异步价值迭代按照$A \to B \to C$的顺序完成状态价值更新。

说明：在上图所有的 action 中，ca 较为特殊，它以 1/4 的概率从状态 C 转移到 A，以 3/4 的概率保持状态 C 不变，保持不变时回报为 0。

2.蒙特卡洛

一个无折现($\gamma = 1$)的马尔可夫回报过程，具有 A 和B 两个状态以及一个终止状态。

(1) 若状态转移图和状态期望回报函数如下图所示，请写出该马尔可夫回报过程的状态价值贝尔曼期望方程，并求解该方程得出状态价值函数 $v(\mathrm{A}), v(\mathrm{B})$。



(2) 若状态转移图及回报函数未知，但已知以下两个观测片段

$$\mathrm{A} \xrightarrow{+3} \mathrm{A} \xrightarrow{+2} \mathrm{B} \xrightarrow{-4} \mathrm{A} \xrightarrow{+4} \mathrm{B} \xrightarrow{-3} \text{terminate}$$

$$\mathrm{B} \xrightarrow{-2} \mathrm{A} \xrightarrow{+3} \mathrm{B} \xrightarrow{-3} \text{terminate}$$

其中 $\mathrm{A} \xrightarrow{+3} \mathrm{A}$ 表示以回报值 $+3$ 从 A 状态转移到 A 状态。请分别使用**首次访问**和**每次访问**的蒙特卡洛预测，估计状态价值函数 $v(\mathrm{A}), v(\mathrm{B})$。

3. 时序差分

考虑下方一个3×3网格图，左上角和右下角为终止状态。非终止状态集合 $\mathrm{S} = \{1, 2, \ldots, 7\}$，每个状态有四种可能的动作 {上，下，左，右}。每个动作会导致状态转移，对于每次转移$\mathrm{R_t} = -1$，但当动作会导致智能体移出网格时，状态保持不变。

	1	2
3	4	5
6	7	

(1) 设初始的 V 值为

0	0	0
0	0	0
0	0	0

观察到的一个 episode 如下:

$$4 \to 1 \to 4 \to 7 \to \text{terminate}$$

取 $\alpha = 0.5, \gamma = 1$，请利用时序差分算法计算该 episode 之后 V 值的更新情况，写出每步的更新过程。

(2) 假设初始状态为 4，初始化的 Q 表如下，其中从左到右每列依次代表状态 $1, 2, \ldots, 7$，从上到下每行依次代表动作上、右、下、左，$Q(terminate, a) = 0$，$\gamma = 1$，$\alpha = 1$。

-4	-3	-1	-3	-4	-2	-4
-3	-3	-2	-4	-2	-3	-3
-4	-3	-4	-2	-2	-3	-4
-3	-2	-3	-3	-4	-3	-2

请写出 SARSA 算法（为了计算方便，假设行为策略和目标策略均由确定性贪心策略给出）在一个 episode 后（即第一次到达终止状态后）更新的 Q 表。