A Bayesian Approach to Generative Adversarial Imitation Learning

NeurIPS 2018

Presenter

Wonseok Jeon @ KAIST

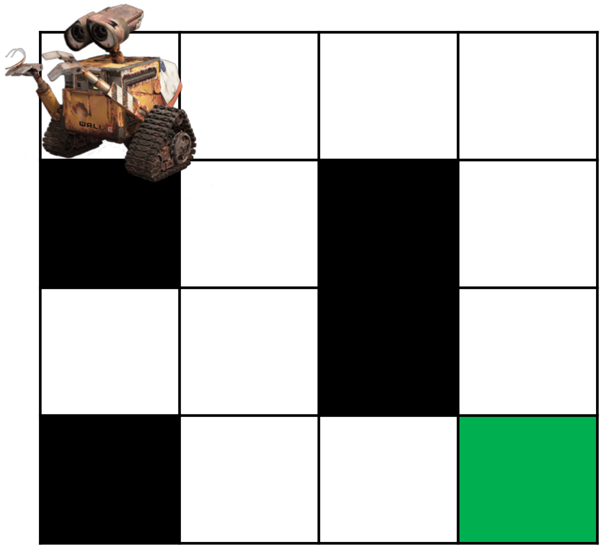
Joint work with

Seokin Seo @ KAIST

Kee-Eung Kim @ KAIST & PROWLER.io

Imitation Learning

- A **Markov decision process** (MDP) $\langle \mathcal{S}, \mathcal{A}, P(s'|s,a), \cancel{c(s,a)} \rangle$
without cost
- A **policy** $\pi(a|s)$

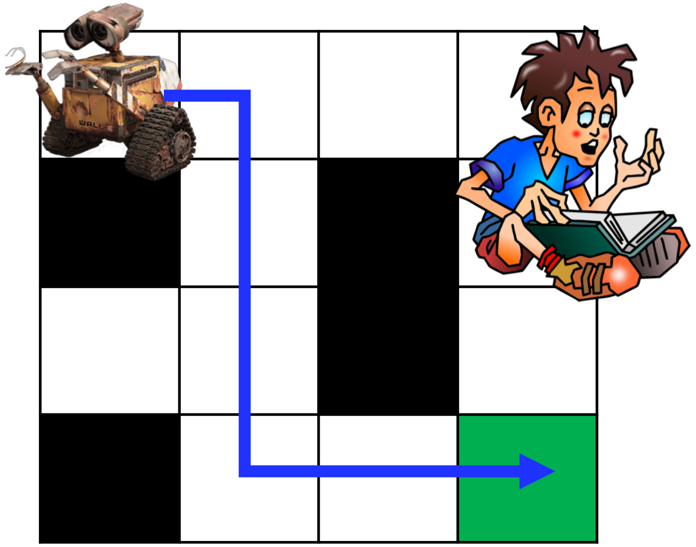


Imitation Learning

- A **Markov decision process** (MDP) $\langle \mathcal{S}, \mathcal{A}, P(s'|s,a), \cancel{c(s,a)} \rangle$
  without cost
- A **policy** $\pi(a|s)$
- Instead, there is a set of **expert's demonstrations**:

$$\{(s_1, a_1, \ldots, s_T)\} \sim \pi_E(a|s)$$

- Learn a policy that mimics $\pi_E(a|s)$ well.



Generative Adversarial Imitation Learning (GAIL)

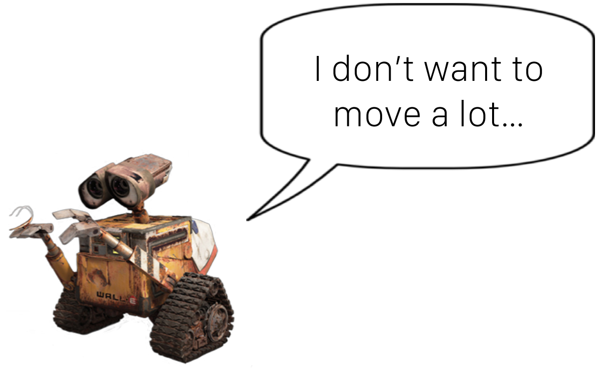
- Use **generative adversarial networks (GANs)** for imitation learning:

$$\min_{\pi} \max_{D} \mathbb{E}_{\pi}\left[\sum_{t=1}^{T} \log D(s_t, a_t)\right] + \mathbb{E}_{\pi_E}\left[\sum_{t=1}^{T} \log(1 - D(s_t, a_t))\right]$$

1. Sample trajectories by using $\pi(a|s)$ and $\pi_E(a|s)$ (expert demonstrations).
2. Train discriminator.
3. Update policy $\pi(a|s)$ by using reinforcement learning (RL), e.g., TRPO, PPO.

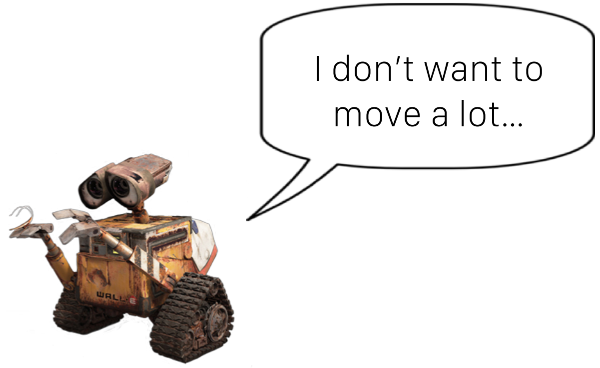
Generative Adversarial Imitation Learning (GAIL)

- GAIL requires model-free RL inner loops.
  - The environment simulation is required.
- Sample-efficiency issues
  - Obtaining trajectory samples from the environment is often very costly, e.g., physical robots in a real world.



Generative Adversarial Imitation Learning (GAIL)

- GAIL requires model-free RL inner loops.
  - The environment simulation is required.
- Sample-efficiency issues
  - Obtaining trajectory samples from the environment is often very costly, e.g., physical robots in a real world.
- **Motivation**
  - For each iteration, the discriminator is updated by using minibatches.
  - How about using Bayesian classification to train discriminator?
    - Expected to make more refined cost function for imitation learning!



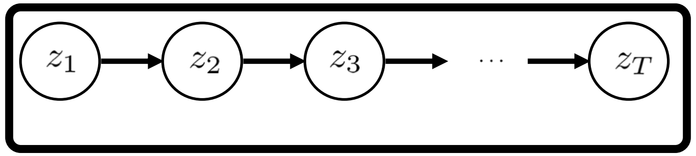
Bayesian Framework for GAIL

- Probabilistic model for trajectories
  - For each trajectories $\tau = (s_1, a_1, s_2, a_2, \ldots, s_T, a_T)$, a sequence of state-action pairs satisfies **Markov property**:

$$p(s_1, a_1) = p(s_1)\pi(a_1|s_1),$$
$$p(s_{t+1}, a_{t+1}|s_t, a_t) = P_T(s_{t+1}|s_t, a_t)\pi(a_{t+1}|s_{t+1})$$

$$z = (s, a)$$

trajectory



τ

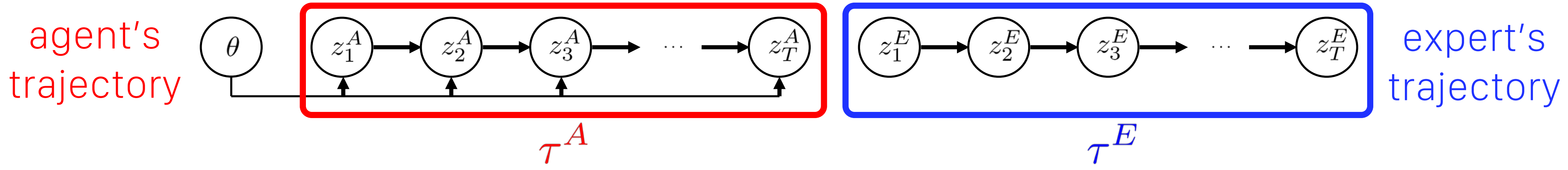
Bayesian Framework for GAIL

- Probabilistic model for trajectories
  - For each trajectories $\tau = (s_1, a_1, s_2, a_2, \ldots, s_T, a_T)$, <u>a sequence of state-action pairs</u> satisfies **Markov property**:

$$p(s_1, a_1) = p(s_1)\pi(a_1|s_1),$$
$$p(s_{t+1}, a_{t+1}|s_t, a_t) = P_T(s_{t+1}|s_t, a_t)\pi(a_{t+1}|s_{t+1})$$

  - Two policies: agent's policy $\pi_\theta(a|s)$, expert's policy $\pi_E(a|s)$

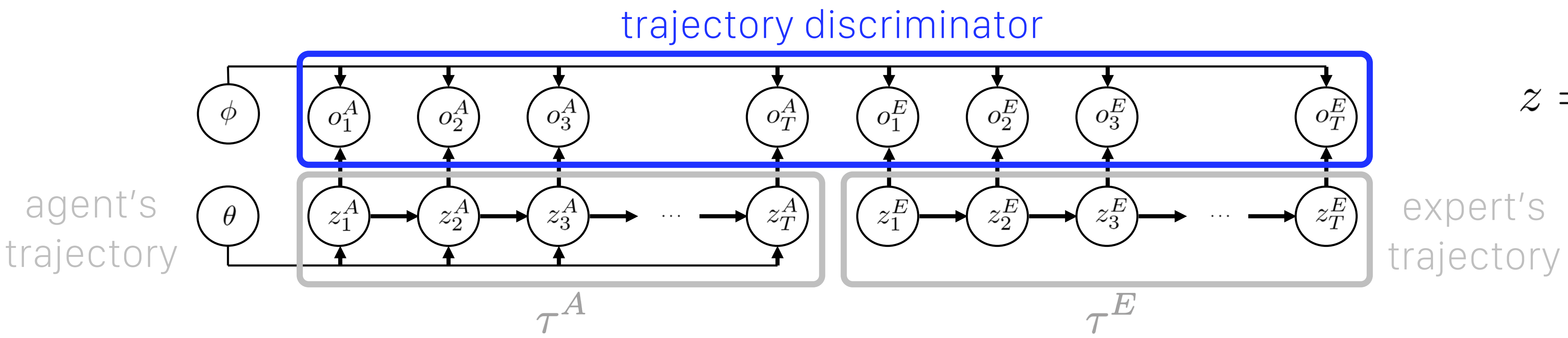
$$z = (s, a)$$



Bayesian Framework for GAIL

- Role of discriminator
  - The probability that models whether (s, a) comes from the expert $(o = 1)$ or the agent $(o = 0)$

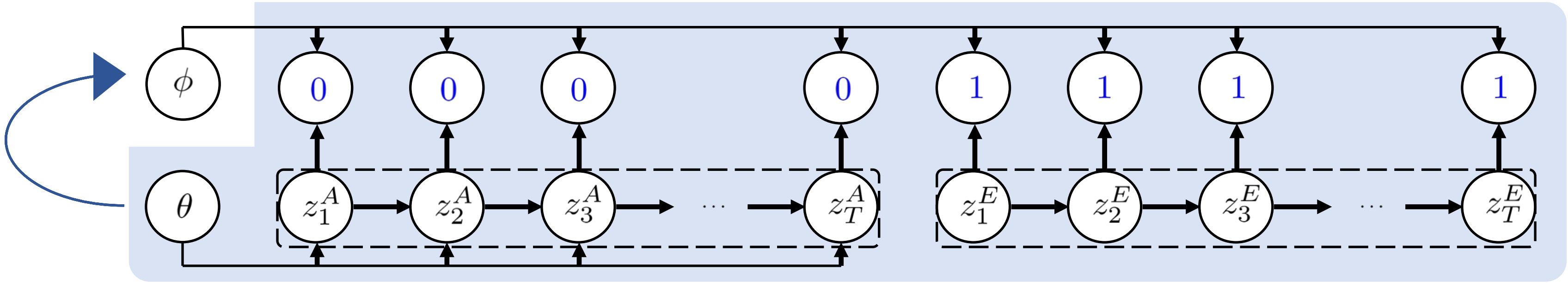
$$p_\phi(o|z) = \begin{cases} 1 - D_\phi(z), & \text{if } o = 1, \\ D_\phi(z), & \text{if } o = 0. \end{cases}$$



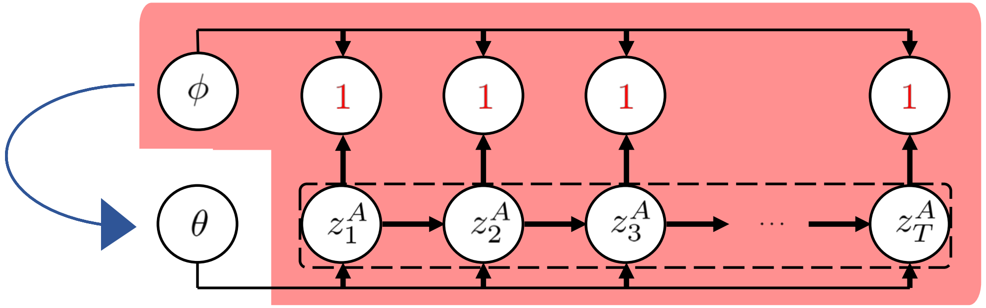
$z = (s, a)$

Bayesian Framework for GAIL

- Posterior distributions
  - Posterior for discriminator (conditioned on perfect trajectory discrimination)



  - Posterior for policy (conditioned on preventing perfect discrimination)

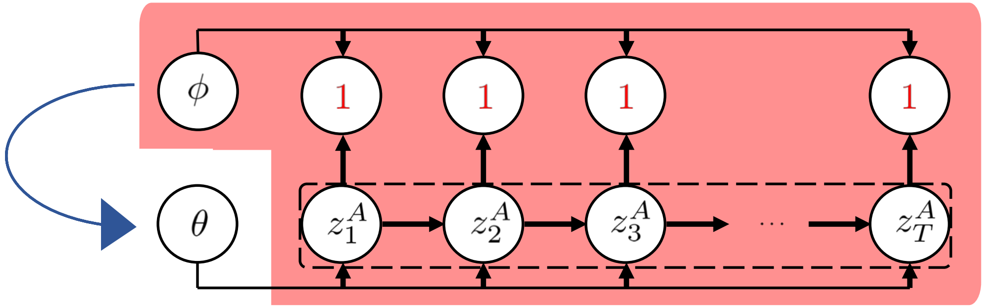


Bayesian Framework for GAIL

- Posterior distributions
  - Posterior for discriminator (conditioned on perfect trajectory discrimination)

GAIL uses maximum likelihood estimation (MLE) for both policy and discriminator updates!

  - Posterior for policy (conditioned on preventing perfect discrimination)



Bayesian GAIL: GAIL with Posterior-Predictive Cost

- The objective is

reinforcement learning

$$\underset{\theta}{\operatorname{argmax}}\ \mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T} \mathbb{E}_{p_{\text{posterior}}(\phi)} \log D_\phi(s_t, a_t)\right]$$

posterior-predictive cost

Bayesian GAIL: GAIL with Posterior-Predictive Cost

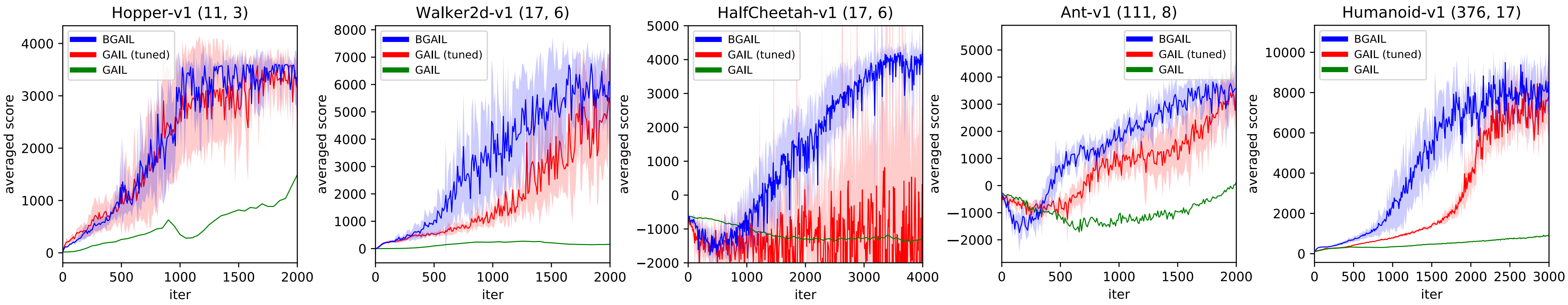
- The objective is

reinforcement learning

$$\underset{\theta}{\operatorname{argmax}}\ \mathbb{E}_{\pi_\theta}\left[\sum_{t=1}^{T}\mathbb{E}_{p_{\text{posterior}}(\phi)}\log D_\phi(s_t, a_t)\right]$$

posterior-predictive cost

- Learning Curve for 5 MuJoCo tasks!



Bayesian GAIL: GAIL with Posterior-Predictive Cost

ive

For more information, please come to our poster session!

Wed Dec 5th 5-7 PM @ Room 210 & 230 AB #129

Thanks!