1.(1) 状态 $S=\{s_0 \cdots s_7\}$ 表示石头被搬到的第8个点，行动为 $A=\{A_0, A_1\}$
A_0 为向上，A_1 为向下

(2) 此时大象向上和下概率相同，转移矩阵为

$$
\begin{array}{c|cccccccc|}
s \backslash s' & s_0 & & & & & & & s_7 \\
s_0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 \\
 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\
 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\
 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\
 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\
 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\
s_7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}
$$

(3) 贝尔曼方程为 $V_s^\pi = r_s^\pi + \gamma P_{ss'}^\pi V_{s'}^\pi$ 　　$r_s^\pi = [0, 0, 0, 0, 0, 0, \frac{0+700}{2}=350, 0]$ 　s_6处 r_s^π 为350，其余为0

$V_0^\pi = V_0^\pi$ 　$V_1^\pi = 0 + \frac{1}{2}V_2^\pi$ 　$V_2^\pi = \frac{V_1^\pi + V_3^\pi}{2}$，$V_3^\pi = \frac{V_2^\pi + V_4^\pi}{2}$，$V_4^\pi = \frac{V_3^\pi + V_5^\pi}{2}$，$V_5^\pi = \frac{V_4^\pi + V_6^\pi}{2}$

$V_6^\pi = \frac{V_5^\pi + V_7^\pi}{2} + 350$ 　$V_7^\pi = V_7^\pi$，6个有效方程，设 $V_0^\pi = 0$，$V_7^\pi = 0$，共8个方程

解得 $V_0^\pi = 0$ 　$V_1^\pi = 100$ 　$V_2^\pi = 200$，$V_3^\pi = 300$，$V_4^\pi = 400$ 　$V_5^\pi = 500$ 　$V_6^\pi = 600$ 　$V_7^\pi = 0$

(4) $q_\pi(s, a_0) = r_s^a + \gamma \sum P_{ss'}^a V_\pi(s')$ 　　向上行动价值为 $q_\pi(s, A_0) = r_s + V_{s+1}$ 　$r_s = \begin{cases} 700, & s = s_7 \\ 0 & s \neq s_7 \end{cases}$

解得 $q_\pi(s, A_0) = \{0, 200, 300, 400, 500, 600, 700\}$

同理，向下行动价值 $q_\pi(s, A_1) = r_s + V_{s-1}$ 　　$r_s = 0$

得 $q_\pi(s, a) = \{0, 0, 100, 200, 300, 400, 500, 0\}$

(5) 蒙特卡洛模拟方法，估计状态价值，之后可通过直接求解线性方程来解出最优策略。

(6) 采用深度强化学习模型，如 Deep-Q Network，优化策略模型，以获得较好的策略或逼近最优策略

3. $V_s^\pi = r_s^\pi + \gamma \sum P_{ss'}^\pi V_{s'}^\pi$，有 $V_A = (\frac{1}{4} \times 3 + \frac{3}{4} \times -3) + \frac{1}{4}V_A + \frac{3}{4}V_B = \frac{1}{4}V_A + \frac{3}{4}V_B - \frac{3}{2}$，同理 $V_B = \frac{1}{2}V_A + \frac{3}{2}$

解得 $V_A = -1$，$V_B = 1$

(2) 首次访问：$V(s_A) = \frac{1}{n}\sum G_i$ 　第一个片段 $G_{A0} = 3+2-4+4-3 = 2$ 　第二个 $G_{A1} = 3-3 = 0$

$V(s_A) = \frac{1}{2}(2+0) = 1$.

同理 第一个片段 $G_{B0} = 1-4+4-3 = -3$，第二个片段 $G_{B1} = -2+3-3 = -2$

$V(s_B) = \frac{1}{2}(-3-2) = -\frac{5}{2}$

每次访问 片段一有3次访问A，$G_{A0} = 3+2-4+4-3 = 2$，$G_{A1} = 2-4+4-3 = -1$，$G_{A2} = 4-3 = 1$

片段二有1次访问A 　$G_{A3} = 3-3 = 0$ 　　$V(s_A) = \frac{1}{4}(2-1+1+0) = \frac{1}{2}$

片段一有2次B：$G_{B0} = -4+4-3 = -3$ 　$G_{B1} = -3$，片段2有2次，$G_{B2} = -2+3-3 = -2$ 　$G_{B3} = -3$

$V(s_B) = \frac{1}{4}(-3-3-2-3) = -\frac{11}{4}$

扫描全能王 创建

2. 同步价值迭代 $V_{k+1}(s) = \max_{a \in A} \{ r^a_s + \gamma \sum_{s' \in S} P^a_{ss'} V_k(s') \}$ V_0 初始全为0

第一轮: $V_1(1) = \max(-8 + 0.5 \times 0) = -8$

$V_1(2) = \max(2 + 0.5 \times 0, -3 + 0.5 \times 0) = 2$

$V_1(3) = \max(0.25 \times 4 + 0.75 \times 0) + 0.5 \times (0.25 \times 0 + 0.75 \times 0), 8 + 0.5 \times 0) = 8$

第二轮 $V_2(1) = \max(-8 + 0.5 \times 2) = -7$

$V_2(2) = \max(2 + 0.5 \times (-8), -3 + 0.5 \times 8) = 1$

$V_2(3) = \max(0.25 \times 4 + 0.75 \times 0) + 0.5 \times (0.25 \times -3 + 0.75 \times 8), 8 + 0.5 \times 2) = 9$

A策略只能为 ab, $q(A, ab) = (8 + 0.5 \times 1) = -7.5$ $\pi(a|A)$ 为 $q(ab)$

B策略为 ba, bc $q(B, ba) = 2 + 0.5 \times -7 = -1.5$ $\pi(a|B)$ 为 $q(bc)$

$q(B, bc) = -3 + 0.5 \times 9 = 1.5$

C策略为 ca, cb $q(C, ca) = 0.25 \times 4 + 0.75 \times 0 + 0.5 \times (0.25 \times -1 + 0.75 \times 9) = 3.5$

$q(C, cb) = 8 + 0.5 \times 1 = 8.5$ $\pi(a|C)$ 为 0 或 1. $\delta(cb)$

采样价值迭代: $V_1(1) = \max(-8 + 0.5 \times 0) = -8$ 这里 $\delta(x) = \begin{cases} 1 & a \neq x \\ 0 & a = x \end{cases}$

$V_1(2) = \max(2 + 0.5 \times -8, -3 + 0.5 \times 0) = -2$

$V_1(3) = \max\{[1 + 0.5 \times (0.25 \times -8 + 0.75 \times 0)], 8 + 0.5 \times -2\} = 7$

$V_2(1) = \max(-8 + 0.5 \times -2) = -9$

$V_2(2) = \max(2 + 0.5 \times -9, -3 + 0.5 \times 7) = 0.5$

$V_3(3) = \max\{[1 + 0.5 \times (0.25 \times -9 + 0.75 \times 7)], 8 + 0.5 \times 0.5\} = 8.25$

$q(A, ab) = -8 + 0.5 \times 0.5 = -7.75$, $\pi(a|A) = \delta(ab)$

$q(B, ba) = 2 + 0.5 \times -9 = -2.5$

$q(B, bc) = -3 + 0.5 \times 8.25 = 1.125$ $\pi(a|B) = \delta(bc)$

$q(C, ca) = 1 + 0.5 \times (0.25 \times -9 + 0.75 \times 8.25) = 2.46875$

$q(C, cb) = 8 + 0.5 \times 0.5 = 8.25$ $\pi(a|C) = \delta(cb)$

扫描全能王 创建