Brain Inspired Artificial Intelligence
8: Sample-efficient generative adversarial imitation learning

Eiji Uchibe

Dept. of Brain Robot Interface

ATR Computational Neuroscience Labs.

Problems of GAIL and AIRL

- GAIL is sample-efficient with respect to the number of demonstrations
  - GAIL outperformed naïve behavior cloning
- However, GAIL's generator adopts on-policy reinforcement learning (TRPO and Proximal Policy Optimization). GAIL is NOT sample-efficient with respect to the number of interactions with the environment
  - It takes a long time to find an optimal policy

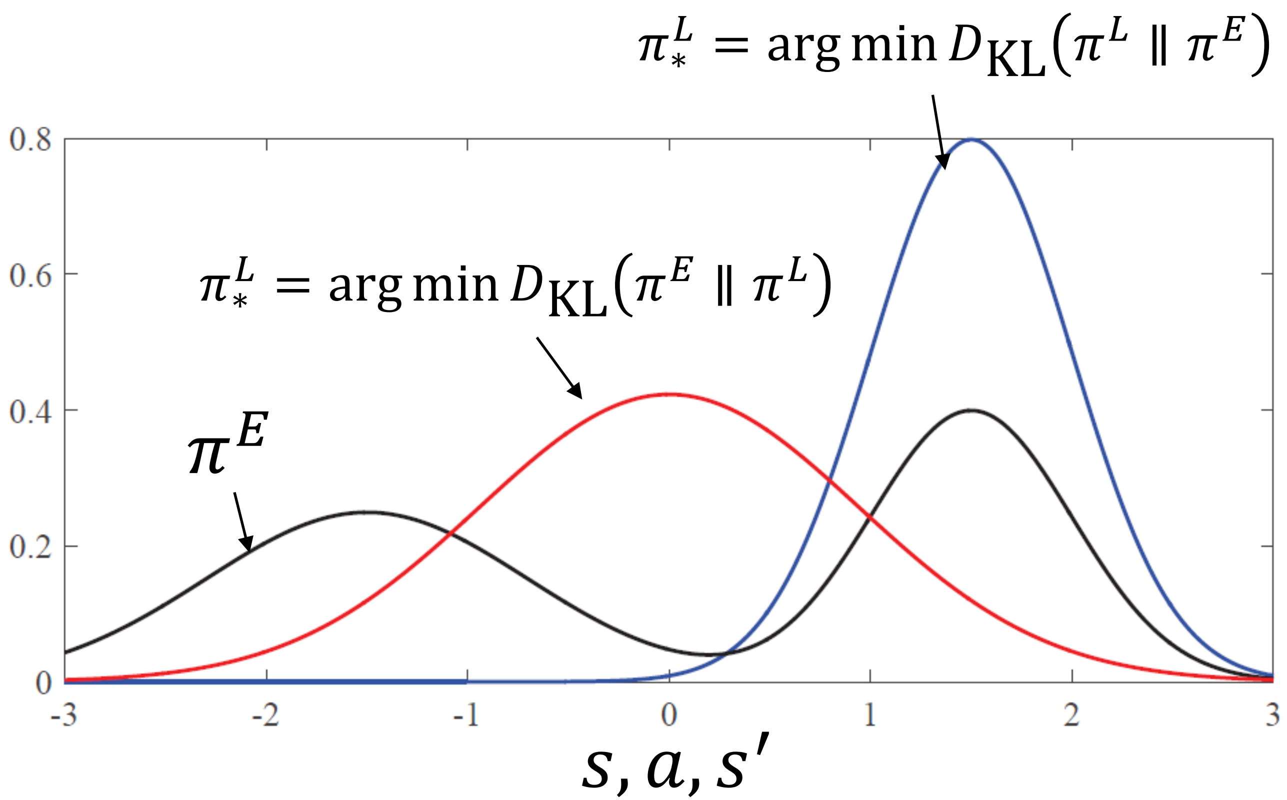
Formulation

- The goal is to minimize the Kullback-Leibler (KL) divergence

$$J(\pi^L) = D_{\text{KL}}(\pi^L \parallel \pi^E) = \int \pi^L(s,a,s') \ln \frac{\pi^L(s,a,s')}{\pi^E(s,a,s')} \mathrm{d}s\mathrm{d}a\mathrm{d}s'$$

  - π^E: (unknown) expert's distribution. We have samples from π^E
  - π^L: learner's distribution
  - π^L/π^E is unknown

- Note: minimizing $D_{\text{KL}}(\pi^E \parallel \pi^L)$ is identical to Behavior Cloning (BC)



Basic idea

- Estimate the log density ratio from samples, and minimize the approximated KL divergence

$$J(\pi^L) = \int \pi^L(s,a,s') \ln \frac{\pi^L(s,a,s')}{\pi^E(s,a,s')} \mathrm{d}s\mathrm{d}a\mathrm{d}s'$$

$$\approx \int \pi^L(s,a,s') \ln \frac{1 - D(s,a,s')}{D(s,a,s')} \mathrm{d}s\mathrm{d}a\mathrm{d}s'$$

density ratio trick
[Sugiyama et al., 2012]

  – $D(s,a,s') = \Pr(\text{Expert} \mid s,a,s')$ is a discriminator
  – The structure of $D(s,a,s')$ is determined by entropy-regularized RL
- Density ratio estimation by logistic regression ➡ Inverse RL
- Minimizing the KL divergence ➡ forward RL

Uchibe, E. (2019). Imitation learning based on entropy-regularized forward and inverse reinforcement learning. Proc. of RLDM.

Inverse RL as density ratio estimation

- The joint distribution can be decomposed under the Markovian assumption

$$\frac{\pi^E(s,a,s')}{\pi^L(s,a,s)} = \overbrace{\frac{p_T(s' \mid s,a)}{p_T(s' \mid s,a)}}^{\text{ratio of state transition}} \times \overbrace{\frac{\pi^E(\,a \mid s\,)}{\pi^L(\,a \mid s\,)}}^{\text{ratio of policies}} \times \frac{\pi^E(s)}{\pi^L(s)}$$

$$\frac{D_k(s,a,s')}{1 - D_k(s,a,s')} \triangleq f_k(s,a,s') \qquad \frac{D_k(s)}{1 - D_k(s)} \triangleq g_k(s)$$

- Two density ratio terms should be estimated

Uchibe, E. (2019). Imitation learning based on entropy-regularized forward and inverse reinforcement learning. Proc. of RLDM.

Entropy-regularized reinforcement learning

- Assumption: the reward function is given by

$$r(s,a) = r_k(s) + \kappa^{-1}\mathcal{H}(\pi) - \eta^{-1} D_{\mathrm{KL}}\left(\pi \parallel \pi_k^L\right)$$

  - $\mathcal{H}(\pi)$: entropy of policy π.
  - $\mathrm{KL}(\pi \parallel \pi_k^L)$: KL divergence between the leaning policy and π_k^L
  - $r_k(x)$: reward function to be estimated
  - κ, η: hyper parameters (Kozuno et al., 2019)
  - $\eta \to \infty$: Soft Q-learning, Soft Actor-Critic (Haarnoja et al., 2018)
  - $\kappa \to \infty$: Dynamic Policy Programming (Azar et al., 2012)

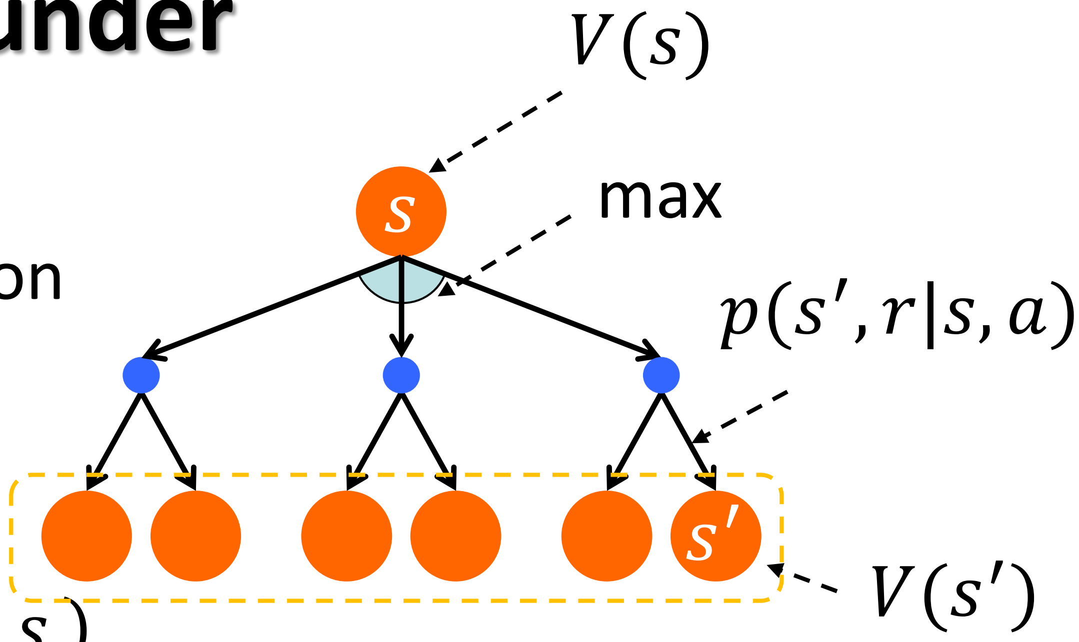
Kozuno, T., Uchibe, E., and Doya, K. (2019). Theoretical analysis of efficiency and robustness of softmax and gap-increasing operators in reinforcement learning. In Proc. of AISTATS.

Bellman optimality equation under entropy regularization

- Relation of the optimal state value function
- The max operator can be solved analytically

$$\frac{1}{\beta}\ln\frac{\pi^E(a \mid s)}{\pi_k^L(a \mid s)} = r_k(s) - \kappa^{-1}\ln\pi_k^L(a \mid s) + \gamma\mathbb{E}_{s'\sim p_T(\cdot|s,a)}[V_k(s')] - V_k(s)$$

$$\beta = \frac{\kappa\eta}{\kappa+\eta}$$



- The log density ratio is represented by the reward, the difference of the state value function, and the policy

Structured discriminator

- Use the previous relations

$$D_k(s,a,s') = \frac{\exp\big(\beta f_k(s,a,s')\big)}{\exp\big(\beta f_k(s,a,s')\big) + \exp\big(\beta\kappa \ln \pi_k^G(\, a \mid s\,)\big)}$$

  – where $f_k(s,a,s') = r_k(s) - \beta^{-1} g_k(s) + \gamma V_k(s') - V_k(s)$

- Relation to previous studies
  – AIRL (Fu et al., 2018): $g_k(s) = 0$ and $\beta = 1, \kappa = 1$
  – LogReg-IRL (Uchibe, 2018): $\kappa = 0$

Uchibe, E. (2019). Imitation learning based on entropy-regularized forward and inverse reinforcement learning. Proc. of RLDM.

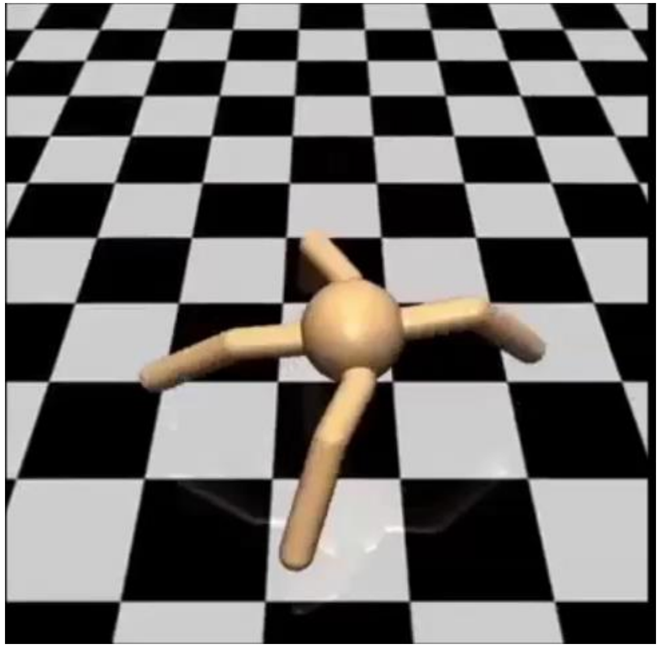
Forward RL as minimizing KL divergence

- Update the baseline policy by minimizing the KL divergence estimated by density ratio estimation

$$\pi_{k+1}^{L} = \arg\min_{\pi^L} \mathbb{E}_{\pi^L}\left[\ln \frac{1 - D(s, a, s')}{D(s, a, s')}\right] = \arg\max_{\pi^L} \mathbb{E}_{\pi^L}\left[\sum_t \gamma^t \tilde{r}(s_t, a_t)\right]$$

  - Identical to optimization of entropy-regularized RL

Experiments: MuJoCo Benchmarks

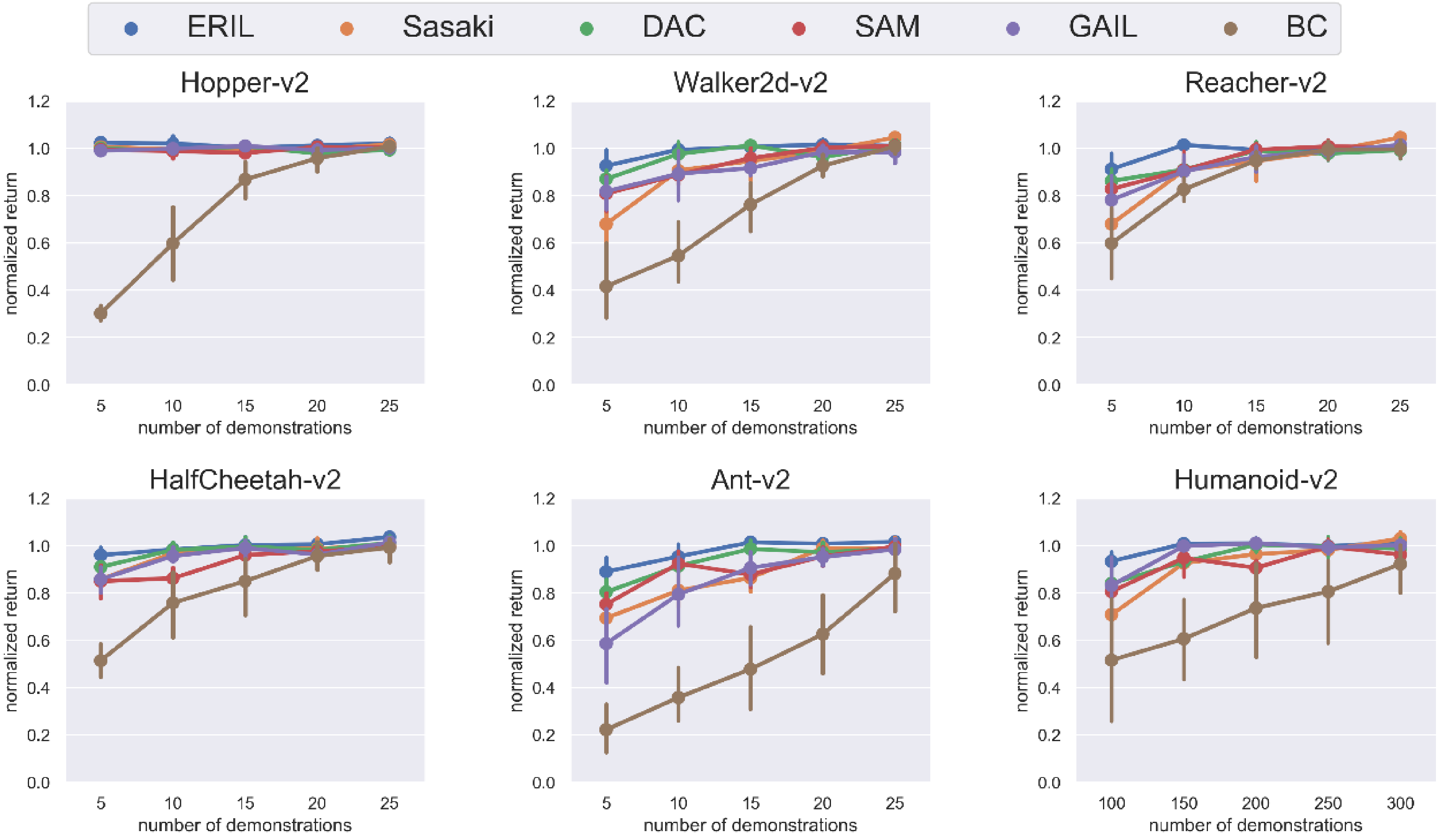


- Task: move as fast as possible
- Original reward function

$$r_t = v_t - c||\boldsymbol{a}_t||_2$$

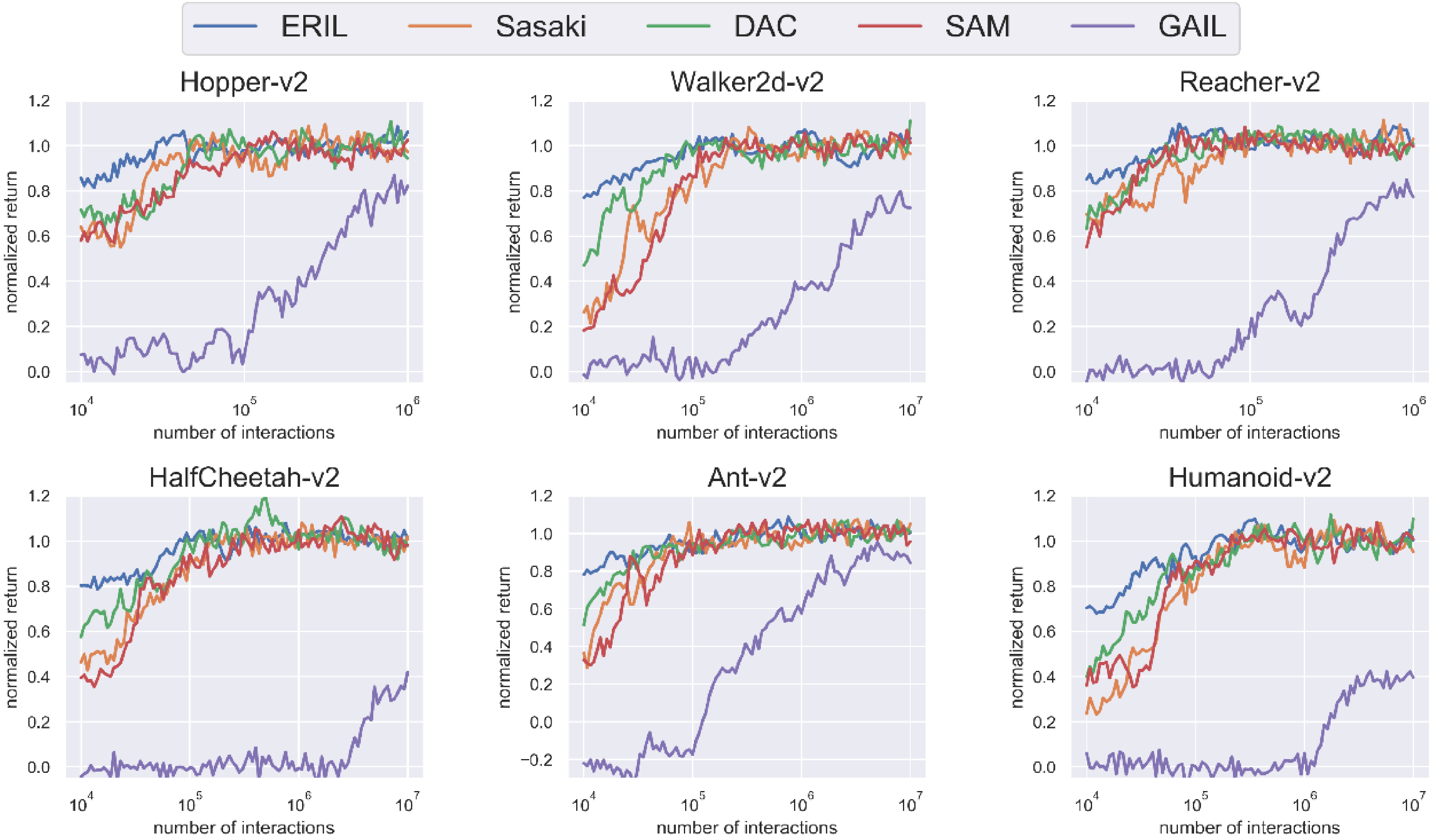
  - v_t: forward velocity. c: robot-specific parameter
- Expert policy
  - Trained by on-policy Trust Region Policy Optimization (TRPO) (Schulman et al., 2015)
- Compare this method with the following methods
  - BC: Behavior Cloning, GAIL
  - (Sasaki et al., 2019), Discriminator-Actor-Critic (Kostrikov et al., 2019), Sample-efficient Adversarial Mimic (Blondé, et al., 2019)

Sample-efficiency w.r.t. the number of experts



Uchibe, E., & Doya, K. (in preparation). Imitation learning based on entropy-regularized forward and inverse reinforcement learning.

Sample efficiency w.r.t. the number of interactions

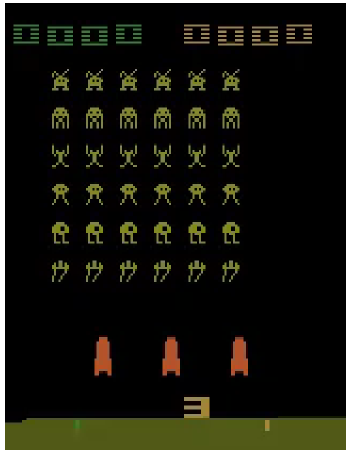


Uchibe, E., & Doya, K. (in preparation). Imitation learning based on entropy-regularized forward and inverse reinforcement learning.

Application of inverse RL to game-play

- Estimate the reward from play-data of three human players
  - Train optimal policies from the estimated rewards
- Evaluate the estimated reward by solving a forward RL from scratch
- ERIL is compared with Behavior Cloning (BC) and
  - LogReg-IRL: model-free version of OptV.
  - PI_IOC: Path-Integral based inverse RL

Space Invaders

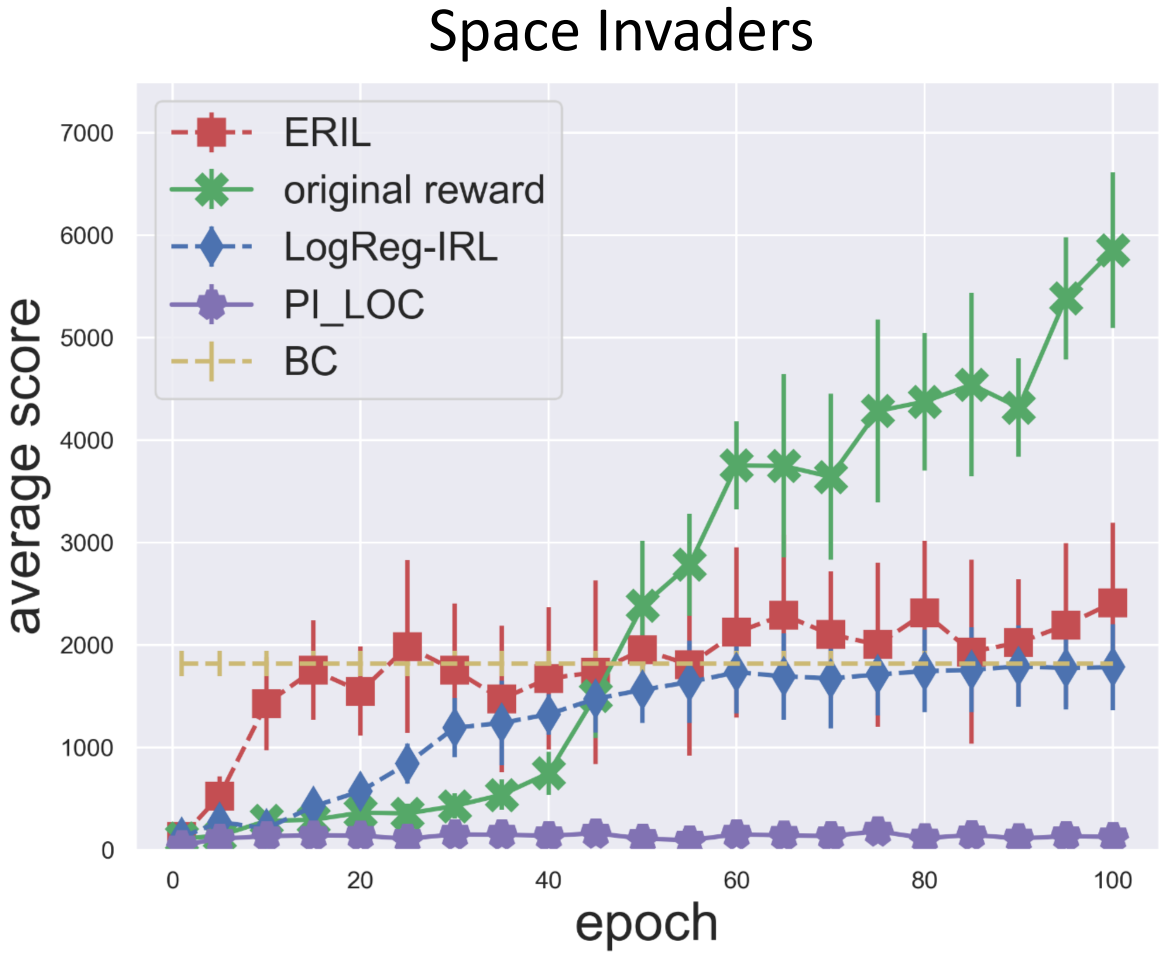


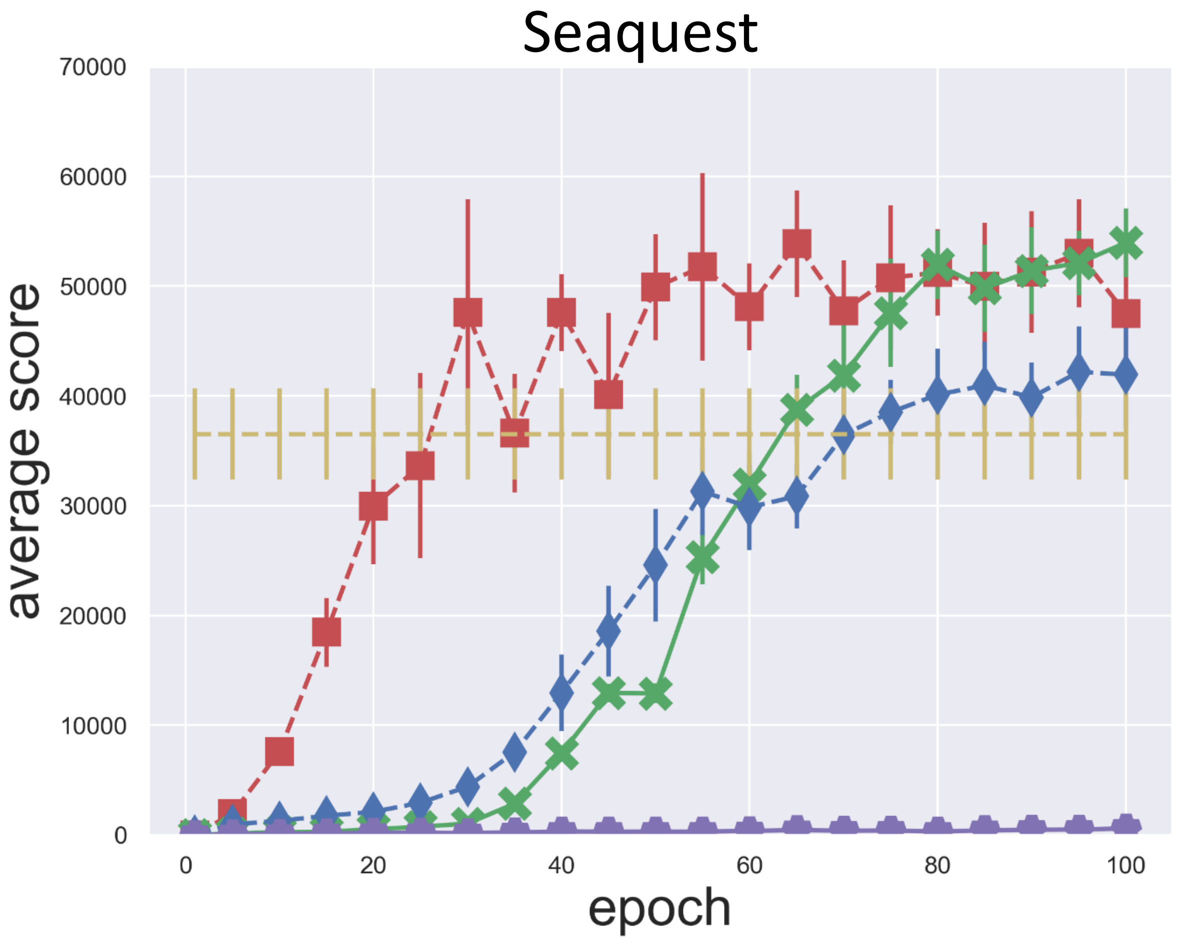
Seaquest

Uchibe, E., & Doya, K. (in preparation). Imitation learning based on entropy-regularized forward and inverse reinforcement learning.

Application of inverse RL to game-play

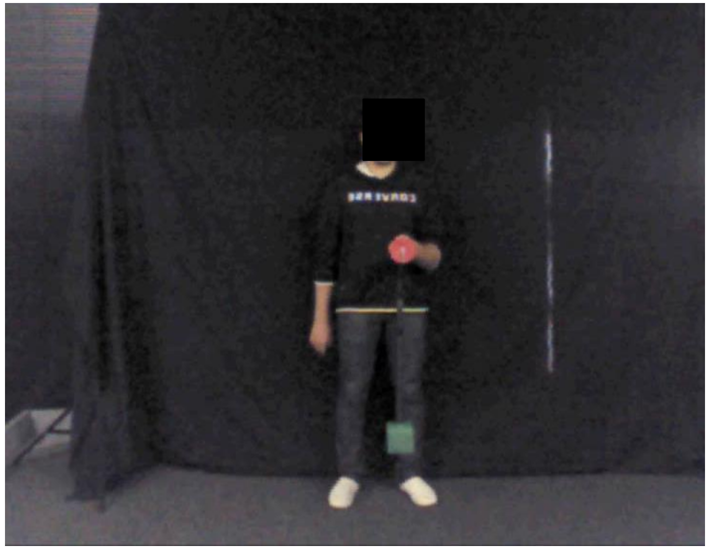
- The estimated reward improved the initial learning period
- Improving baseline was important



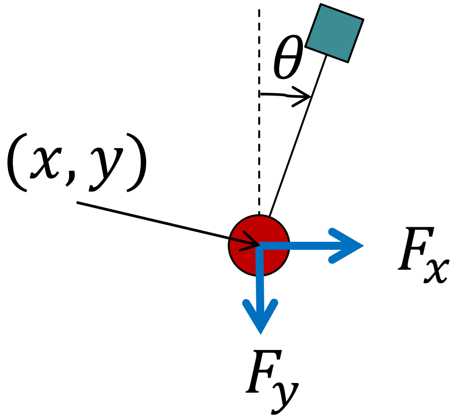


Analyzing inverted pendulum task

- The goal is to swing up and keep the pole upright for more than 3 [s]
- Task conditions:
  - length: long (73 cm), short (29 cm)
  - 15 trials for each pole
  - 40 [s] for each trial
  - 7 subjects (5: right-handed, 2: left-handed)
  - Action is not observed
- ERIL is compared with BC and
  - GAIfO: GAN-based imitation
  - OptV1: good baseline
  - OptV2: bad baseline



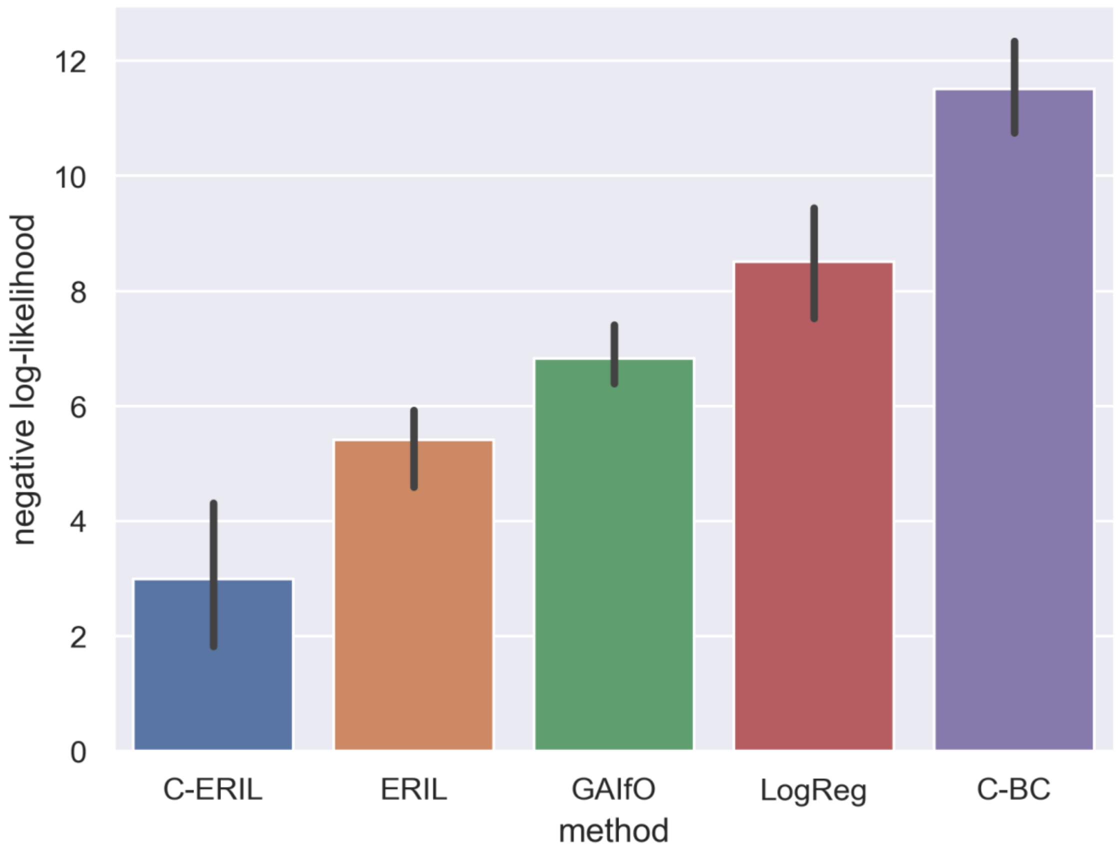
- State: $(x, \dot{x}, y, \dot{y}, \theta, \dot{\theta})$
- Action: (F_x, F_y)



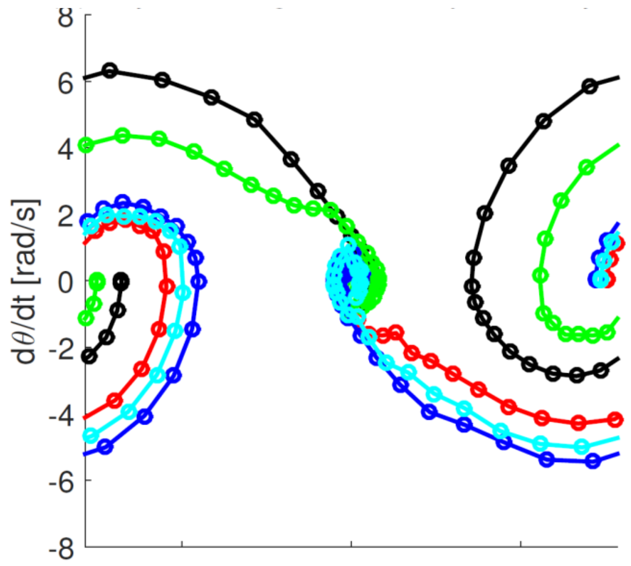
[Uchibe and Doya, in preparation]

Analyzing inverted pendulum task

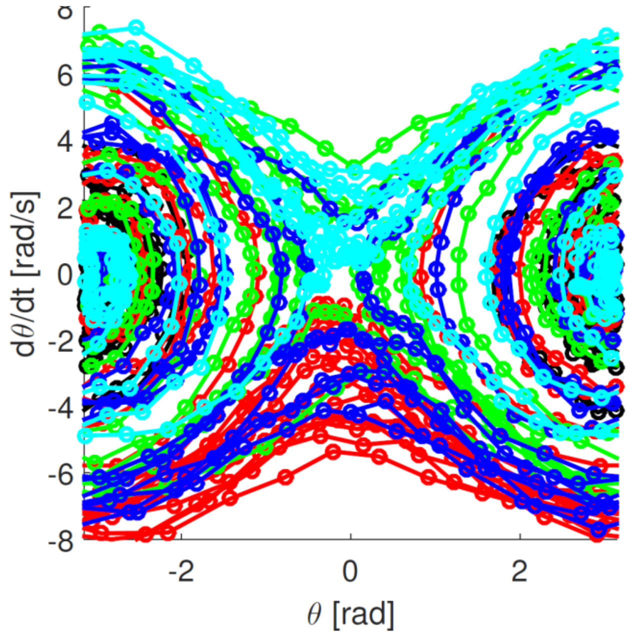
- Update the baseline by reinforcement learning with the estimated reward
  ➡ improve the performance



observed trajectories



generated trajectories



Report

- Please select one topic and write your report

1. Consider the application of reinforcement learning
2. Consider the application of inverse reinforcement learning

References

- Blondé, L., & Kalousis, A. (2019). Sample-Efficient Imitation Learning via Generative Adversarial Nets. *Proc. of the 22nd International Conference on Artificial Intelligence and Statistics*, 3138–48.
- Finn, C., Christiano, P., Abbeel, P., and Levine, S. (2016). A Connection Between Generative Adversarial Networks, Inverse Reinforcement Learning, and Energy-Based Models. NIPS 2016 Workshop on Adversarial Training.
- Fu, J., Luo, K., and Levine, S. (2018). Learning robust rewards with adversarial inverse reinforcement learning. In Proc. of ICLR.
- Fujimoto, S., van Hoof, H., & Meger, D. (2018). Addressing Function Approximation Error in Actor-Critic Methods. *Proc. of the 35th International Conference on Machine Learning.*
- Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. NIPS29.
- Kostrikov, I., Agrawal, K.K., Dwibedi, D., Levine, S., & Tompson, J. (2019). Discriminator-Actor-Critic: Addressing Sample Inefficiency and Reward Bias in Adversarial Imitation Learning. Proc. of the 7th ICLR
- Kozuno, T., Uchibe, E., and Doya, K. (2019). Theoretical analysis of efficiency and robustness of softmax and gap-increasing operators in reinforcement learning. In Proc. of AISTATS.

References

- Sasaki, F., Yohira, T., & Kawaguchi, A. (2019). Sample Efficient Imitation Learning for Continuous Control. *Proc. of the 7th International Conference on Learning Representations*.
- Uchibe, E. & Doya, K. (2014). Inverse reinforcement learning using dynamic policy programming. In Proc. of ICDL and Epirob.
- Uchibe, E. (2018). Model-Free Deep Inverse Reinforcement Learning by Logistic Regression. Neural Processing Letters, 47(3): 891-905.
- Uchibe, E. (2019). Imitation learning based on entropy-regularized forward and inverse reinforcement learning. Proc. of RLDM.
- Uchibe, E., & Doya, K. (in preparation). Imitation learning based on entropy-regularized forward and inverse reinforcement learning.
- Ziebart, B.D., Maas, A., Bagnell, J.A., & Dey, A. (2008). Maximum entropy inverse reinforcement learning. In Proc. of AAAI, 1433-1438.