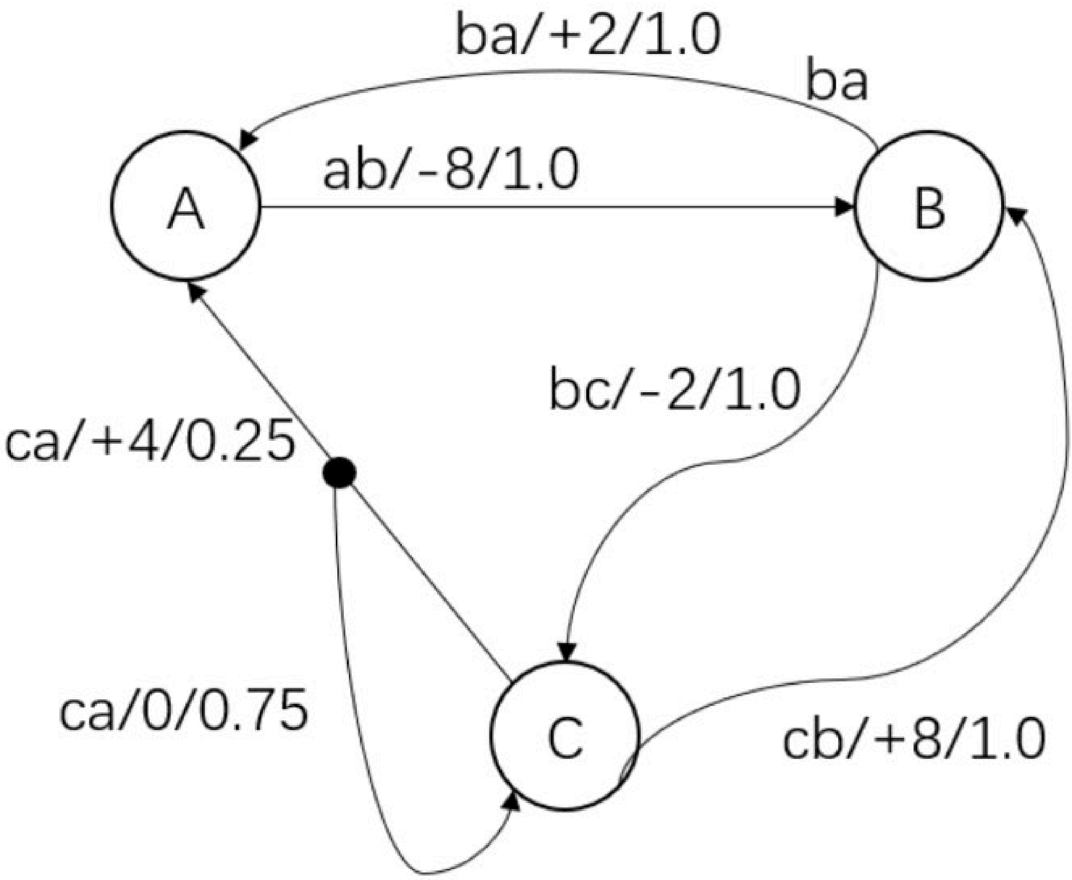
1 价值迭代

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。

解：

（1）**同步价值迭代**：先用**旧值** $V_1(\cdot)$ 计算新一轮的 **状态价值** $V_2(\cdot)$，统一再替换

- 对于状态 A，可能的行动为 A→B，则：

$$V_2(A) = -8 + \gamma \times V_1(B) = -8 + 0.5 \times 4 = -6$$

- 对于状态 B，可能的行动为 B→A 以及 B→C，其中 B→A 的行动价值为：

  $$2 + \gamma \times V_1(A) = 2 + 0.5 \times 4 = 4$$

  B→C 的行动价值为：

  $$-2 + \gamma \times V_1(C) = -2 + 0,5 \times 4 = 0$$

  贪心策略取 max

  $$V_2(B) = \max(4, 0) = 4$$

- 对于状态C，有两条行动路径，选择 C→B 的行动价值为：

  $$8 + \gamma \times V_1(B) = 8 + 0.5 \times 4 = 10$$

  执行 ca 后，执行特殊跳转，以 $\frac{1}{4}$ 概率到 A，回报 4，后续价值 $\gamma \times V_2(A)$，以 $\frac{3}{4}$ 概率留在 C，回报 0，后续价值 $\gamma \times V_2(C)$。

  因此，选择 C→A 的行动价值为

  $$\frac{1}{4} \times (4 + \gamma \times V_1(A)) + \frac{3}{4} \times (0 + \gamma \times V_1(C)) = 3$$

  则：

  $$V_2(C) = \max(10, 3) = 10$$

状态 S	V_1	V_2
A	4	-6
B	4	4
C	4	10

确定性贪心策略 $\pi_2(a|s)$

基于上面计算出的新一轮的状态价值，计算每个状态下可能动作的动作价值 $q_\pi(s, a)$，然后选值最大的动作，构成贪心策略。

- 对于 A，最优策略只能为 A→B，因此

  动作价值：

  $$q_\pi(A, ab) = -8 + 0.6 \times V_2(B) = -6$$

  因为只有一种动作，所以

  $$\pi_2(ab|A) = 1$$

- 对于B，有两种动作 B → A 和 B → C

$$q_\pi(B, ba) = 2 + 0.5 \times (-6) = -1$$
$$q_\pi(B. bc) = -2 + 0.5 \times 10 = 3$$

比较两个值

$$q_\pi(B, ba) < q_\pi(B. bc)$$

所以确定性贪心策略选 B→C：

$$\pi_2(bc|B) = 1$$

- 对于 C，有两种动作 C → A 和 C → B

执行 ca 后，执行特殊跳转，以 $\frac{1}{4}$ 概率到 A，回报 4，后续价值 $\gamma \times V_2(A)$，以 $\frac{3}{4}$ 概率留在 C，回报 0，后续价值 $\gamma \times V_2(C)$，

$$q_\pi(C, ca) = \frac{1}{4} \times (4 + \gamma \times V_2(A)) + \frac{3}{4} \times (0 + \gamma \times V_2(C)) = 4$$
$$q_\pi(C, cb) = -8 + 0.5 \times V_2(B) = 10$$

比较两个值

$$q_\pi(C, ca) < q_\pi(C. cb)$$

所以确定性贪心策略选 C→B：

$$\pi_2(cb|C) = 1$$

综上所述：$\pi_2(a|s)$ 中有 $\pi_2(ab|A) = 1$、$\pi_2(bc|B) = 1$、$\pi_2(cb|C) = 1$ 其余均为 0

（2）**异步迭代**：按顺序逐个更新状态价值，每算出一个新值 $V_2(.)$，立刻用来算后续值.题目设定异步顺序是：**A → B → C**

- 对于状态 A，可能的行动为 A→B，则：

$$V_2(A) = -8 + \gamma \times V_1(B) = -8 + 0.5 \times 4 = -6$$

- 对于状态 B，可能的行动为 B→A 以及 B→C，其中 B→A 的行动价值为：

$$2 + \gamma \times V_2(A) = 2 + 0.5 \times (-6) = -1$$

B→C 的行动价值为：

$$-2 + \gamma \times V_1(C) = -2 + 0,5 \times 4 = 0$$

贪心策略取 max

$$V_2(B)=\max(-1,0)=0$$

- 对于状态 C，有两条行动路径，选择 C→B 的行动价值为：

$$8+\gamma\times V_2(B)=8+0.5\times 0=8$$

执行 ca 后，执行特殊跳转，以 $\frac{1}{4}$ 概率到 A，回报 4，后续价值 $\gamma\times V_2(A)$，以 $\frac{3}{4}$ 概率留在 C，回报 0，后续价值 $\gamma\times V_2(C)$。

因此，选择 C→A 的行动价值为

$$\frac{1}{4}\times(4+\gamma\times V_2(A))+\frac{3}{4}\times(0+\gamma\times V_1(C))=1.75$$

则：

$$V_2(C)=\max(8,1.75)=8$$

状态 **S**	V_1	V_2
A	4	-6
B	4	0
C	4	8

异步迭代后贪心策略 $\pi_2'(a|s)$

- 对于 A，最优策略只能为 A→B，因此

动作价值：

$$q_\pi(A,ab)=-8+0.6\times V_2(B)=-6$$

因为只有一种动作，所以

$$\pi_2'(ab|A)=1$$

- 对于B，有两种动作 B → A 和 B → C

$$q_\pi(B,ba)=2+0.5\times(-6)=-1$$
$$q_\pi(B.bc)=-2+0.5\times 8=2$$

比较两个值

$$q_\pi(B,ba)<q_\pi(B.bc)$$

所以确定性贪心策略选 B→C：

$$\pi_2'(bc|B)=1$$

- 对于 C，有两种动作 C → A 和 C → B

执行 ca 后，执行特殊跳转，以 $\dfrac{1}{4}$ 概率到 A，回报 4，后续价值 $\gamma \times V_2(A)$，以 $\dfrac{3}{4}$ 概率留在 C，回报 0，后续价值 $\gamma \times V_2(C)$，

$$q_\pi(C,ca) = \frac{1}{4} \times (4+\gamma{\times}V_2(A)) + \frac{3}{4} \times (0+\gamma{\times}V_2(C)) = 4.25$$
$$q_\pi(C,cb) = -8 + 0.5 \times V_2(B) = 8$$

比较两个值

$$q_\pi(C,ca) < q_\pi(C.cb)$$

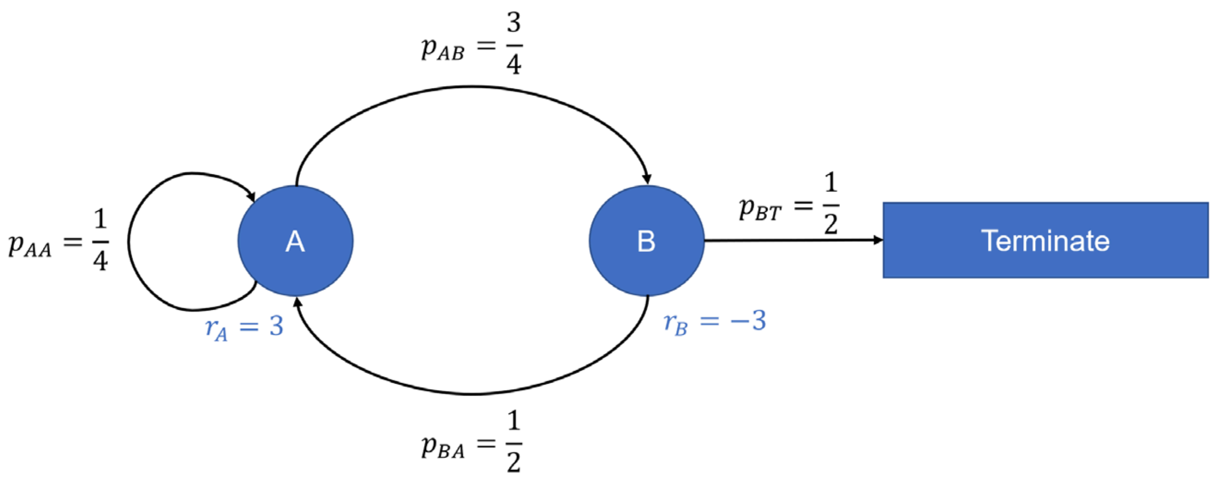
所以确定性贪心策略选 C→B：

$$\pi'_2(cb|C) = 1$$

综上所述：$\pi'_2(a|s)$ 中有 $\pi'_2(ab|A) = 1$、$\pi'_2(bc|B) = 1$、$\pi'_2(cb|C) = 1$ 其余均为 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 $v(A), v(B)$。



(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

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

其中 $A \overset{+3}{\to} A$ 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 $v(A), v(B)$。

（1）**贝尔曼期望方程：**

【方法一】

对于每个状态 s：

$$v(s)=\mathbb{E}[R_{t+1}+\gamma v(S_{t+1})|S_t=s]$$

- 对于 A

$$v_{k+1}(A)=p_{AA}\times(r_A+\gamma\times v_k(A))+p_{AB}\times(r_A+\gamma\times v_k(B))$$

- 对于 B

$$v_{k+1}(B)=p_{BA}\times(r_B+\gamma\times v_k(A))+p_{BT}\times(r_B+\gamma\times 0)$$

即对于马尔可夫回报过程有

$$\begin{aligned}v(A)&=r_A+\gamma(p_{AA}v(A)+p_{AB}v_k(B))\\ v(B)&=r_B+\gamma(p_{BA}v(B)+p_{BT}v_k(T))\end{aligned}$$

化简得

$$\begin{cases}v(A)=3+\dfrac{1}{4}v(A)+\dfrac{3}{4}v(B)\\ v(B)=-3+\dfrac{1}{2}v(A)\end{cases}\quad\Rightarrow\quad\begin{cases}v(A)=2\\ v(B)=-2\end{cases}$$

【方法二】

贝尔曼方程写成矩阵形式时，r 是即时奖励向量，P 是转移概率矩阵，得到稳态方程：

$$v=r+\gamma Pv\quad\Rightarrow\quad v=(I-\gamma P)^{-1}r$$

其中 $r=\begin{bmatrix}3\\-3\end{bmatrix}$, $\gamma=1,\ P=\begin{bmatrix}\frac{1}{4}&\frac{3}{4}\\ \frac{1}{2}&0\end{bmatrix}$

$$v=\begin{bmatrix}\dfrac{3}{4}&-\dfrac{3}{4}\\ -\dfrac{1}{2}&1\end{bmatrix}^{-1}\begin{bmatrix}3\\-3\end{bmatrix}=\begin{bmatrix}2\\-2\end{bmatrix}$$

即 $v(A)=2$，$v(B)=-2$（两种解法结果一样）

（2）已知两个观测片段，采用：

- 首次访问（First-Visit MC）
- 每次访问（Every-Visit MC）

【首次访问】

- 对片段 1：

$$G_1(A) = 3 + 2 - 4 + 4 - 3 = 2$$
$$G_1(B) = -4 + 4 - 3 = -3$$

- 对片段 2：

$$G_2(A) = 3 - 3 = 0$$
$$G_2(B) = -2 + 3 - 3 = -2$$

$$v(A) = \frac{G_1(A) + G_2(A)}{2} = 1, \quad v(B) = \frac{G_1(B) + G_2(B)}{2} = -2.5$$

【每次访问】

- 对片段 1：

$$\begin{aligned}
&G_{11}(A) = 3 + 2 - 4 + 4 - 3 = 2 \\
&G_{12}(A) = 2 - 4 + 4 - 3 = -1 \\
&G_{13}(A) = 4 - 3 = 1 \\
&G_{11}(B) = -4 + 4 - 3 = -3 \\
&G_{12}(B) = -3
\end{aligned}$$

- 对片段 2：

$$\begin{aligned}
&G_{21}(A) = 3 - 3 = 0 \\
&G_{21}(B) = -2 + 3 - 3 = -2 \\
&G_{23}(B) = -3
\end{aligned}$$

$$v(A) = \frac{G_{11}(A)G +_{12} (A) + G_{13}(A) + G_{21}(A)}{4} = \frac{2 - 1 + 1 + 0}{4} = 0.5$$

$$v(B) = \frac{G_{11}(B) + G_{12}(B) + G_{21}(B) + G_{22}(B)}{4} = \frac{-3 - 3 - 2 - 3}{4} = -2.75$$