人工智能基础

作业 8

注意：

1） 请在网络学堂提交**电子版**；

2） 请在 **12 月 22 日晚 23:59:59** 前提交作业，**不接受补交**；

3） 如有疑问，请联系助教：

杨鹏帅：yps18@mails.tsinghua.edu.cn

鄞启进：yqj17@mails.tsinghua.edu.cn

崔雪建：cuixj19@mails.tsinghua.edu.cn

高子靖：gzj21@mails.tsinghua.edu.cn

鲁永浩：yonghao.lu@foxmail.com

江澜：jiangl20@mails.tsinghua.edu.cn

牛家赫：njh20@mails.tsinghua.edu.cn

尹小旭：yxx21@mails.tsinghua.edu.cn

请从以下四题中任选两题作答。

1. 策略改进

(1) 对于任意策略π，基于确定性贪心策略进行策略改进得到π'

$$\pi'(a|s)=\begin{cases}1, & \text{if } a=\underset{a\in A}{\operatorname{argmax}}\, q_\pi(s,a)\\ 0, & \text{otherwise.}\end{cases}$$

请证明，改进后的策略π'优于原策略π。即证明：$v_{\pi'}(s)\geq v_\pi(s)$。

(2) 对于任意策略π，基于$\varepsilon-$greedy策略进行策略改进得到π'

$$\pi'(a|s)=\begin{cases}1-\varepsilon+\dfrac{\varepsilon}{|A|}, & \text{if } a=\underset{a\in A}{\operatorname{argmax}}\, q_\pi(s,a)\\ \dfrac{\varepsilon}{|A|}, & \text{otherwise.}\end{cases}$$

请证明，改进后的策略π'优于原策略π。

(3) 你觉得$\varepsilon-$greedy策略相较于确定性贪心策有什么优势。

2. 价值迭代

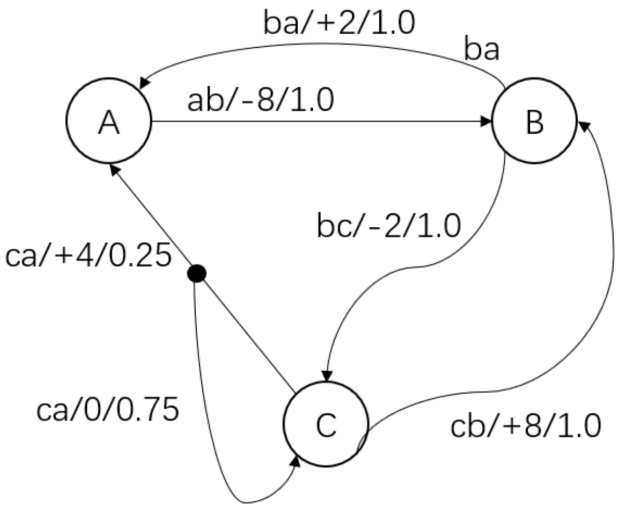
考虑如下图所示的马尔可夫决策过程(MDP)，折现因子$\gamma=0.5$。图中大写字母表示状态；状态之间的有向边表示转移；边上的三元组“actions/rewards/probability”给出了各个转移的名称、回报及转移概率。

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$V_1(a)=V_1(b)=V_1(c)=2$，请给出：

(1) 通过同步迭代和确定性贪心策略得到的策略$\pi_2(a|s)$。

(2) 通过异步迭代和确定性贪心策略得到的策略$\pi_2{}'(a|s)$。

说明：在下图所有的 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。



3. 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，其状态转移矩阵及回报函数未知。现在观测到了两个行动过程如下。

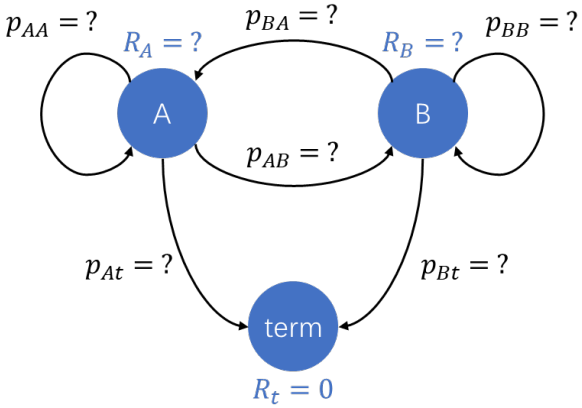
$$\mathrm{A} \xrightarrow{+3} \mathrm{A} \xrightarrow{+2} \mathrm{B} \xrightarrow{-4} \mathrm{A} \xrightarrow{+4} \mathrm{B} \xrightarrow{-3} \text{terminate}$$

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

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。

(1) 分别使用**首次访问**与**每次访问**的蒙特卡洛预测，估计状态价值函数$v(\mathrm{A}), v(\mathrm{B})$。

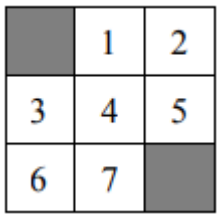
(2) 绘出上述过程的马尔可夫回报过程图，并在图中标出状态 A 与 B 的平均回报与平均状态转移概率(如图所示)，使其可以最佳拟合表示观测到的两个行动过程(即数据的最大化似然估计)。



(3) 写出该马尔可夫回报过程的状态价值贝尔曼期望方程，并利用(b)中估计的状态回报与状态转移概率求解该方程得出状态价值函数$v(\mathrm{A}), v(\mathrm{B})$。

4. 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$S=\{1,2,\ldots,7\}$，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$R_t=-1$，但当动作会导致智能体移出网格时，状态保持不变。



(1) 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：

$$4\rightarrow5\rightarrow4\rightarrow3\rightarrow\text{terminate}$$

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,...,7，从上到下每行依次代表动作上，右，下，左，$Q(\text{terminate},\ a)=0$，$\gamma=1$，$\alpha=1$。

-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

请分别写出 SARSA 算法和 Q-learning 算法（均采用贪心方式选择动作）在一个 episode 后（即第一次到达终止状态后）更新的 Q 表。