No 1. (a)

$Q(3, left) = Q(3, left) + \alpha[r + \gamma Q(2, right) - Q(3, left)]$

则有 $7 + [-1 + 0.9 \times 8 - 7]$

得 $Q(3, left) = 6.2$

(b) ① 状态1 → right → 状态2

$Q(1, right) = Q(1, right) + \alpha[r + \gamma Q(2, up) - Q(1, right)]$

则有 $3 + 0.2[-1 + 0.9 \times 6 - 3]$

得 $Q(1, right)$ 更新为 3.16

② 状态2 → up → 状态5

$Q(2, up) = Q(2, up) + \alpha[r + \gamma Q(5, right) - Q(2, up)]$

则有 $6 + 0.2 \times [-1 + 0.8 \times 8 - 6]$

得 $Q(2, up)$ 更新为 5.88

③ 状态5 → right → 状态6

$Q(5, right) = Q(5, right) + \alpha[r + \gamma Q(6) - Q(5, right)]$

则有 $8 + 0.2 \times [10 + 0 - 8]$

得 $Q(5, right)$ 更新为 8.4

No2. 设?处状态为"S_3"

(a) $Q(S_3, \text{study}) = R + \gamma V = 10 + 1 \times 0 = 10$

$Q(S_3, \text{Pub}) = R + \gamma V$

$= 1.02 + 0.2V_1 + 0.4V_2 + 0.4V_3$

$= 1.84 + 0.4V_3$

$V_3 = \pi(S, \text{study}) \cdot Q(S_3, \text{study}) + \pi(\text{Pub}) \cdot Q(?, \text{Pub})$

$= 0.5 \times 10 + 0.5 \times (1.84 + 0.4V_3)$

解得 V_3 Ψ

(b) 由(a) $Q(S_3, \text{Study}) = 10$

$Q(S_3, \text{Pub}) = 1.84 + 0.4 \times 7.4 = 4.8$

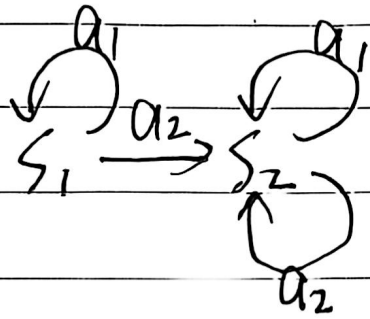
$Q(S, a) = R(S, a) + \gamma V_S$

S_1: $Q(S_1, \text{Play}) = -1 - 2.3 = -3.3$ $Q(S_1, \text{Study}) = -2 + 2.7 = 0.7$

S_2: $Q(S_2, \text{Sleep}) = 0 + 0 = 0$ $Q(S_2, \text{Study}) = -2 + \Psi = 5.4$

S_4: $Q(S_4, \text{Play}) = -1 - 2.3 = -3.3$ $Q(S_4, \text{Quit}) = 0 + 1.3 = -1.3$

3. 分析(1)问题

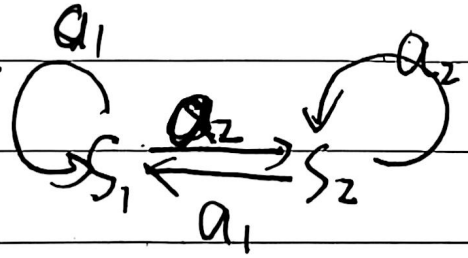


$S_1 \xrightarrow{a_2} S_2$ 时奖励1

(2) 问题转移函数同(1)

只在S_2执行任意动作时奖励1

(3) 问题及



S_1执行a_1, S_2执行a_2时奖励1

综上(1) MDP1 不存在无限大S_1值函数，最大在S_1选择a_2 $V(S_1$

MDP2 存在无限大策略，在S_1选择a_2

MDP3 存在无限大策略，在S_1选择a_1在S_2选择a_2

(4) 当$\gamma^2=1$，值函数收敛。

MDP1: $V(S_2)=0$ $Q(S_1,a_1)=\gamma^2 V(S_1)$

$Q(S_1,a_2)=\gamma^2 V(S_2)$

$V(S_1)=\max(\gamma^2 V(S_1), 1)=1$

最优策略：S_1选择a_2 S_2任意。

MDP2: $Q(s_1, a_1) = 0 + \gamma V(s_1)$ $Q(s_2, a_1) = 1 + \gamma^2 V(s_2)$

$Q(s_1, a_2) = 0 + \gamma^2 V(s_2)$ $Q(s_2, a_2) = 1 + \gamma^2 V(s_2)$

$V(s_2) = 1 + \gamma^2 V(s_2)$

$V(s_1) = \max(\gamma^2 V(s_1), \frac{\gamma^2}{1-\gamma^2})$ $V(s_2) = \frac{1}{1-\gamma^2}$

$= \frac{\gamma^2}{1-\gamma^2}$

最优策略：s_1 选择 a_2，s_2 任意。

MDP3: $Q(s_1, a_1) = 1 + \gamma^2 V(s_1)$ $Q(s_2, a_1) = \gamma^2 V(s_1)$

$Q(s_1, a_2) = \gamma^2 V(s_2)$ $Q(s_2, a_2) = 1 + \gamma^2 V(s_2)$

$V(s_1) = \max(\gamma V(s_2), 1 + \gamma V(s_1))$

$V(s_2) = \max(\gamma V(s_1), 1 + \gamma V(s_2))$

假设 $V(s_1) = 1 + \gamma V(s_1)$ 则 $V(s_1) = V(s_2) = \frac{1}{1-\gamma^2}$

$Q(s_1, a_2) = \frac{\gamma^2}{1-\gamma^2} < Q(s_1, a_1) = \frac{1}{1-\gamma^2}$ 选择 a_1

同理 s_2 选择 a_2

最优策略：s_1 选择 a_1

s_2 选择 a_2