Policy Optimization with Demonstrations

Wei-Rong Chen
Department of Computer Science
National Yang Ming Chiao Tung University
wrchen.c@nycu.edu.tw

1 Introduction

In Reinforcement Learning (RL), we train the agent by the experiences which are from interacting with the environment and gaining the rewards. However, it will be hard to learn meaningful policies when the rewards are sparse and rare. This is because we can't effectily measure the value of the states or actions and find the interesting states in that case.

Some recently works are basically rooted in the following two ideas to tackle the exploration problems in RL. 1) Reshape the original reward function by encouraging the agent to visit states never seen before, driven by intrinsic curiosity (Pathak et al. [2017]) or information gain (Houthooft et al. [2016]). 2) Use demonstration trajectories sampled from an expert policy to guide the learning procedure, by either putting the demonstrations into a replay memory (Nair et al. [2018]) or using them to pretrain the policy in a supervised manner (Silver et al. [2016]). This paper proposes to combine the above two ideas by reshaping the reward function with demonstration information. It means that when the reward is not available, the agent will tend to follow the demonstrated behavior in early learning stages for exploration. Relatively, the agent will explore new states on its own after acquiring sufficient skills.

In summary, this paper has the following contributions:

- Propose a novel Policy Optimization from Demonstration (POfD) method, which can acquire knowledge from the demonstration data to boost exploration, even though the data are scarce and imperfect.
- Theoretically analyze the benefits brought by POfD to vanilla policy gradient ones, in terms of improvement over the expected return.
- Establish an optimization-friendly lower bound for the proposed objective.
- Show that existing replay memory-based learning from demonstration methods (Hester et al. [2017]) can be interpreted as degenerated cases of this method in terms of how to leverage the demonstration data.
- Compare POfD against 5 state-of-the-art baselines in sparse-reward environments and demonstrate that POfD surpasses all the well-established baselines in the experiments.

2 Problem Formulation

2.1 Preliminaries

Given a stochastic policy $\pi(a|s) = p(a|s;\pi)$ mapping from states to action probabilities, the performance of π is usually evaluated by its expected discounted reward $\eta(\pi)$:

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

where $(s0, a0, s1, ...)$ is a trajectory generated by executing policy π,*i.e.*, $s_0 \sim p_0$, $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim \mathcal{P}(\cdot|st, at)$.

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:*

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

Then the unnormalized distribution of state-action pairs $\rho_\pi(s, a) = \rho_\pi(s)\pi(a|s)$ is called occupancy measure of policy π.

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

2.2 Problem Definition

Generally, RL tasks and environments provide rewards just when the goal is reached. However, existing RL algorithms usually fail to explore and collect useful information in such sparse-reward settings. This paper contributes to solving this challenge by developing a method capable of boosting exploration with demonstrations and learning from sparse rewards.

Specifically, the agent is provided with some (and possibly imperfect) demonstrations $\mathcal{D}^E = \{\tau_1, \tau_2, ..., \tau_N\}$, where the i-th trajectory $\tau_i = \{(s_0^i, a_0^i), (s_1^i, a_1^i), ..., (s_T^i, a_T^i)\}$ is generated from executing an unknown expert policy π_E in the environment. The goal is to effectively leverage $\mathcal{D}^E$ and maximize $\eta(\pi)$ in Eqn. (1). In addition, there is an assumption on the quality of the expert policy:

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.*,*

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

3 Theoretical Analysis

To make the policy explore the environment by following the demonstration $\mathcal{D}^E$, this paper introduces a demonstration-guided exploration term $\mathcal{L}_M(\pi_\theta, \pi_E) = D_{JS}(\pi_\theta, \pi_E)$ to the learning objective. $\mathcal{L}_M(\pi_\theta, \pi_E)$ is defined over Jensen-Shannon divergence between the current policy and π_θ and the expert one π_E. 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. However, since π_E is unknown, we utilize the one-to-one correspondence (theorem 2 of Syed et al. [2008]) between the policy and the occupancy measure to modify the exploration term, and we get $\mathcal{L}_M \triangleq D_{JS}(\rho_\theta, \rho_E)$, where ρ_θ and ρ_E are short for ρ_{π_θ} and ρ_{π_E}. The finally proposed demonstration guided learning objective is

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

In particular, instead of performing optimization on the difficult Jensen-Shannon divergence directly, we optimize its lower bound given as follows.

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

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

With the above theorem, the occupancy measure matching objective $\mathcal{L}_M$ can be written as

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

where $D(s,a) = \log(\frac{1}{1+e^{-U(s,a)}}) : \mathcal{S} \times \mathcal{A} \to (0,1)$ is an arbitrary mapping function followed by a sigmoid activation function. Suppose D is parameterized by w, the labeling expert state-action pairs as true ("1") and policy state-action pairs as false ("0"), we get the following objective,

$$\min_{\theta} \max_{w} \mathcal{L} = -\eta(\pi_\theta) - \lambda_2 H(\pi_\theta) + \lambda_1 (\mathbb{E}_{\pi_\theta}[\log(D_w(s,a))] + \mathbb{E}_{\pi_E}[\log(1 - D_w(s,a))]),$$

where $H(\pi_\theta)$ is entropy term to avoid potential over-fitting risks. We can rewrite the function above by substitute η with Eqn. (1) and (2) and get the reshaped reward function as

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

where $r'(s,a) = r(a,b) - \lambda_1 \log(D_w(s,a))$ is the reshaped reward function. With demonstration information in reward function, the agent will tend to follow the expert policy π_E when the exploration is insufficient. In other words, the divergence of π and π_E is minimized.

However, it seems strange that if π is far away from π_E, $\mathcal{D}_w(s,a)$ will be close to 0 and $-\log(\mathcal{D}_w(s,a))$ will be large. It makes the agent prefer to stay away from π_E when the reward is not available because it will receive more reward r'.

We can optimize the above objective by updating policy parameters θ and discriminator parameters w. The gradient is given by

$$\mathbb{E}_{\pi_\theta}[\nabla_w \log(D(s,a))] + \mathbb{E}_{\pi_E}[\nabla_w \log(1 - D(s,a))].$$

The reshaped policy gradient is

$$\nabla_\theta \mathbb{E}_{\pi_\theta}[r'(s,a)] = \mathbb{E}_{\pi_\theta}[\nabla_\theta \log \pi_\theta(a|s) Q'(s,a)],$$

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

The gradient for causal entropy regularization is given by

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

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

4 Conclusion

It is a significant problem that exploring the environment in sparse-reward settings. This paper proposes a method, POfD, that combines sparse rewards and demonstrations to prompt the agent to explore the environment more efficiently. With demonstration information in the reshaped reward function, the agent is capable of dynamically adjusting the tendency to update the policy between own exploration and expert demonstrations. To authors' best knowledge, POfD is the first one that can acquire knowledge from few and imperfect demonstration data to aid exploration in environments with sparse feedback.

References

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

Ashvin Nair, Bob McGrew, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6292–6299. IEEE, 2018.

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):484–489, 2016.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Andrew Sendonaris, Gabriel Dulac-Arnold, Ian Osband, John Agapiou, et al. Learning from demonstrations for real world reinforcement learning. 2017.

Umar Syed, Michael Bowling, and Robert E Schapire. Apprenticeship learning using linear programming. In *Proceedings of the 25th international conference on Machine learning*, pages 1032–1039, 2008.