Problem 1: MDP Warm-up

Consider an MDP problem. There are four states $\{S_A, S_B, S_C, S_D\}$, at each of which two actions $\{+, -\}$ are available, and the state transition and reward have no randomness. All the (action, reward) pairs are described in Figure 1. Assume all the episodes have length 3 (e.g. $S_A \xrightarrow{+} S_B \xrightarrow{-} S_A \xrightarrow{-} S_A$).

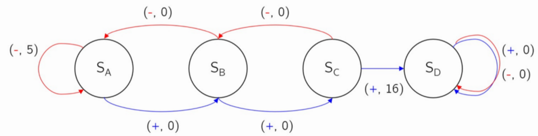

Figure 1: MDP problem with (action, reward) pairs.

Problem 1a [2 points]

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

Problem 1b [2 points]

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

Problem 1c [2 points]

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

① $\pi(S_A) = +$ 인 경우

episode는 다음과 같이 총 4가지가 나올수 있다.

$+ + + \Rightarrow 16\gamma^2$

$+ + - \Rightarrow 0$

$+ - + \Rightarrow 0$

$+ - - \Rightarrow 5\gamma^2$

value는 expected utility 이므로 $\frac{21\gamma^2}{4}$ 이다.

② $\pi(S_B) = -$ 인 경우

episode는 다음과 같이 총 4가지가 나올수 있다

$- + + \Rightarrow 5$

$- + - \Rightarrow 5$

$- - + \Rightarrow 5+5\gamma$

$- - - \Rightarrow 5+5\gamma+5\gamma^2$

value $= \frac{20+10\gamma+5\gamma^2}{4}$

Problem 1a

$\gamma = 0.001 \Rightarrow \frac{21(0.001)^2}{4} < \frac{20+10\times 0.001+5(0.001)^2}{4}$ 이므로 optimal policy = $-$

Problem 1b

$\gamma = 0.999 \Rightarrow \frac{21(0.999)^2}{4} < \frac{20+10\times 0.999+5(0.999)^2}{4}$ 이므로 optimal policy = $-$

S_B에 대해서도 마찬가지로 계산해보면

① $\pi(S_A) = +$ 인 경우

episode는 다음과 같이 총 4가지가 나올수 있다.

$+ + + \Rightarrow 16\gamma$

$+ + - \Rightarrow 16\gamma$

$+ - + \Rightarrow 0$

$+ - - \Rightarrow 0$

value $= \frac{32\gamma}{4}$

② $\pi(S_B) = -$ 인 경우

episode는 다음과 같이 총 4가지가 나올수 있다

$- + + \Rightarrow 0$

$- + - \Rightarrow 0$

$- - + \Rightarrow 5\gamma$

$- - - \Rightarrow 5\gamma + 5\gamma^2$

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

$\boxed{\text{Problem 1c}}$ $\gamma \in (0,1)$ 이면

$32\gamma > 10\gamma + 5\gamma^2$

$22\gamma > 5\gamma^2$

$\frac{22}{5} > \gamma$

즉 $\gamma \in (0,1)$ 이면 항상 $\frac{10\gamma + 5\gamma^2}{4}$가 $\frac{32}{4}\gamma$보다 작다.

$\Rightarrow$ optimal policy $= +$