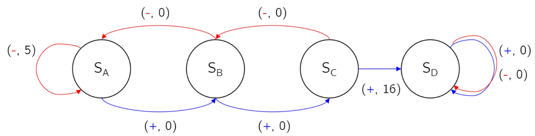

Figure 1: MDP problem with (action, reward) pairs.

Problem 1a [2 points]

Find the optimal policy at the initial state S_A with discount factor $\gamma = 0.001$. Justify your answer.

episode의 length가 3으로 제한되어 있으므로 모든 policy에 대한 utility를 직접 계산한다.

$\pi(S_A) = -$ reward

$S_A \xrightarrow{-} S_A \xrightarrow{-} S_A \longrightarrow S_A$ $\quad 5+\gamma\cdot 5+\gamma^2\cdot 5$

$S_A \xrightarrow{-} S_A \xrightarrow{-} S_A \xrightarrow{+} S_B$ $\quad 5+\gamma\cdot 5$

$S_A \xrightarrow{-} S_A \xrightarrow{+} S_B \xrightarrow{-} S_A$ $\quad 5$

$S_A \xrightarrow{-} S_A \xrightarrow{+} S_B \xrightarrow{+} S_C$ $\quad 5$

total utility of $\pi(S_A) = -$: $20+10\gamma+5\gamma^2$

value : $\dfrac{20+10\gamma+5\gamma^2}{4}$

$\pi(S_A) = +$

$S_A \xrightarrow{+} S_B \xrightarrow{+} S_C \xrightarrow{+} S_D$ $\quad 0+\gamma\cdot 0+\gamma^2\cdot 16$

$S_A \xrightarrow{+} S_B \xrightarrow{+} S_C \xrightarrow{-} S_B$ $\quad 0+\gamma\cdot 0+\gamma^2\cdot 0$

$S_A \xrightarrow{+} S_B \xrightarrow{-} S_A \xrightarrow{+} S_B$ $\quad 0+\gamma\cdot 0+\gamma^2 0$

$S_A \xrightarrow{+} S_B \xrightarrow{-} S_A \xrightarrow{-} S_A$ $\quad 0+\gamma\cdot 0+\gamma^2\cdot 5$

total utility of $\pi(S_A) = +$: $21\gamma^2$

value : $\dfrac{21\gamma^2}{4}$

when $\gamma = 0.001$, $\pi(S_A) = -$ $\Rightarrow$ $V_\pi(S_A) \approx \frac{20}{4} = 5$

$\pi(S_A) = +$ $\Rightarrow$ $V_\pi(S_B) \approx 0$

$\therefore$ $\pi_{opt}(S_A) = -$ when $\gamma = 0.001$

Problem 1b [2 points]

Find the optimal policy at the initial state S_A with discount factor $\gamma = 0.999$. Justify your answer.

When $\gamma = 0.999$, $\pi(S_A) = -$ $\Rightarrow$ $V_\pi(S_A) \approx 8.75$

$\pi(S_A) = +$ $\Rightarrow$ $V_\pi(S_B) \approx 5.24$

$\therefore$ $\pi_{opt}(S_A) = -$ when $\gamma = 0.999$

Problem 1c [2 points]

What is the optimal policy at the initial state S_B? Explain your answer in terms of discount factor $\gamma \in (0, 1)$.

$\pi(S_B) = -$ 인경우

reward

$S_B \xrightarrow{-} S_A \xrightarrow{-} S_A \xrightarrow{-} S_A$ $\quad 0 + \gamma \cdot 5 + \gamma^2 \cdot 5$

$S_B \xrightarrow{-} S_A \xrightarrow{-} S_A \xrightarrow{+} S_B$ $\quad 0 + \gamma \cdot 5 + \gamma^2 \cdot 0$

$S_B \xrightarrow{-} S_A \xrightarrow{+} S_B \xrightarrow{-} S_A$ $\quad 0 + \gamma \cdot 0 + \gamma^2 \cdot 0$

$S_B \xrightarrow{-} S_A \xrightarrow{+} S_B \xrightarrow{+} S_C$ $\quad 0 + \gamma \cdot 0 + \gamma^2 \cdot 0$

utility $= 10\gamma + 5\gamma^2$

Value $= \frac{10\gamma + 5\gamma^2}{4}$

$\pi(S_B) = +$ 인경우

reward

$S_B \xrightarrow{+} S_C \xrightarrow{+} S_D \xrightarrow{+} S_D \quad 0 + r \cdot 16 + \gamma^2 \cdot 0$

$S_B \xrightarrow{+} S_C \xrightarrow{+} S_D \xrightarrow{-} S_C \quad 0 + \gamma \cdot 16 + \gamma^2 \cdot 0$

$S_B \xrightarrow{+} S_C \xrightarrow{-} S_B \xrightarrow{+} S_C \quad 0 + \gamma \cdot 0 + \gamma^2 \cdot 0$

$S_B \xrightarrow{+} S_C \xrightarrow{-} S_B \xrightarrow{-} S_A \quad 0 + \gamma \cdot 0 + \gamma^2 \cdot 0$

utility $= 32r$
value $= 8r$

$0 < \gamma < 1, \qquad 8r > \dfrac{10\gamma + 5\gamma^2}{4}$

$\therefore \quad \pi_{opt}(S_B) = +$