Research Statement

Wonseok Jeon
✉ *jeonwons@mila.quebec*
https://wsjeon.github.io/
(Last Update: Nov 23, 2020)

My (co-)first-author publications [1, 2, 3, 4] since 2018 have been focused on the problems of reinforcement learning (RL), especially on *(1) Imitation Learning (IL)*—learning to mimic expert's behavior—*(2) Inverse Reinforcement Learning (IRL)*—acquiring a reward that rationalizes expert's behavior—and *(3) adversarial IL and IRL*—IL and IRL algorithms with adversarial training objectives [5] (See *Figure 1*).

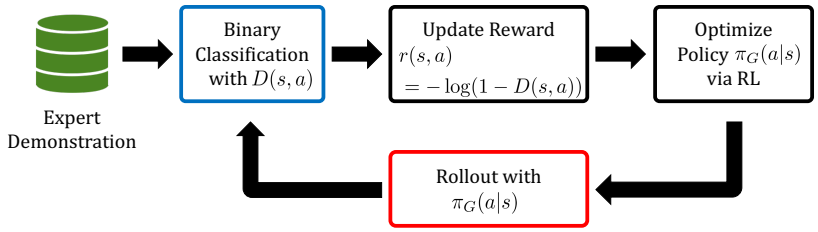

Figure 1. Iterative training of adversarial imitation learning

Practically, IL and IRL are crucial to learn from human demonstrations, enabling us to solve many practical sequential decision-making problems where well-designed rewards are not available, and thus, RL is not directly applicable. My publications have solved various issues of IL or IRL, and their contributions are respectively summarized below.

- *Bayesian perspective for adversarial IL [1, NeurIPS 2018 Spotlight]*
  We propose a probabilistic graphical model for adversarial IL and show that discriminator training of adversarial IL can be seen as finding out the point estimate for maximum likelihood. Motivated by this observation, we consider posterior distribution over discriminators and use posterior predictive reward for policy optimization. We empirically show this approach highly **improves the sample efficiency of adversarial IL**.
- *Simplified adversarial IL without reinforcement learning [2, NeurIPS 2020 Spotlight]*
  Existing adversarial ILs commonly require the internal policy optimization through reinforcement learning, which increases algorithmic complexity and may cause unstable training. In this work, a structured discriminator was proposed, where the discriminator involves policy networks as its internal component. Since learning with the proposed discriminator simultaneously optimizes policies, it is possible to **imitate expert's behavior only via supervised learning**.
- *Generalization of maximum-entropy IRL [3]*
  Regularized RL [6] generalizes entropy-regularized RL—that regularizes Shannon entropy of learner's policy—with a class of arbitrary convex policy regularizers. For the regularized IRL—a problem of seeking a reward such that an expert becomes optimal with regularized RL— we propose tractable solutions and a practical algorithm, which has not been done previously. This **generalizes maximum-entropy IRL** [7] which has motivated many modern IL and IRL algorithms.
- *Scalable and sample-efficient multi-agent adversarial IRL [4]*
  Existing adversarial IL and IRL for multi-agent problems have been validated for a small number of agents. We empirically combine multi-agent RL algorithms with various discriminator models and validate the multi-agent scalability of each combination. As a result, Multi-agent Actor-Attention-Critic (MAAC) [8] with decentralized discriminator for each agent highly outperforms other combinations as well as baseline algorithms, shown to be **scaled up to tens of agents**.

My plans for future works include *(1) seeking policy regularizers for robust IL*, *(2) exploiting the reward optimality condition for RL*, and *(3) IL with probabilistic planning*. More details on my publications and plans are explained in the following sections.

Preliminaries on Imitation Learning

Reinforcement learning (RL) has been successfully applied to many challenging tasks including robotics and games, and sometimes it has significantly outperformed human strategies. One crucial assumption of RL is a well-defined reward describing desirable behaviors in each task, which comes from the mathematical framework of Markov decision processes [9]. However, unlike video games where one can easily define the reward from scores, designing appropriate rewards is often cumbersome in many real-world applications.

Alternative to RL from scratch, we assume there is an expert, instead of the reward in RL. The learner's objective is to mimic the expert's behavior, mostly with the expert demonstration, e.g., human movements and doctor's prescription. Traditionally, behavioral cloning via supervised learning has been regarded as a simple way of IL. However, behavioral cloning was shown to suffer from the covariate shift issue [10], and an IL via Inverse Reinforcement Learning (IRL) [11]—learning the expert's reward from the demonstration followed by applying RL on that reward—has been considered. Particularly, Maximum-Entropy IRL (MaxEntIRL) [7]—an IRL method that penalizes Shannon entropy of the learner's policy during training—has been widely used [12].

More recently, Generative Adversarial Imitation Learning (GAIL) [13] and its variant for IRL, Adversarial Inverse Reinforcement Learning (AIRL) [14], have been proposed. Based on MaxEntIRL, GAIL and AIRL have applied generative adversarial networks [5] for IL and IRL, respectively, and have shown remarkable performances on high-dimensional tasks. Motivated by generative adversarial networks, both GAIL and AIRL iteratively optimize (1) a discriminator—a binary classifier to differentiate whether behavioral samples are from the expert demonstration or not—and (2) the learner's policy by using deep RL and the reward defined by the discriminator (See *Figure 1*). Later, GAIL and AIRL are extended to be applicable for multi-agent IL and IRL problems [15, 16].

A Bayesian Approach to GAIL [1]

Due to GAIL's internal RL procedure—which requires the samples from environment interactions—GAIL suffers from the problem of sample efficiency as general RL algorithms do. To enhance the sample efficiency of GAIL, we consider the probabilistic graphical model of GAIL and its relevant Bayesian perspective.

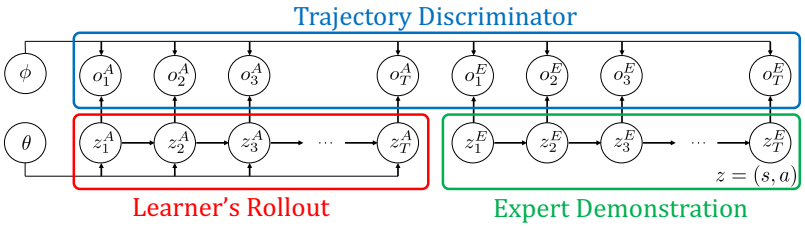

Figure 2. A probabilistic graphical model for GAIL

As shown in *Figure 2*, the graphical model contains random variables such as state-action pairs $z_t = (s_t, a_t)$ and labels o_t which become 1, if the state-action pair comes from the expert, and 0, otherwise, for $t = 1, ..., T$. Additionally, each state-action pair is assumed to be sampled from either the learner (superscript A) or the expert (superscript E). For such a graphical model, we define the *discrimination optimality event* $\mathcal{E}_D = \{o_1^A = \cdots = o_T^A = 0, o_1^E = \cdots = o_T^E = 1\}$—a probabilistic event that indicates the perfect classification—and the *imitation optimality event* $\mathcal{E}_I = \{o_1^A \neq 0 \text{ or } \cdots \text{ or } o_T^A \neq 0\}$—a probabilistic event that says the discriminator is fooled.

With such a graphical model, GAIL's discriminator and policy updates are shown to be the *maximum likelihood estimations* with events $\mathcal{E}_D$ and $\mathcal{E}_I$, respectively. Instead of the point estimate from maximum likelihood, we consider the *posterior distribution* for discriminators and use a posterior

predictive reward for the RL procedure. It can be empirically shown that using the posterior predictive reward is shown to highly improve the sample efficiency.

Adversarial IL without RL [2]

Both GAIL and AIRL iteratively optimize a discriminator and the learner's policy via RL. However, such an alternated optimization is known to be delicate in practice since it compounds unstable adversarial training and uses deep RL that requires considerable resources for engineering.

We remove the burden of the RL part by leveraging a novel discriminator formulation. The contribution of our work comes from using *the discriminator* $D(s, a; \pi, \pi_G)$ *explicitly conditioned on two policies*: $\pi_G(a|s)$—a policy from the previous iteration and fixed during discriminator training—and $\pi(a|s)$—a policy which can be updated during discriminator training. For example, one of our methods uses the discriminator $D(s, a; \pi, \pi_G) = \frac{\pi(a|s)}{\pi(a|s)+\pi_G(a|s)}$ and is optimized by updating the policy π of D *when binary classification loss is minimized for training* D. After D is optimized, its inherent policy π is shown to recover the expert's policy.

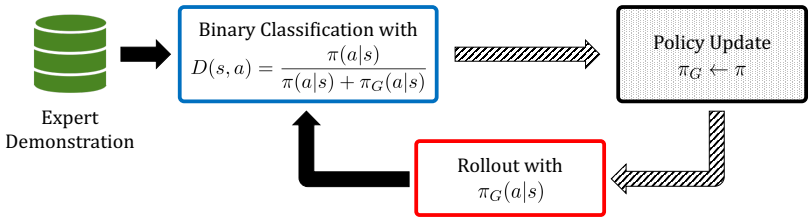

Figure 3. Our method without RL

This approach only requires the binary classification during training, enabling us not to use the policy optimization through deep RL algorithms (See *Figure 3*). That is, our formulation effectively cuts by half the implementation and computational burden of GAIL and AIRL by removing RL steps. We show on a variety of tasks that our simpler approach is competitive to existing adversarial IL methods.

Regularized Inverse Reinforcement Learning [3]

While Shannon entropy is often used as a policy regularizer [7], Geist et al. [6] recently proposed a theoretical foundation of *regularized Markov decision processes* (MDPs)—a framework that uses arbitrary strongly convex functions as policy regularizers. Here, one crucial advantage is that an optimal policy *uniquely* exists, whereas multiple optimal policies may exist in the absence of policy regularization.

Thanks to this advantage, [6] showed that IRL in regularized MDPs—*regularized IRL*—does not contain degenerate solutions—any constant can be a solution for IRL in unregularized MDPs. Nonetheless, analytical solutions and practical algorithms for regularized IRL—other than maximum-(Shannon-)entropy IRL (MaxEntIRL) [7]—have not yet been proposed.

We propose tractable solutions for regularized IRL problems. We show that RL with our IRL solutions (rewards) is equivalent to *minimizing average Bregman divergence* [17] between the learner and the expert policies. Our approach covers MaxEntIRL when Shannon entropy is considered to be a policy regularizer.

We devise Regularized Adversarial Inverse Reinforcement Learning (RAIRL), a practical sample-based method for policy imitation and reward learning in regularized MDPs, which generalizes AIRL, and empircially validate RAIRL on both discrete and continuous control tasks.

Scalable Multi-Agent Inverse Reinforcement Learning [4]

Multi-Agent AIRL (MA-AIRL) [15, 16] is a recent approach that applies single-agent AIRL to multi-agent problems where we seek to recover both policies for multiple learners and reward functions that promote

expert-like behavior. However, MA-AIRL has only been validated empirically for small numbers of agents—its applicability to many agents remains an open question.

We seek a multi-agent IRL method that is more sample-efficient and is applicable to larger numbers of agents than previous works. Specifically, we employ Multi-agent Actor Attention Critic (MAAC) [8]—a scalable off-policy multi-agent RL method via attention mechanism [18]—for the RL inner loop of the IRL procedure. Then, we empirically tested our method with various discriminator models—centralized/decentralized/attention-based discriminators—which are used for the reward learning of the IRL procedure.

In doing so, we find out the MA-AIRL method that is highly sample-efficient compared to state-of-the-art baselines for the environments up to *tens of agents* where the baselines struggle to learn. Moreover, the RL agents trained with the rewards acquired by our method better match the experts than those trained on the rewards derived from the baselines. Finally, our method requires far fewer environment interactions, particularly as the number of agents increases.

Conclusion and Future Works

There has been a huge advance in modern IL and IRL along with a success with deep RL algorithms, and I believe IL and IRL will become more crucial to utilize human knowledge on practical domains. This motivates me to think of tentative research directions which are deeply related to my publications:

- *Robust imitation learning with novel policy regularizer and imitation loss*
  In my work on regularized IRL [3], the key ingredient of generalization is the Bregman divergence [17] which covers a wide range of probabilistic divergences and has often appeared machine learning literature other than IRL. Recently in [19], a novel loss function motivated by the Bregman divergence was proposed for supervised learning, shown to effectively improve the robustness of classification w.r.t. outliers. Since human demonstrations are often noisy in practice—mostly coming from multiple human experts—it would be interesting to consider which policy regularizers and loss functions can make **IL robust for the imperfect demonstrations**.
- *Exploiting the reward optimality condition to improve RL*
  From the mathematical derivation of my work [3], it can be shown that there exists a condition that the reward, value and policy should satisfy at the optimality—where maximum expected return is achieved in regularized RL—but existing works have not utilized this condition to the best of my knowledge. Motivated by this observation, we may use structured networks, e.g., value network that inherits policy network as was done in [20]. I believe this approach may highly **simplify neural network models and improve the performance of RL**.
- *IL with probabilistic planning*
  Recent advances in simulators and Sim2Real transfer make learning from simulators portable to real-world applications, and I expect the role of planning becomes more crucial in practice. However, only a few works on IL with planning (e.g. [21]) have been proposed. One of my plans is to bridge IL with control-as-inference frameworks [22] in a rigorous way and apply probabilistic planning [23] to planning-based IL with simulators so that we can exploit existing sequential Bayesian inference methods for IL. I believe this approach will **outperform model-free IL baselines and be more computationally-efficient than planning-based IL.**

References

[1] Wonseok Jeon, Seokin Seo, and Kee-Eung Kim. A Bayesian approach to generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. paper link.

[2] Paul Barde*, Julien Roy*, Wonseok Jeon*, Joelle Pineau, Christopher Pal, and Derek Nowrouzezahrai. Adversarial soft advantage fitting: Imitation learning without policy optimization.

In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. *Equal Contribution, paper link.

[3] Wonseok Jeon, Chen-Yang Su, Paul Barde, Thang Doan, Derek Nowrouzezahrai, and Joelle Pineau. Regularized inverse reinforcement learning. In *NeurIPS Deep Reinforcement Learning Workshop (DRLW)*, 2020. paper link.

[4] Wonseok Jeon, Paul Barde, Derek Nowrouzezahrai, and Joelle Pineau. Scalable multi-agent inverse reinforcement learning via actor-attention-critic. In *AAAI Workshop on Reinforcement Learning in Games (AAAI-RLG)*, 2020. paper link.

[5] Ian Goodfellow, Jean Pouget Abadie, Mehdi Mirza, Bing Xu, David Warde Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

[6] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized Markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2160–2169, 2019.

[7] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, 2008.

[8] Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 2961–2970, 2019.

[9] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[10] Stéphane Ross and Drew Bagnell. Efficient reductions for imitation learning. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[11] Andrew Y Ng and Stuart J Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, 2000.

[12] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pages 49–58, 2016.

[13] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[14] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 2018.

[15] Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. Multi-agent generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7461–7472, 2018.

[16] Lantao Yu, Jiaming Song, and Stefano Ermon. Multi-agent adversarial inverse reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019.

[17] Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

[19] Ehsan Amid, Manfred KK Warmuth, Rohan Anil, and Tomer Koren. Robust bi-tempered logistic loss based on bregman divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 15013–15022, 2019.

[20] Brendan O'Donoghue, Remi Munos, Koray Kavukcuoglu, and Volodymyr Mnih. Combining policy gradient and q-learning. *arXiv preprint arXiv:1611.01626*, 2016.

[21] Nicholas Rhinehart, Rowan McAllister, and Sergey Levine. Deep imitative models for flexible inference, planning, and control. *arXiv preprint arXiv:1810.06544*, 2018.

[22] Sergey Levine. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

[23] Alexandre Piché, Valentin Thomas, Cyril Ibrahim, Yoshua Bengio, and Chris Pal. Probabilistic planning with sequential Monte Carlo methods. In *International Conference on Learning Representations*, 2018.