3. 首次访问的蒙特卡洛预测

(1).

$v(A)=\frac{1}{2}[G_1+G_2]$.

$G_1=3+2-4+4-3=2$ $G_2=3-3=0$.

故 $v(A)=1$

$v(B)=\frac{1}{2}[G_1+G_2]$.

$G_1=-4+4-3=-3$ $G_2=-2+3-3=-2$

故 $v(B)=-2.5$

每次访问的蒙特卡洛预测:

$v(A)=\frac{1}{4}(G_1+G_2+G_3+G_4)$

$G_1=3+2-4+4-3=2$. $G_2=2-4+4-3=-1$ $G_3=4-3=1$. $G_4=3-3=0$

故 $v(A)=0.5$

$v(B)=\frac{1}{4}(G_1+G_2+G_3+G_4)$

$G_1=-4+4+3=3$ $G_2=-3$ $G_3=-2+3-3=-2$. $G_4=-3$.

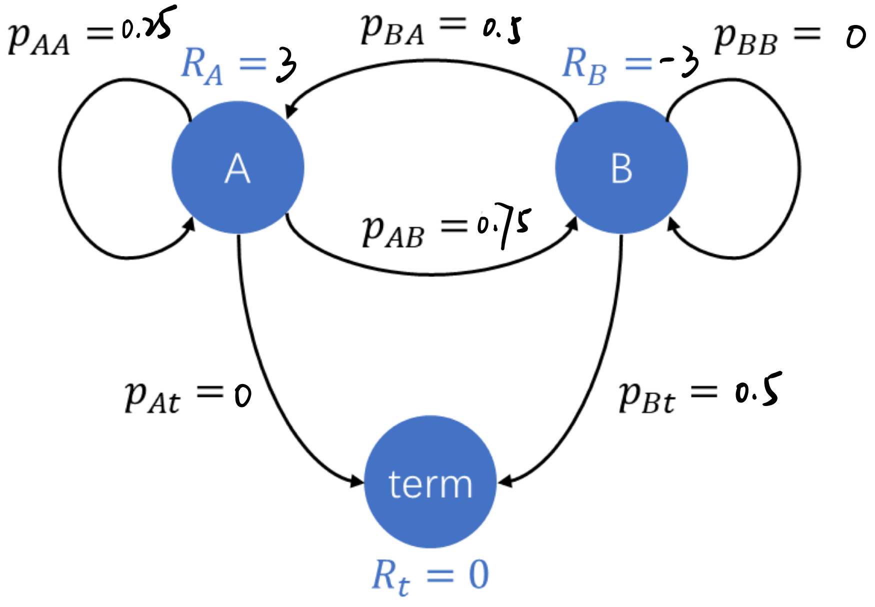
故 $v(B)=-2.75$

(2). $R_A=\frac{1}{4}(3+2+4+3)=3$. $R_B=\frac{1}{4}(-4-3-2-3)=-3$.

由统计可估计: $P_{AA}=0.25$. $P_{AB}=0.75$

$P_{BA}=0.5$. $P_{Bt}=0.5$.

故马尔可夫回报过程图应为:



(3). 由于:

$$\begin{bmatrix} v_A \\ v_B \\ v_t \end{bmatrix} = \begin{bmatrix} R_A \\ R_B \\ R_t \end{bmatrix} + r \cdot P \cdot \begin{bmatrix} v_A \\ v_B \\ v_t \end{bmatrix}$$

即:

$$\begin{bmatrix} v_A \\ v_B \\ v_t \end{bmatrix} = \begin{bmatrix} 3 \\ -3 \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{4} & \frac{3}{4} & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_A \\ v_B \\ v_t \end{bmatrix}.$$

由于 $v_t = R_t + v_t = v_t$

不妨设 $v(t) = 0$

则: $\begin{cases} v_A = 3 + \frac{1}{4} v_A + \frac{3}{4} v_B \\ v_B = -3 + \frac{1}{2} v_A + \frac{1}{2} v_t \end{cases} \Rightarrow \begin{cases} v_A = 2 \\ v_B = -2 \end{cases}$

4.

(1). 考虑每步: $4 \to 5$: $v(4) = v(4) + \alpha(R_t + rV(5) - V(4))$

$= 0.5(-1) = -0.5$

$5 \to 4$: $v(5) = v(5) + \alpha(R_t + rV(4) - v(5))$

$= 0.5(-1-0.5) = -0.75$

$4 \to 3$: $v(4) = v(4) + \alpha(R_t + rv(3) - v(4))$

$= -0.5 + 0.5(-1+0.5) = -0.75$

$3 \to term$: $v(3) = v(3) + \alpha(R_t + rV(term) - v(3))$

$= 0.5(-1) = -0.5$

V值:

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

(2). SARSA算法与Q-learning算法在贪心方式下结果相同.

$S_1 = 4$. A_1 = 向下,

$S_2 = 7$ A_2 = 向左

$Q(4,3) = Q(4,3) + \alpha(R_2 + rQ(7,4) - Q(4,3))$

$$= -2 - 1 - 2 + 2 = -3 .$$

$S_3 = 6$， $A_3 =$ 向上.

$$Q(7,4) = Q(7,4) + \alpha\big(R_3 + rQ(6,1) - Q(7,4)\big)$$

$$= -2 - 1 - 2 + 2 = -3$$

$S_4 = 3$， $A_4 =$ 向上.

$$Q(6,1) = Q(6,1) + \alpha\big(R_4 + rQ(3,1) - Q(6,1)\big)$$

$$= -2 - 1 - 1 + 2$$

$$= -2$$

$S_5 = term$.

$$Q(3,1) = Q(3,1) + \alpha\big(R_5 + rQ(term, a) - Q(3,1)\big)$$

$$= -1 - 1 + 1$$

$$= -1$$

综上，更新后的Q表：

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