동아대학교 DONG-A UNIVERSITY

[실증적SW개발프로젝트]

RLHF기반 로봇 팔 제어 프로그램 개발

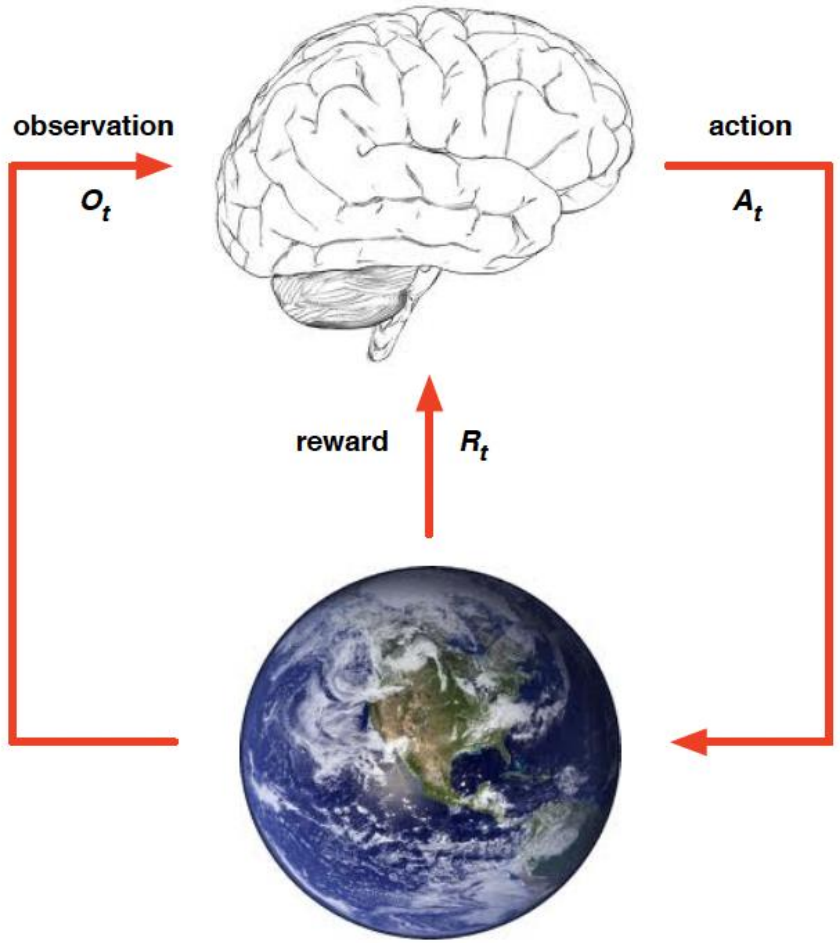
2143841 권은주
1824751 진현석
2051505 조현진

CONTENTS

1. RL introduction
2. Markov Decision Process
3. RLIF 논문 리뷰
4. 금주 활동내역

- **Agent and Environment**

  - ✓ Observation
  - ✓ Reward
  - ✓ Action



▪ History and State

$$H_t = O_1, R_1, A_1, ..., A_{t-1}, O_t, R_t$$

- State is the information used to determine what happens next

$$S_t = f(H_t)$$

▪ Information state

An information state (a.k.a. Markov state) contains all useful information from the history.

Definition

A state S_t is Markov if and only if

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, \ldots, S_t]$$

- "The future is independent of the past given the present"

- Agent의 구성요소

  - **Policy**(Agent 행동에 근거)

    - A policy is the agent's behaviour

  - **Value Function**

    - Value function is a prediction of future reward

$$v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s \right]$$

- Agent의 구성요소

- **Model(환경이 어떻게 될 지 예측하는 것)**
  - A model predicts what the environment will do next
  - Model Free
    - Policy and/or Value Function
    - No Model
  - Model Based
    - Policy and/or Value Function
    - Model

▪ Markov Property

An information state (a.k.a. Markov state) contains all useful information from the history.

Definition

A state S_t is Markov if and only if

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, ..., S_t]$$

- "The future is independent of the past given the present"

▪ Markov Process

A Markov process is a memoryless random process, i.e. a sequence of random states $S_1, S_2, \ldots$ with the Markov property.

Definition

A *Markov Process* (or *Markov Chain*) is a tuple $\langle \mathcal{S}, \mathcal{P} \rangle$

- $\mathcal{S}$ is a (finite) set of states
- $\mathcal{P}$ is a state transition probability matrix,
  $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s]$

$$\mathcal{P} = \text{from} \begin{bmatrix} & & \text{to} \\ \mathcal{P}_{11} & \cdots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{n1} & \cdots & \mathcal{P}_{nn} \end{bmatrix}$$

▪ Markov Reward Process

A Markov reward process is a Markov chain with values.

Definition

A *Markov Reward Process* is a tuple $\langle \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

- $\mathcal{S}$ is a finite set of states
- $\mathcal{P}$ is a state transition probability matrix, $\mathcal{P}_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s]$
- $\mathcal{R}$ is a reward function, $\mathcal{R}_s = \mathbb{E}[R_{t+1} \mid S_t = s]$
- γ is a discount factor, $\gamma \in [0, 1]$

▪ Value Function

An information state (a.k.a. Markov state) contains all useful information from the history.

Definition

A state S_t is Markov if and only if

$$\mathbb{P}[S_{t+1} \mid S_t] = \mathbb{P}[S_{t+1} \mid S_1, ..., S_t]$$

- "The future is independent of the past given the present"

- **Bellman Equation for MRP**

$$
\begin{aligned}
v(s) &= \mathbb{E}\left[G_t \mid S_t = s\right] \\
&= \mathbb{E}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s\right] \\
&= \mathbb{E}\left[R_{t+1} + \gamma\left(R_{t+2} + \gamma R_{t+3} + \ldots\right) \mid S_t = s\right] \\
&= \mathbb{E}\left[R_{t+1} + \gamma G_{t+1} \mid S_t = s\right] \\
&= \mathbb{E}\left[R_{t+1} + \gamma v(S_{t+1}) \mid S_t = s\right]
\end{aligned}
$$

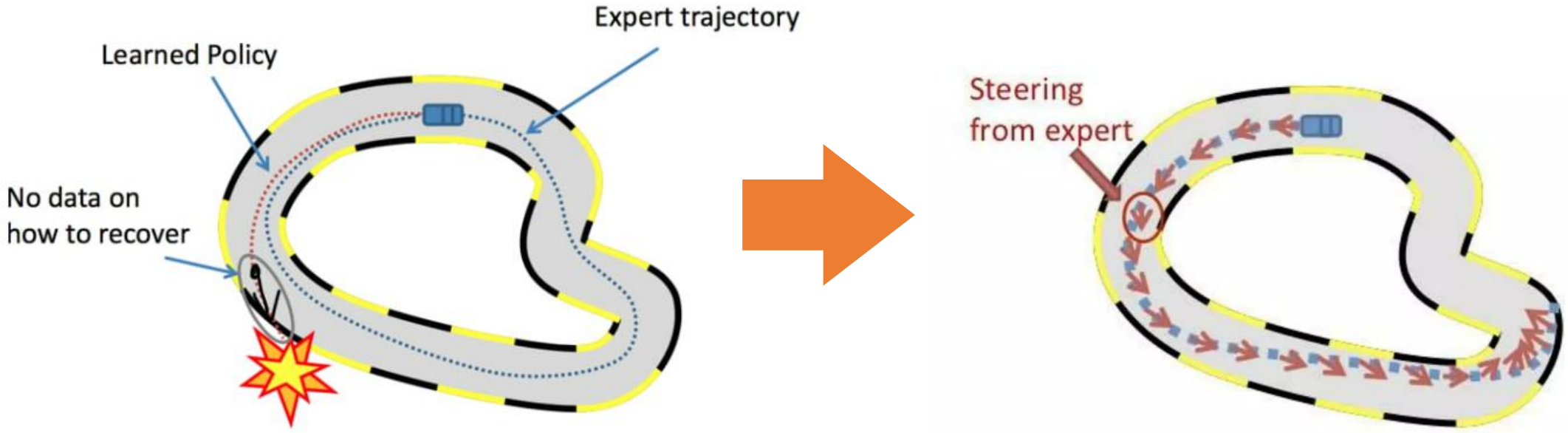
▪ Markov Decision Process

Definition

A *Markov Decision Process* is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$

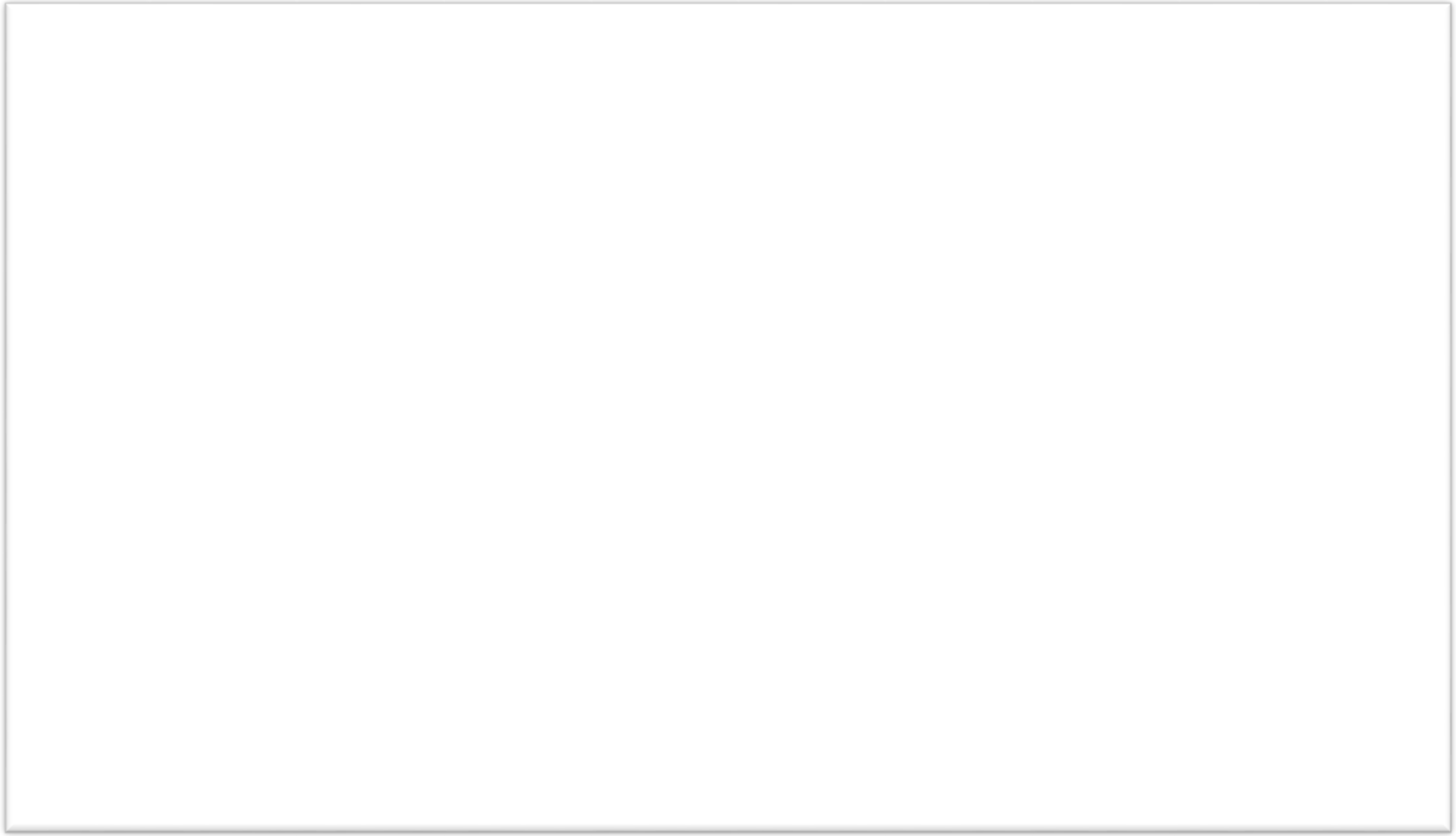
- $\mathcal{S}$ is a finite set of states
- $\mathcal{A}$ is a finite set of actions
- $\mathcal{P}$ is a state transition probability matrix, $\mathcal{P}^{a}_{ss'} = \mathbb{P}\left[S_{t+1} = s' \mid S_t = s, A_t = a\right]$
- $\mathcal{R}$ is a reward function, $\mathcal{R}^{a}_{s} = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$
- γ is a discount factor $\gamma \in [0, 1]$.

- BC(Behavior Cloning), DAgger(Dataset Aggregation)
  - 강화학습의 보상함수 설계에 대한 어려움을 극복하기 위해 나온 방법
  - BC: expert의 Demo Trajectory를 학습시키는 방법
  - DAgger: expert가 학습된 정책을 관찰하고 개입하여 수정을 제공하는 방법



BC DAgger

- **RLIF: INTERACTIVE IMITATION LEARNING AS REINFORCEMENT LEARNING**
  - **인간이 하는 개입 자체는 정보가 부족하고 최적이 아님**
  - **사용자 개입 신호를 reward로 사용하는 방식**



- **RLIF: INTERACTIVE IMITATION LEARNING AS REINFORCEMENT LEARNING**
  - 사용자 개입으로 이어지는 행동에 부정적인 보상(-1 reward)을 부여함
  - Agent는 현재의 policy(현재 state에서 한 행동)에 대해 피드백을 받을 수 있음

Algorithm 1 Interactive imitation

Require: π, π^{exp}, D

for trial $i = 1$ to N **do**
  Train π on D via supervised learning
  for timestep $t = 1$ to T **do**
    if π^{exp} intervenes at t **then**
      append $(s_t, a_t^{\pi^{\text{exp}}})$ to D_i
    end if
  end for
  $D \leftarrow D \cup D_i$
end for

Dagger의 Pseudocode

Algorithm 2 RLIF

Require: π, π^{exp}, D

for trial $i = 1$ to N **do**
  Train π on D via reinforcement learning.
  for timestep $t = 1$ to T **do**
    if π^{exp} intervenes at t **then**
      label (s_{t-1}, a_{t-1}, s_t) with -1 reward, append to D_i
    else
      label (s_{t-1}, a_{t-1}, s_t) with 0 reward, append to D_i
    end if
  end for
  $D \leftarrow D \cup D_i$
end for

RLIF의 Pseudocode

- **RLIF: INTERACTIVE IMITATION LEARNING AS REINFORCEMENT LEARNING**
  - **전문가가 언제 개입을 선택하느냐에 따라 결과가 달라짐**
  - **Ground truth Reward signal(행동에 의해 발생하는 실제 보상)이 필요 없음**

실증적AI프로젝트 금주 활동내역

주제: RLHF를 이용한 협동 로봇 제어 프로그램 개발

금주 활동계획	**1. 강화학습 1~2주차 정리, 5주차 학습 및 RLIF 논문 리뷰**		
	팀장 (권은주)	팀원 1 (조현진)	팀원 2 (진현석)
금주 개인별 활동내역	1. 1~2주차 개념정리 2. 5주차 학습 및 논문리뷰 3. 활동내역을 바탕으로 한 스터디 진행	1. 1~2주차 개념정리 2. 5주차 학습 및 논문리뷰 3. 활동내역을 바탕으로 한 스터디 진행	1. 1~2주차 개념정리 2. 5주차 학습 및 논문리뷰 3. 활동내역을 바탕으로 한 스터디 진행
차주 활동계획	1. Prof. David Silver 강화학습 3~4주차 정리 후 스터디 2. RLHF 논문 리뷰		

QUESTIONS & ANSWERS

Dept. of AI, Dong-A University

권은주 (kkkoj4284@donga.ac.kr)

진현석 (cpu132465@donga.ac.kr)

조현진 (gkfkgkdh@naver.com)

Github (https://github.com/eunjuyummy/AI_Project_CoRLHF)