Policy Optimization with Demonstration

Chia-Hung, Liao
Department of Computer Science
National Chiao Tung University
aiallen.cs07g@nctu.edu.tw

1 Introduction

- The main research challenges tackled by the paper:
  Exploration still remains a significant challenge to reinforcement learning algorithms, especially the reward signals from the environment are sparse. Learning from demonstrations (LfD) (Hester et al. (2018); Vecerik et al. (2017)) is a common approach for addressing exploration difficulties in sparse reward tasks. However, existing LfD methods is limited by only treating the demonstrations in the same way as self-generated data. Besides, the traditional LfD usually require a large number of high-quality demonstrations which are difficult to collect at scale.
- The high-level technical insights into the problem of interest
  The intuition of LfD is that the agent could imitate the expert demonstrations when the reward signal sparse in early learning stages instead of random exploration. After acquiring enough assistances, the agent can explore for even better policy on its own. In other words, LfD can be viewed as a dynamic intrinsic reward mechanism. That is, one can introduce demonstrations for reshaping native rewards in RL. Therefore, this research proposes a novel Policy Optimization from Demonstration (POfD) method (Kang et al. (2018)), which can acquire knowledge from demonstration data to boost exploration, even though the data are scarce and imperfect.
- The main contributions of the paper (compared to the prior works)
  1. The research successfully proves that POfD induces implicit dynamic reward shaping and brings significant benefits for policy improvement.
  2. POfD is generally compatible with most policy gradient algorithms.
  3. It also shows that existing LfD methods Vecerik et al. (2017) can be interpreted as degenerated cases of POfD in terms of how to leverage the demonstration data.
- Your personal perspective on the proposed method
  In my opinion, this POfD method looks similar to the inverse reinforcement learning method (IRL) (Fu et al. (2017)). But the POfD will be better than the typical generative IRL in some sense due to its mechanism. When the environment feedback reward is sparse, we usually let the agent learn from the expert's demonstration. In addition, we don't know the reward function for the expert policy. Instead, IRL can adopt generative methods such as generative adversarial networks (GANs) (Goodfellow et al. (2014)) to learn the reward function for expert policy by using the expert trajectories even though the expert demonstration is few. However, the typical IRL regards expert demonstration as the best policy. When it comes to POfD, it reshapes the reward function during the policy optimization process. Furthermore, POfD does not regard expert demonstration as the best policy. Instead, this new reshaped reward function can guide the agent to imitate expert behavior when the environment reward is sparse, and explore independently when it can get the environment reward value. In this way, the expert demonstration can be more fully utilized, and there is no need to ensure that the expert policy is the optimal one, which is a great improvement compared to the previous methods.

Preprint. Under review.

2 Problem Formulation

Please present the formulation in this section. You may want to cover the following aspects:

- Your notations (e.g. MDPs, value functions, function approximators,...etc)
- The optimization problem of interest
- The technical assumptions

2.1 Preliminaries and Notations

Definition 1. *(Occupancy measure) Let $\rho_\pi(s) : \mathcal{S} \to \mathbb{R}$ denote the unnormalized distribution of state visitation by following policy π in the environment:*

$$\rho_\pi(s) = \sum_{t=0}^{\infty} \gamma^t P(s_t = s|\pi).$$

From the Def. 1, for better exploiting demonstrations, this research first converts the expected discounted reward from Eq. (1) to a Eq. (2) defined on occupancy measure (Ho Ermon, 2016).

$$\eta(\pi) = \mathbb{E}_\pi \left[r(s,a)\right] = \mathbb{E}_{(s_0,a_0,s_1,\dots)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)\right], \tag{1}$$

$$\begin{aligned} \mathbb{E}_\pi \left[r(s,a)\right] &= \sum_{t=0}^{\infty} \sum_s P(s_t = s|\pi) \sum_a \pi(a|s)\gamma^t r(s,a) \\ &= \sum_s \rho_\pi(s) \sum_a \pi(a|s) r(s,a) \\ &= \sum_{s,a} \rho_\pi(s,a) r(s,a), \end{aligned} \tag{2}$$

In the following lemma (Syed and Schapire (2008)), the occupancy measure has an important property that it uniquely specifies a policy.

Lemma 1. *(Theorem 2 of (Syed et al., 2008)) Suppose ρ is the occupancy measure for $\pi_\rho(a|s) \triangleq \frac{\rho(s,a)}{\sum_{a'} \rho(s,a')}$. Then π_ρ is the only policy whose occupancy measure is ρ.*

Besides, in addition to sparse rewards from environments, the agent is also provided with a few demonstrations $D^E = \{\tau_1, \tau_2, ..., \tau_N\}$, where the i-th trajectory $\tau_i = \left\{(s_0^i, a_0^i), (s_1^i, a_1^i), ..., (s_T^i, a_T^i)\right\}$ is generated from executing an unknown expert policy π_E.

2.2 The optimization problem of interest

Besides maximizing the expected return $\eta(\pi_\theta)$ through learning from sparse feedback, this research also encourages the agent to explore by "following" the demonstrations DE. Then an introduce demonstration-guided exploration term $L_M(\pi_\theta, \pi_E) = D_{JS}(\pi_\theta, \pi_E)$ is introduced to the vanilla objective $\eta(\pi_\theta)$. This gives a new learning objective:

$$\mathcal{L}(\pi_\theta) = -\eta(\pi_\theta) + \lambda_1 D_{JS}(\pi_\theta, \pi_E),$$

Where λ_1 is a trading-off parameter, and D_{JS} is the Jensen-Shannon divergence between current policy π_{theta} and the expert one π_E. However, this divergence measure is infeasible since π_E is unknown. Instead, we redefine the divergence over the occupancy measures $\rho_\pi(s,a)$ and $\rho_{\pi_E}(s,a)$ by Lemma 1. And finally the proposed demonstration guided learning objective is changed to the Eq. (3)

$$\mathcal{L}(\pi_\theta) = -\eta(\pi_\theta) + \lambda_1 D_{JS}(\rho_\theta, \rho_E). \tag{3}$$

2.3 The technical assumptions

Although the expert policy may not best one, but its quality is usually better than the agent policy in early training stage. Therefore, this paper presents a reasonable and necessary assumption, and it will be used to prove the benefits of exploration with demonstrations.

Assumption 1. *In early learning stages, we assume acting according to expert policy π_E will provide higher advantage value with a margin as least δ over current policy π,* i.e.,

$$\mathbb{E}_{a_E \sim \pi_E, a \sim \pi}[A_\pi(s, a_E) - A_\pi(s, a)] \geq \delta.$$

And the second assumption is the same when using the form of surrogate objective when proving the TRPO algorithm (Schulman et al. (2015)).

$$J_{\pi_{\text{old}}}(\pi) = \eta(\pi_{\text{old}}) + \sum_s \rho_{\pi_{\text{old}}}(s) \sum_a \pi(a|s) A_{\pi_{\text{old}}}(s,a)$$

The assumption is, the reason why ρ_π is replaced by $\rho_{\pi_{old}}$ is that the change in state distribution can be ignored due to policy update.

3 Theoretical Analysis

Please present the theoretical analysis in this section. Moreover, please formally state the major theoretical results using theorem/proposition/corollary/lemma environments. Also, please clearly highlight your new proofs or extensions (if any).

3.1 Proof for Benefits of Exploration with Demonstrations

First, it is quite important to know whether adding the JS-divergence term as learning regularization could improve the agent policy eventually. So, starting from the surrogate function and the learning objective Eq. (3), one can prove the following Theorem:

Theorem 1. *Let* $\alpha = D_{KL}^{\max}(\pi_{old}, \pi) = \max_s D_{KL}(\pi(\cdot|s), \pi_{old}(\cdot|s))$, $\beta = D_{JS}^{\max}(\pi_E, \pi) = \max_s D_{JS}(\pi(\cdot|s), \pi_E(\cdot|s))$, *and* π_E *is an expert policy satisfying Assumption 1. Then we have*

$$\eta(\pi) \geq J_{\pi_{old}}(\pi) - \frac{2\gamma(4\beta\epsilon_E + \alpha\epsilon_\pi)}{(1-\gamma)^2} + \frac{\delta}{1-\gamma},$$

where $\epsilon_E = \max_{s,a} |A_{\pi_E}(s,a)|, \epsilon_\pi = \max_{s,a} |A_\pi(s,a)|$.

Since it needs three pages to prove this inequality function of Theorem 1, and it has been provided by the paper's supplement, so I decide to skip the details in this report. Most importantly, one can use Theorem 1 to further prove the benefits of adding the demonstration guided regularization term.
let $M_i(\pi) = J_{\pi_i}(\pi) - C_{\pi_E} D_{JS}^{max}(\pi, \pi_E) - C_\pi D_{KL}^{max}(\pi, \pi_i) + \hat{\delta}$
where $C_{\pi_E} = \frac{8\gamma\epsilon_E}{(1-\gamma)^2}$, $C_\pi = \frac{2\gamma\epsilon_\pi}{(1-\gamma)^2}$, $\hat{\delta} = \frac{\delta}{1-\gamma}$.
And substitue the above $M_i(\pi)$ into the function of Theorem 1, we can derive:
$\eta(\pi_{i+1}) \geq M_i(\pi_{i+1})$.
Then, since the KL divergence between the same policies will be zero, so
$\eta(\pi_i) = M_i(\pi_i) + C_{\pi_E} D_{JS}^{max}(\pi_i, \pi_E) - \hat{\delta}$.
Therefore, from these above two equations, one can derive:
$\eta(\pi_{i+1}) - \eta(\pi_i) \geq M_i(\pi_{i+1}) - M_i(\pi_i) - C_{\pi_E} D_{JS}^{max}(\pi_i, \pi_E) + \hat{\delta}$
This result recalls the classic monotonic improvement guarantees for the policy gradient algorithm. But POfD brings another two terms $C_{\pi_E} D_{JS}^{max}(\pi_i, \pi_E)$ and $\hat{\delta}$. This implies when the agent follows the demonstrations, the JS divergence is close to zero, and therefore the advantage term $\hat{\delta}$ from the reasonable assumption 1 can guarantee the monotonic policy improvement.

3.2 Main Optimization Objective

Since the JS divergence is quite hard to optimize, so this paper changes the way to optimize its lower bound which is given as Theorem 2:

Theorem 2. *Let* $h(u) = \log(\frac{1}{1+e^{-u}})$, $\bar{h}(u) = \log(\frac{e^{-u}}{1+e^{-u}})$ *and* $U(s,a) : \mathcal{S} \times A \to \mathbb{R}$ *be an arbitrary function. Then we have*

$$D_{JS}(\rho_\pi, \rho_E) \geq \sup_U \left(\mathbb{E}_{\rho_\pi}[h(U(s,a))] + \mathbb{E}_{\rho_E}[\bar{h}(U(s,a))]\right) + \log 4.$$

Also, this theorem has been proved in the supplement provided by the paper author. Therefore, the occupancy measure matching objective can be written as

$$\mathcal{L}_M \triangleq \sup_{D \in (0,1)^{\mathcal{S}\times\mathcal{A}}} \mathbb{E}_{\pi_\theta}[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))],$$

That is, the supremum will range over $D(s,a) = \frac{1}{1+e^{-U(s,a)}}$ which is an arbitrary mapping function followed by a sigmoid activation funciton. And this objective can be regarded as the binary classification loss for distinguishing π_θ and π_E w.r.t. state-action pairs sampled from the occupancy measure ρ_θ and ρ_E.

To avoid the overfitting risks, this paper introduce the causal entropy $-H(\pi_\theta)$ (Ziebart (2010)). And thus the objective becomes the following form:

$$\min_\theta \max_w \mathcal{L} = -\eta(\pi_\theta) - \lambda_2 H(\pi_\theta) + \lambda_1 \left(\mathbb{E}_{\pi_\theta}[\log(D_w(s,a))] + \mathbb{E}_{\pi_E}[\log(1 - D_w(s,a))]\right). \tag{6}$$

It is actually a minimax problem similar to the GANs. In this POfD case, the true distribution is ρ_E, and the generator is to learn ρ_θ. D represents the discriminator parameterized by w, and we label the state-action pairs from expert as true ("1") while the policy as false ("0").

Moreover, by substituting Eq. (1) and Eq. (2) into Eq. (6), the objective becomes:

$$\min_\theta \max_w - \mathbb{E}_{\pi_\theta}[r'(s,a)] - \lambda_2 H(\pi_\theta) + \lambda_1 \mathbb{E}_{\pi_E}[\log(1 - D_w(s,a))], \tag{7}$$

This function results in a dynamic reward reshaping mechanism. And the reshaped reward function is:
$r'(s,a) = r(s,a) - \lambda_1 log(D_w(s,a))$

It can augment the environment reward with aid of demonstrations. In other words, when the environment feedback is sparse, this reshaped reward can force the policy to generate similar trajectory as π_E. So, such way can make the agent to explore the environment more efficiently.

3.3 POfD Algorithm

The Eq. (7) can be optimized by alternately updating the parameters w and θ of the discriminator and the agent policy, respectively. The overall optimization details are summarized in Alg. 1.

Algorithm 1 Policy optimization with demonstrations

Input: Expert demonstrations $\mathcal{D}_E = \{\tau_1^E, \ldots, \tau_N^E\}$, initial policy and discriminator parameters θ_0 and w_0, regularization weights λ_1, λ_2, maximal iterations I.
for $i = 1$ to I **do**
  Sample trajectories $\mathcal{D}_i = \{\tau\}, \tau \sim \pi_{\theta_i}$.
  Sample expert trajectories $\mathcal{D}_i^E \subset \mathcal{D}^E$.
  Update discriminator parameters from w_i to w_{i+1} with the gradient

$$\hat{\mathbb{E}}_{\mathcal{D}_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\mathcal{D}_i^E}[\nabla_w \log(1 - D_w(s,a))]$$

  Update the rewards in $\mathcal{D}_i$ with

$$r'(s,a) = r(a,b) - \lambda_1 \log(D_{w_i}(s,a)), \forall (s,a,r) \in \mathcal{D}_i$$

  Update the policy with policy gradient method (*e.g.*, TRPO, PPO) using the following gradient

$$\hat{\mathbb{E}}_{\mathcal{D}_i}[\nabla_\theta \log \pi_\theta(a|s) Q'(s,a)] - \lambda_2 \nabla_\theta H(\pi_{\theta_i})$$

end for

Note that the reshaped policy gradient is given by:

$$\nabla_\theta \mathbb{E}_{\pi_\theta}[r'(s,a)] = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q'(s,a)],$$
$$\text{where} \quad Q'(\bar{s},\bar{a}) = \mathbb{E}_{\pi_\theta}[r'(s,a)|s_0 = \bar{s}, a_0 = \bar{a}].$$

And the gradient of causal entropy is:

$$\nabla_\theta \mathbb{E}_{\pi_\theta}[-\log \pi_\theta(a|s)] = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q^H(s,a)],$$
$$\text{where } Q^H(\bar{s},\bar{a}) = \mathbb{E}_{\pi_\theta}[-\log \pi_\theta(a|s)|s_0 = \bar{s}, a_0 = \bar{a}].$$

4 Conclusion

Please provide succinct concluding remarks for your report. You may discuss the following aspects:

- The potential future research directions
- Any technical limitations
- Any latest results on the problem of interest
  Observing the new reward function $r'(s,a)$ in Eq. (7), we can find that if the environment feedback rewards are very sparse, then the agent behaves like an expert, which can guide the agent to learn from the expert; but when the environment itself has rewards, it is directly Relying on these environment rewards itself for learning, not relying on expert demonstrations. Therefore, it realizes learning experts first and then self-learning.
  However, the optimization problem of POfD begins with the demonstration guided regularization term which is a typical JS-divergence, and then is deduced into the form similar to GANs. As I know, the GAN method with the original JS-divergence-based objective is usually hard to be optimized and converge. So this POfD may have similar technical limitations like the early type of GAN, maybe it can refer to the subsequent developments of GAN to find a more useful technique to further finetune its current optimization method and get even better performance.

References

Fu, J., Luo, K., and Levine, S. (2017). Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial networks. *arXiv preprint arXiv:1406.2661*.

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Osband, I., et al. (2018). Deep q-learning from demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Kang, B., Jie, Z., and Feng, J. (2018). Policy optimization with demonstrations. In *International Conference on Machine Learning*, pages 2469–2478. PMLR.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. (2015). Trust region policy optimization. In *International conference on machine learning*, pages 1889–1897. PMLR.

Syed, U. and Schapire, R. E. (2008). A game-theoretic approach to apprenticeship learning. In *Advances in neural information processing systems*, pages 1449–1456.

Vecerik, M., Hester, T., Scholz, J., Wang, F., Pietquin, O., Piot, B., Heess, N., Rothörl, T., Lampe, T., and Riedmiller, M. (2017). Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*.

Ziebart, B. D. (2010). Modeling purposeful adaptive behavior with the principle of maximum causal entropy.