2.

1) 同步：$V_{k+1}(s) = \max_{a\in A}\left[r_s^a + \gamma \sum_{s\in S} P_{ss'}^a V_k(s')\right]$

$V_1(A) = -8 + 0.5\cdot 1\cdot V_0(B) = -8$

$V_1(B) = \max\{-3 + 0.5\cdot 1\cdot V_0(C),\ 2 + 0.5\cdot 1\cdot V_0(A)\} = 2$

$V_1(C) = \max\{0.25(4 + 0.5\cdot V_0(A)) + 0.75(0 + 0.5 V_0(C)),\ 8 + 0.5\cdot 1\cdot V_0(B)\} = 8$

$V_2(A) = -8 + 0.5\cdot 1\cdot V_1(B) = -7$

$V_2(B) = \max\{-3 + 0.5\cdot 1\cdot V_1(C),\ 2 + 0.5\cdot 1\cdot V_1(A)\} = 1$

$V_2(C) = \max\{0.25(4 + 0.5\cdot V_1(A)) + 0.75(0 + 0.5 V_1(C)),\ 8 + 0.5\cdot 1\cdot V_1(B)\} = 9$

贪心策略：$\pi_2(a=ab \mid s|A) = 1$

$\pi_2(a=bc \mid s=B) = 1$ $\quad \pi_2(a=ba \mid s=B) = 1$

$\pi_2(a=cb \mid s=C) = 1$ $\quad \pi_2(a=ca \mid s=C) = 0$

2) 异步：$V(s) = \max_{a\in A}\left[r_s^a + \gamma \sum_{s\in S} P_{ss'}^a V(s')\right]$

$V(A) = -8 + 0.5\cdot 1\cdot V(B) = -8$

$V(B) = \max\{-3 + 0.5\cdot 1\cdot V(C),\ 2 + 0.5\cdot 1\cdot V(A)\} = -2$

$V(C) = \max\{0.25(4 + 0.5\cdot V(A)) + 0.75(0 + 0.5 V(C)),\ 8 + 0.5\cdot 1\cdot V(B)\} = 7$

$V(A) = -8 + 0.5\cdot 1\cdot V(B) = -9$

$V(B) = \max\{-3 + 0.5\cdot 1\cdot V(C),\ 2 + 0.5\cdot 1\cdot V(A)\} = 0.5$

$V(C) = \max\{0.25(4 + 0.5\cdot V(A)) + 0.75(0 + 0.5 V(C)),\ 8 + 0.5\cdot 1\cdot V(B)\} = 8.25$

贪心策略：$\pi_2(a=ab \mid s|A) = 1$

$\pi_2(a=bc \mid s=B) = 1$ $\quad \pi_2(a=ba \mid s=B) = 0$

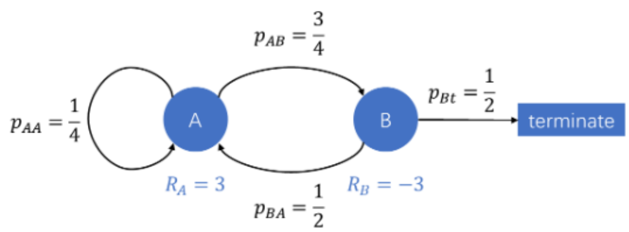
$\pi_2(a=cb \mid s=C) = 1$ $\quad \pi_2(a=ca \mid s=C) = 0$

3.

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。

1）若状态转移图和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$v(A), v(B)$。



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

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

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

其中$A \xrightarrow{+3} A$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**与**每次访问**的蒙特卡洛预测，估计状态价值函数$v(A), v(B)$。

(1)

状态价值的贝尔曼期望方程为：$V_\pi(s) = \sum_{a \in A} \pi(a|s) \left(r_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_\pi(s') \right)$

故：$V_\pi(A) = P_{AA} \cdot (3 + V_\pi(A)) + P_{AB} \cdot (-3 + V_\pi(B))$

$V_\pi(B) = P_{BA} (3 + V_\pi(A)) + P_{Bt} \cdot 0$

$\Rightarrow \begin{cases} V_\pi(A) = -1 \\ V_\pi(B) = 1 \end{cases}$

(2)

首次访问：

episode 1 : $V_A = 2$ $V_B = -3$

episode 2 : $V_A = 0$ $V_B = -2$

avg : $V_A = 1$ $V_B = -2.5$

故： $V(A) = 1$ $V(B) = -2.5$

每次访问：

episode 1 : $V_A = [2, -1, 1]$ $V_B = [-3, -3]$

episode 2 : $V_A = [0]$ $V_B = [-2, -3]$

avg : $V_A = 0.5$ $V_B = -2.75$

故： $V(A) = 0.5$ $V(B) = -2.75$

4. 时序差分.

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$S=\{1,2,\dots,7\}$，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	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：

$$4\rightarrow1\rightarrow4\rightarrow7\rightarrow \text{terminate}$$

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。

1)

时序差分算法：$V(s)\leftarrow V(s)+\alpha[R+\gamma V(s')-V(s)]$

$4\longrightarrow1\longrightarrow4\longrightarrow7\longrightarrow$ terminate

0	0	0
0	-0.5	0
0	0	0

0	-0.75	0
0	-0.5	0
0	0	0

0	-0.75	0
0	-0.75	0
0	0	0

0	-0.75	0
0	-0.75	0
0	-0.5	0

2)

	1	2	3	4	5	6	7
上	-4	-3	-1	-3	-4	-2	-4
右	-3	-3	-2	-4	-2	-3	-3
下	-4	-3	-4	-2	-2	-3	-4
左	-3	-2	-3	-3	-4	-3	-2

SARSA算法：$Q(s,A)\leftarrow Q(s,A)+\alpha[R+\gamma Q(s',A')-Q(s,A)]$

$4\longrightarrow7\longrightarrow6$

	1	2	3	4	5	6	7
上							
右							
下				-3			
左							

$Q(4,\downarrow)\leftarrow R+Q(7,\leftarrow)$

	1	2	3	4	5	6	7
上							
右							
下							
左							-3

$Q(7,\leftarrow)\leftarrow R+Q(6,\uparrow)$

	1	2	3	4	5	6	7
上						-2	
右							
下							
左							

$Q(6,\uparrow)\leftarrow R+Q(3,\uparrow)$

$6 \longrightarrow 3 \longrightarrow \text{terminate}$

	1	2	3	4	5	6	7
上			-1				
右							
下							
左							

$Q(3,\uparrow) \leftarrow R + Q(\text{terminal})$

故1个episode后Q表为：

	1	2	3	4	5	6	7
上	-4	-3	-1	-3	-4	-2	-4
右	-3	-3	-2	-4	-2	-3	-3
下	-4	-3	-4	-3	-2	-3	-4
左	-3	-2	-3	-3	-4	-3	-3