Non-Adversarial Inverse Reinforcement Learning by Distribution Matching

I. Introduction and Preliminaries

Reinforcement Learning is a popular technique that enables machines to autonomously learn how to solve challenging tasks such as playing video games [1], [2] or robotic grasping [3]. For learning effective policies reinforcement learning relies on reward functions that assess the quality of the current behavior. However, designing reward functions that induce the correct behavior is difficult even for expert. In contrast, inverse reinforcement learning (IRL [4]) enables non-expert users to program machines (such as robots) by demonstrating the desired behavior. IRL aims to infer a reward function that explains the demonstrations and that can subsequently be optimized by the machine to solve the task. In this paper we focus on three major challenges of current IRL methods that are related to efficiency, generalizability and applicability.

Many early methods for IRL [5], [6], [7] required to iteratively solve a reinforcement learning problem and were therefore only applicable to low-dimensional and discrete Markov Decision Processes (MDPs). Instead, modern methods [8], [9], [10], which relate to *Generative Adversarial Networks* (GANs [11]) interleave inverse reinforcement learning and reinforcement learning. These methods can solve the inverse reinforcement learning problem with similar computational costs compared to reinforcement learning and are applicable to high-dimensional problems and complex function approximators–such as neural networks–for policy and reward function. However, these methods have only been applied in combination with on-policy reinforcement learning methods such as TRPO [12] or PPO [13] that require a large amount of system interactions and are, thus, not applicable to robotic applications. An imitation learning method that is similar to the aforementioned methods, *GAIL* [14], was recently successfully applied to off-policy methods [15]. However, based on our preliminary experiments combining existing IRL methods such as AIRL [10], with off-policy reinforcement learning is not straightforward. In this work, we propose an IRL method that interleaves IRL with the off-policy reinforcement learning algorithm *SAC* [16] thereby achieving higher sample efficiency than existing IRL methods.

An additional challenge for modern IRL methods is to learn a reward function that correctly encodes the goal of the task in order to generalize to changes in the environment. Indeed, a reward function may not contain any more information than a policy. For example, in a maximum entropy RL [7], [16] setting a reward function defined by $r(\mathbf{s}, \mathbf{a}) = \log \pi(\mathbf{a}|\mathbf{s})$ will induce the policy $\pi(\mathbf{a}|\mathbf{s})$. Such reward function rewards the agent for closely following the reference policy $\pi(\mathbf{a}|\mathbf{s})$ which may fail to solve the task in case of changes in the environment. AIRL addresses this problem by enforcing a state-only reward function such that

$$r(\mathbf{s}) - V(\mathbf{s}) + \gamma V(\mathbf{s}') = \log \pi(\mathbf{a}|\mathbf{s}), \qquad (1)$$

where $\mathbf{s}'$ denotes the state that was reached after applying action $\mathbf{a}$ in state $\mathbf{s}$, $V(\mathbf{s})$ denotes the value in state $\mathbf{s}$ and γ corresponds to the discount factor. Although Equation 1 is only sensible for deterministic MDPs, AIRL showed that it may recover meaningful reward functions also for stochastic MDPs. However, AIRL depends on a specific form of the discriminator which is only applicable if the demonstrations include direct observations of states and actions. Our method also uses Equation 1 to learn a robust reward function, but it does not pose any constraints on the discriminator making it applicable to arbitrary observations.

II. IRL by Information Projection

Similar to MaxCausalEnt-IRL [7] our inverse reinforcement method is derived from an imitation learning problem, i.e., we aim to learn a policy that behaves similar to the demonstrations. However, in contrast to MaxCausalEnt-IRL which can be shown to minimize the forward Kullback-Leibler divergence, $\mathrm{KL}(p_{\text{expert}}(\mathbf{o})||p_\pi(\mathbf{o}))$, between the observation-distribution of the expert, $p_{\text{expert}}(\mathbf{o})$, and the distribution induced by the agent's policy $p_\pi(\mathbf{o})$, we aim to minimize the reverse KL, $\mathrm{KL}(p_\pi(\mathbf{o})||p_{\text{expert}}(\mathbf{o}))$, that is,

$$\arg\max_{\pi(\mathbf{a}|\mathbf{s})} \int_{\mathbf{o}} p_\pi(\mathbf{o}) \log \frac{p_{\text{expert}}(\mathbf{o})}{p_\pi(\mathbf{o})} d\mathbf{o} + \alpha H(\pi(\mathbf{a}|\mathbf{s})), \qquad (2)$$

where $\alpha \geq 0$ can be used for regularization by increasing the policy's entropy $H(\pi(\mathbf{a}|\mathbf{s}))$. Compared to minimizing the forward KL, minimizing the reverse KL can be especially beneficial when multi-modal demonstrations are to be matched by a uni-modal (e.g. Gaussian) policy [17]. A similar, yet constraint-based formulation was used by Arenz et al. [18], where they showed that their Lagrangian multiplier can be regarded as reward function with its optimum at $r_{\text{opt}}(\mathbf{s}, \mathbf{a}) \propto \log p_{\text{expert}}(\mathbf{o}(\mathbf{s}, \mathbf{a})) - \log p_{\pi_{\text{opt}}}(\mathbf{o}(\mathbf{s}, \mathbf{a}))$, resulting in a mutual dependency between the optimal reward function and the policy that optimizes their objective function. Our method (as well as AIRL) use a similar reward function. However, whereas the work by Arenz et al. [18] involved iteratively solving the reinforcement learning problem, our work directly optimizes Equation 2 based on a lower bound decomposition making it applicable to more complex MDPs and function approximators. More specifically, we iteratively increase a lower bound of Equation 2,

$$\arg\max_{\pi(\mathbf{a}|\mathbf{s})} \int_{\mathbf{o}} p_\pi(\mathbf{s}, \mathbf{a}) \log \frac{p_{\text{expert}}(\mathbf{o})}{p_{\pi_{\text{ref}}}(\mathbf{o})} d\mathbf{o} \\ + E_\pi\left[\log \pi_{\text{ref}}(\mathbf{a}|\mathbf{s})\right] + (1+\alpha) H(\pi(\mathbf{a}|\mathbf{s})),$$

and tighten it by setting $\pi_{\text{ref}} = \pi$. For increasing the lower bound we can use any reinforcement learning algorithm that can use entropy regularization, e.g. SAC or TRPO, and perform few policy updates optimizing a reward function of the form

$$r(\mathbf{s}, \mathbf{a}) = \sigma(\mathbf{o}(\mathbf{s}, \mathbf{a})) + r_{\pi_{\text{ref}}}(\mathbf{s}) - V_{\pi_{\text{ref}}}(\mathbf{s}) + \gamma V_{\pi_{\text{ref}}}(\mathbf{s}'). \qquad (3)$$

Here $r_{\pi_{\text{ref}}}$ and $V_{\pi_{\text{ref}}}$ are trained via supervised learning to approximate $\log \pi_{\text{ref}}$ by means of Equation 1; σ corresponds to the logit produced by a discriminator that is trained via binary cross-entropy loss to classify between samples that were collected from the expert and samples that were collected using the policy π_{ref}. As shown by Sugiyama et al. [19] these logits approximate the log density-ratio, $\sigma(\mathbf{o}) \approx \log(p_{\text{expert}}(\mathbf{o})/p_{\pi_{\text{ref}}}(\mathbf{o}))$. By using a discriminator, our method closely connects to AIRL. However, our method is non-adversarial and we do not need to solve a min-max game. Instead, optimizing the function approximators in Equation 3 corresponds to tightening a lower-bound objective which does not depend on the agent's policy π. For a policy that matches the expert demonstrations, the log density-ratio should become zero and–as shown by Ng et al. [20]–$r_{\pi_{\text{ref}}}(\mathbf{s})$ would induce the same optimal behavior as $r(\mathbf{s}, \mathbf{a})$ on deterministic systems. In practice, the discriminator might still produce large values especially on areas with few samples and the system dynamics are often stochastic. However, preliminary experiments on the *Maze* [10] experiment show that $r_{\pi_{\text{ref}}}(\mathbf{s})$ still induces the desired behavior even after changes in the environment.

REFERENCES

[1] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[2] Oriol Vinyals, Igor Babuschkin, Junyoung Chung, Michael Mathieu, Max Jaderberg, Wojtek Czarnecki, Andrew Dudzik, Aja Huang, Petko Georgiev, Richard Powell, Timo Ewalds, Dan Horgan, Manuel Kroiss, Ivo Danihelka, John Agapiou, Junhyuk Oh, Valentin Dalibard, David Choi, Laurent Sifre, Yury Sulsky, Sasha Vezhnevets, James Molloy, Trevor Cai, David Budden, Tom Paine, Caglar Gulcehre, Ziyu Wang, Tobias Pfaff, Toby Pohlen, Dani Yogatama, Julia Cohen, Katrina McKinney, Oliver Smith, Tom Schaul, Timothy Lillicrap, Chris Apps, Koray Kavukcuoglu, Demis Hassabis, and David Silver. AlphaStar: Mastering the Real-Time Strategy Game StarCraft II. https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/, 2019.

[3] Stephen James, Paul Wohlhart, Mrinal Kalakrishnan, Dmitry Kalashnikov, Alex Irpan, Julian Ibarz, Sergey Levine, Raia Hadsell, and Konstantinos Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12627–12637, 2019.

[4] A. Ng and S. Russell. Algorithms for Inverse Reinforcement Learning. In *in Proceeedings of the 17th International Conference on Machine Learning (ICML)*, 2000.

[5] P. Abbeel and A. Ng. Apprenticeship learning via Inverse Reinforcement Learning. In *Proceedings of the 21st International Conference on Machine Learning (ICML)*, 2004.

[6] N. Ratliff, A. Bagnell, and M. Zinkevich. Maximum Margin Planning. In *In Proceedings of the 23rd International Conference on Machine Learning (ICML)*, 2006.

[7] B. Ziebart, A. Bagnell, and A. Dey. Modeling Interaction via the Principle of Maximum Causal Entropy. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

[8] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 49–58, 2016.

[9] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *CoRR*, abs/1611.03852, 2016.

[10] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[12] John Schulman, Jonathan Ho, Cameron Lee, and Pieter Abbeel. Learning from demonstrations through the use of non-rigid registration. In *Robotics Research*, pages 339–354. Springer, 2016.

[13] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[14] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4565–4573, 2016.

[15] Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. 2019.

[16] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[17] Liyiming Ke, Matt Barnes, Wen Sun, Gilwoo Lee, Sanjiban Choudhury, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. *arXiv preprint arXiv:1905.12888*, 2019.

[18] O. Arenz, H. Abdulsamad, and G. Neumann. Optimal control and inverse optimal control by distribution matching. In *Proceedings of the International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[19] Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density ratio estimation in machine learning*. Cambridge University Press, 2012.

[20] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.