Republic of Korea Air Force Academy

Intelligence Systems - Problem Sheet 1

Lecturer - Sungkwon On

October 2023

1. Briefly define the Markov Decision Process(MDP) with an aid of a diagram.
(다이어그램을 활용하여 Markov Decision Process(MDP)에 대해 서술하시오.)

2. Briefly define the Markov Sequence and state how it helps to simplify the MDP model used in Reinforcement Learning algorithms.
(Markov Sequence에 대해 서술하고, Markov Sequence가 어떻게 Reinforcement Learning에서 쓰이는 MDP 모델을 간략화 하는데 사용되는지 서술하시오.)

3. The following diagram shows a simplified procedure of an MDP. The two functions in the diagram, $f(x|c)$, and $g(y|d)$ both denote a probability function of each variable given each condition. State appropriate names of the functions and what the variables x, y, c, and d denotes. Note that a variable could be a vector with multiple elements.
(아래 다이어그램은 간략화된 MDP 절차를 보여준다. 두개의 함수는 확률 함수이며 각자 주어진 변수에 대한 확률을 각각에 해당하는 조건 변수에 따라 출력한다. 각 함수의 적절한 명칭을 제시하고, 각 변수 x, y, c, d가 무엇을 의미하는지 서술하라. 해당 변수는 요소가 하나 이상인 vector일 수도 있다.)

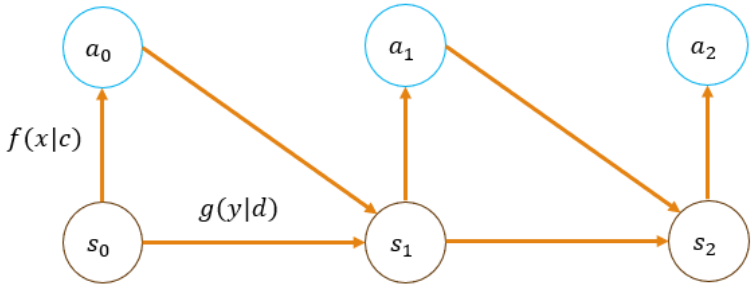

Figure 0.1: MDP Procedure

4. State both the verbal and mathematical definition of a ”Return” in the context of Reinforcement Learning. State the two roles of the discount rate and prove mathematically that the discount rate prevents the return from reaching infinity for an infinitely long sequence of rewards.
(Reinforcement Learning에서의 ”Return”의 사전적 정의와 수학적 정의를 제시하라. Discount rate(감가율)의 두가지 역할을 설명하고 감가율이 Return이 무한한 값으로 발산함을 막는 역할을 하는걸 수학적으로 증명하라.)

5. State both the verbal and mathematical definition of a Value function($V(s)$) and an Action-Value function(also called a Q function)($Q(s,a)$). Prove mathematically that $V(s)$ is an expectation of $Q(s,a)$ over all possible actions in the action space.

(Value function($V(s)$)과 Action-Value function(Q function이라고도 불리는)($Q(s,a)$)의 사전적 정의와 수학적 정의를 제시하라. $V(s)$가 모든 행동들에 대한 $Q(s,a)$의 기댓값이라는걸 수학적으로 증명하라.)

6. Consider the following diagram. The diagram shows a simple graphical model of an environment. At each state, there are only two possible actions, α and β. The actions are chosen based on the current policy defined as; $\pi(\alpha|S_B) = 0.7$ and uniform at all other states. The state transition probability is as follows;

(아래 그래프로 나타낸 환경 다이어그램을 참고하라. 모든 state에서는 두개의 선택 가능한 action들이(α, β) 있다. 해당 action들은 현재의 policy에 따라 선택된다. Policy는 다음과 같이 정의된다; $\pi(\alpha|S_B) = 0.7$ and uniform at all other states. State transition probability는 다음과 같다;)

$$p(S_D|S_B,\alpha) = 0.4, p(S_F|S_B,\beta) = 0.2, p(S_H|S_C,\alpha) = 1, p(S_I|S_C,\beta) = 0.1$$

a. Compute the values, V, at state S_A, S_B, and S_C

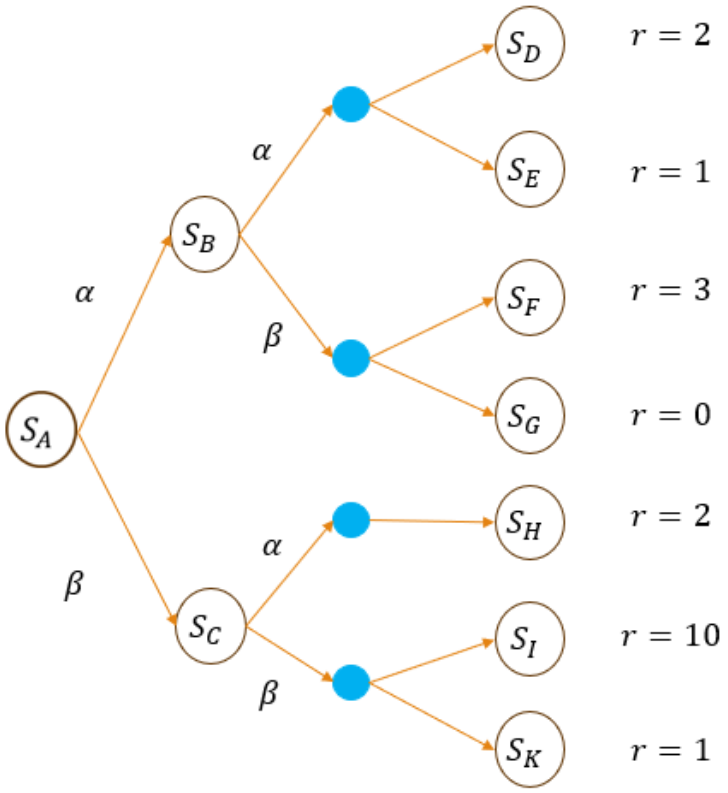

Figure 0.2

(State, S_A, S_B, S_C에서의 Value, V를 계산하라.)

b. Compute $Q(S_C,\alpha)$ and $Q(S_C,\beta)$. Hence, show their relationship with the value, $V(S_C)$.

($Q(S_C,\alpha)$ and $Q(S_C,\beta)$를 계산하라. 그리고 해당 값들과 $V(S_C)$의 상관관계에 대해 설명하라.)

7. Consider the following Bellman Equation;

(다음 공식을 참고하라.)

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{s_{t+1} \sim p_{\pi}(s_{t+1}|s_t, a_t)}[r_t + \gamma V^{\pi}(s_{t+1})|s_t, a_t]$$

From the above equation derive the following equation;

(위 공식을 활용하여 다음 공식을 유도하라.)

$$Q^{\pi}(S_t, a_t) = r_t + \mathbb{E}_{a_{t+1} \sim \pi(a_t|s_t), s_{t+1} \sim p_{\pi}(s_{t+1}|s_t, a_t)}[\gamma Q^{\pi}(s_{t+1}, a_{t+1})|s_t, a_t]$$

8. State two disadvantages of the Monte-Carlo(MC) algorithm and state how the Temporal Difference(TD) algorithm can overcome the disadvantages.

(Monte-Carlo(MC)알고리즘의 단점 두가지를 제시하고, 해당 단점들이 Temporal Difference(TD)알고리즘에서는 어떻게 해결되는지 설명하시오.)

9. State the difference between SARSA algorithm and Q-Learning algorithm. Hence also describe the differences between On-Policy and Off-Policy algorithms.

(SARSA 알고리즘과 Q-Learning 알고리즘의 차이를 설명하시오. 이여서 On-Policy와 Off-Policy알고리즘의 차이도 설명하시오.)