A Note on Policy Optimization with Demonstrations

Yun-Ming, Chan
Department of Computer Science
National Yang Ming Chiao Tung University
yunming.cs07@nctu.edu.tw

1 Introduction

Exploration is still a big challenge in reinforcement learning, especially in environments with sparse rewards. Common exploration strategies, like ϵ-greedy or noise-based method, can't explore effectively to learn meaningful policy in an environment with sparse rewards because random exploration can rarely find some good states.

Some works use expert demonstration trajectories, by adding expert data into replay memory or pre-training the policy, to guide the agent to explore in sparse reward environments. The performance of these methods seems to be great. However, these methods require a lot of high-quality demonstration data.

To deal with these difficult problems, Policy Optimization from Demonstration (POfD) method was proposed in this paper (Kang et al. [2018]). POfD can learn from demonstration data to help exploration in environments with sparse reward even when the demonstration data are few and imperfect. The experiments in the paper also shows that POfD can just learn from one imperfect demonstration trajectory, but can reach the same performance as TRPO (Schulman et al. [2015]) trained in the same environment with dense reward.

2 Problem Formulation

2.1 Preliminaries

$(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ is the MDP, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P(s'|s,a)$ is the transition distribution, $r(s,a)$ is the reward function, and $\gamma \in (0,1)$ is the discount factor.

$\pi(a|s)$ is a stochastic policy. The expected discounted reward by following π is $\eta(\pi)$

$$\eta(\pi) = \mathbb{E}_\pi \left[r(s,a) \right] = \mathbb{E}_{(s_0,a_0,s_1,a_1,\ldots)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \tag{1}$$

where $(s_0, a_0, s_1, a_1, ...)$ is a trajectory generated by the policy π. The value function $V^\pi(s) = \mathbb{E}_\pi \left[r(\cdot,\cdot) | s_0 = s \right]$, the action value function $Q^\pi(s,a) = \mathbb{E}_\pi \left[r(\cdot,\cdot) | s_0 = s, a_0 = a \right]$, and the advantage function $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$. $\eta(\pi) = \mathbb{E}_{s\sim\mu}[V^\pi(s)]$, where μ is the initial state distribution. $D_{KL}(p,q)$ is Kullback–Leibler divergence. $D_{JS}(p,q)$ is Jensen–Shannon divergence.

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.

Then, we can write $\eta(\pi)$ as

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

Lemma 1. *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.*

Lemma 2. *(Lemma 1 in Schulman et al. [2015]) Given two policies π, $\tilde{\pi}$,*

$$\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{\tau\sim\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi}(s_t,a_t)\right]$$

This expectation is taken over trajectories $\tau := (s_0, a_0, s_1, a_1, \dots)$, and the notation $\mathbb{E}_{\tau\sim\tilde{\pi}}[\dots]$ indicates that actions are sampled from $\tilde{\pi}$ to generate τ.

2.2 Proposed Objective

Let $\mathcal{D}^E$ denotes a set of expert demonstration trajectories, where each trajectory is sampled from an unknown expert policy π_E.

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

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

The expert policy that used to generate demonstration data only needs to satisfy Assumption 1. That is, the expert policy do not need to be perfect. Also, we do not need a lot of demonstration data.

Suppose π_θ is a θ-parameterized policy and is differentiable. In policy gradient methods, we want to maximize the expected discounted reward $\eta(\pi_\theta)$. POfD introduces a demonstration-guided exploration term $D_{JS}(\pi_\theta, \pi_E)$ into the objective of policy gradient methods. It gives the objective function

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

where λ_1 is a trading-off parameter. Thus, the policy π_θ was encouraged to explore around the expert policy π_E. However, $D_{JS}(\pi_\theta, \pi_E)$ is hard to compute since π_E is an unknown expert policy. Fortunately, by Lemma 1, we can change the polices to occupancy measures, which gives

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

ρ_π, ρ_E are abbreviations for $\rho_{\pi_\theta}, \rho_{\pi_E}$.

3 Theoretical Analysis

3.1 Benefits of Exploration with Demonstrations

With Lemma 2, we have

$$\begin{aligned}\eta(\pi) &= \eta(\pi_{old}) + \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty}\gamma^t A^{\pi_{old}}(s_t,a_t)\right] \\ &= \eta(\pi_{old}) + \sum_{t=0}^{\infty}\sum_{s} P(s_t = s|\pi)\sum_{a}\pi(a|s)\gamma^t A^{\pi_{old}}(s,a) \\ &= \eta(\pi_{old}) + \sum_{s}\rho_{\pi}(s)\sum_{a}\pi(s|a)A^{\pi_{old}}(s,a)\end{aligned}$$

And we use the surrogate function $J_{\pi_{old}}(\pi)$ to locally approximate $\eta(\pi)$.

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

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

With Theorem 1, 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 $C_{\pi_E} = \frac{8\gamma\epsilon_E}{(1-\gamma)^2}, C_\pi = \frac{2\gamma\epsilon_\pi}{(1-\gamma)^2}, \hat{\delta} = \frac{\delta}{1-\gamma}$. Then, we have

$$\eta(\pi_{i+1}) \geq M_i(\pi_{i+1}), \tag{3}$$

$$\eta(\pi_i) = M_i(\pi_i) + C_{\pi_E} D_{JS}^{max}(\pi_i, \pi_E) - \hat{\delta}, \tag{4}$$

$$\eta(\pi_{i+1}) - \eta(\pi_i) \geq M_i(\pi_{i+1}) - M_i(\pi_i) - C_{\pi_E} D_{JS}^{max}(\pi_i, \pi_E) + \hat{\delta} \tag{5}$$

Since RHS of (5) may be negative, we do not have a monotonic policy improvement guarantee. But it brings improvement with a margin $\hat{\delta}$ over the pure policy gradient methods when π is close to π_E.

3.2 Optimization Algorithm

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

$$D_{JS}(\rho_\pi, \rho_E) \geq \frac{1}{2} \sup_U \left(\mathbb{E}_{\rho_\pi}[h(U(s,a))] + \mathbb{E}_{\rho_E}[\bar{h}(U(s,a))] \right) + \log 2$$

The theorem given in the paper is wrong. In the paper, it gives

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

which is different from the correct one in a constant $\frac{1}{2}$. I will show it later. Nevertheless, it does not affect the following discussion.

With Theorem 2 and equation (2), we can rewrite the objective function into

$$\begin{aligned}
\mathcal{L}(\pi_\theta) &= -\eta(\pi_\theta) + \lambda_1 \sup_{D \in (0,1)^{\mathcal{S} \times \mathcal{A}}} \left(\mathbb{E}_{\rho_\pi}[\log(D(s,a))] + \mathbb{E}_{\rho_E}[\log(1 - D(s,a))] \right) \\
&= -\eta(\pi_\theta) + \lambda_1 \sup_{D \in (0,1)^{\mathcal{S} \times \mathcal{A}}} \left(\mathbb{E}_{\pi_\theta}[\log(D(s,a))] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))] \right)
\end{aligned}$$

where $D(s,a) = \frac{1}{1+e^{-U(s,a)}} : \mathcal{S} \times \mathcal{A} \to (0,1)$. The supremum over $D(s,a)$ thus represent the optimal binary classification of distinguishing state-action pairs sampled from ρ_π and ρ_E, where the expert state-action pairs are labeled as 1 and the policy state-action pairs are labeled as 0.Suppose D is parameterized by w, the objective becomes

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

It is a minimax objective similar to Generative Adversarial Networks (GANs). When maximizing over w, it makes π_θ close to π_E. When minimizing over θ, it maximizes $\eta(\pi_\theta)$ and pushes π_θ away from π_E. Therefore, it can help π_θ to explore around the π_E.

Besides, POfD also introduce causal entropy $-H(\pi_\theta)$ into the objective to avoid overfitting. Therefore, the objective of proposed POfD is

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

Moreover, we can write the objective as

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

where $r'(s,a) = r(s,a) - \lambda_1 \log(D(s,a))$ is the reshaped reward function. Reward reshaping is another common approach to do better exploration in environment with sparse rewards. Thus, we can find out that POfD is a method that combining learning from demonstrations and dynamic reward reshaping.

Now, we have a optimization algorithm Algorithm 1 which is compatible with any policy gradient methods. Let $\hat{\mathbb{E}}_{\mathcal{D}}$ denote the empirical expectation estimated from trajectories $\mathcal{D}$. The reshaped policy gradient is

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

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

The gradient for causal entropy $\nabla_\theta H(\pi_\theta)$ is given by

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

Algorithm 1 Policy Optimization with Demonstration

Input: Expert demonstration demonstration $\mathcal{D}_E = \{\tau_1^E, \ldots, \tau_N^E\}$, 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

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

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

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

  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

3.3 Proof of Theorem 2

Proof. Let $f(v) = v \log(v) - (v+1)\log(v+1)$. The domain of $f(v)$ is $(0, +\infty)$ and is convex. The second derivative of $f(v)$ is

$$f''(v) = (\log v + 1 - \log(v+1) - 1)' = \frac{1}{v} - \frac{1}{1+v} = \frac{1}{v(1+v)} \geq 0$$

Therefore, $f(v)$ is a convex function.

Let f^* be the convex conjugate function of $f(v)$ and given by $f^*(t) = \sup_{v \in \text{dom}_f} \{vt - f(v)\}$. Since $f(v)$ is convex and continuous, $f(v) = f^{**}(v) = \sup_{t \in \text{dom}_{f^*}} \{tv - f^*(t)\}$, where f^{**} is the convex conjugate function of f^*.

$f^*(t) = \sup_{v \in \text{dom}_f} \{vt - f(v)\}$. The supremum is reached when $\frac{\partial}{\partial v}(vt - f(v)) = 0$.

$$
\begin{aligned}
&\frac{\partial}{\partial v}(vt - f(v)) = 0 \\
\Rightarrow &\frac{\partial}{\partial v}(vt - f(v)) = t - (\log v + 1 - \log(v+1) - 1) = 0 \\
\Rightarrow &t = \log \frac{v}{v+1} = -\log(1 + \frac{1}{v}) \\
\Rightarrow &v = \frac{1}{-1 + e^{-t}}
\end{aligned}
$$

Substituting the result into $f^*(t)$, gives

$$
\begin{aligned}
f^*(t) &= \sup_{v \in \text{dom}_f} \{vt - f(v)\} \\
&= \frac{t}{-1+e^{-t}} - \frac{1}{-1+e^{-t}} \log \frac{1}{-1+e^{-t}} + \frac{e^{-t}}{-1+e^{-t}} \log \frac{e^{-t}}{-1+e^{-t}} \\
&= \frac{te^t}{1-e^t} - \frac{e^t}{1-e^t} \log \frac{e^t}{1-e^t} + \frac{1}{1-e^t} \log \frac{1}{1-e^t} \\
&= \frac{1}{1-e^t} \left(te^t - e^t \log e^t + e^t \log(1-e^t) - \log(1-e^t) \right) \\
&= -\log(1-e^t) = \log \frac{1}{1-e^t}
\end{aligned}
$$

The domain of $f^*(t)$ is $(-\infty, 0)$.

$$
\begin{aligned}
&D_{JS}(\rho_\pi, \rho_E) \\
&= \frac{1}{2} D_{KL}(\rho_\pi, \frac{\rho_\pi + \rho_E}{2}) + \frac{1}{2} D_{KL}(\rho_E, \frac{\rho_\pi + \rho_E}{2}) \\
&= \frac{1}{2} \int_{\mathcal{S} \times \mathcal{A}} \rho_\pi \log \frac{2\rho_\pi}{\rho_\pi + \rho_E} + \rho_E \log \frac{2\rho_E}{\rho_\pi + \rho_E} dsda \\
&= \frac{1}{2} \int_{\mathcal{S} \times \mathcal{A}} \rho_\pi \log \rho_\pi + \rho_E \log \rho_E - (\rho_\pi + \rho_E) \log(\rho_\pi + \rho_E) dsda + \log 2 \\
&= \frac{1}{2} \int_{\mathcal{S} \times \mathcal{A}} \rho_\pi \log \rho_\pi - \rho_\pi \log \rho_E + \rho_\pi \log \rho_E + \rho_E \log \rho_E - (\rho_\pi + \rho_E) \log(\rho_\pi + \rho_E) dsda + \log 2 \\
&= \frac{1}{2} \int_{\mathcal{S} \times \mathcal{A}} \rho_\pi \log \frac{\rho_\pi}{\rho_E} - (\rho_\pi + \rho_E) \log \frac{\rho_\pi + \rho_E}{\rho_E} dsda + \log 2 \\
&= \frac{1}{2} \int_{\mathcal{S} \times \mathcal{A}} \rho_E \left(\frac{\rho_\pi}{\rho_E} \log \frac{\rho_\pi}{\rho_E} - (\frac{\rho_\pi}{\rho_E} + 1) \log(\frac{\rho_\pi}{\rho_E} + 1) \right) dsda + \log 2 \\
&= \frac{1}{2} \int_{\mathcal{S} \times \mathcal{A}} \rho_E f(\frac{\rho_\pi}{\rho_E}) dsda + \log 2 \\
&= \frac{1}{2} \int_{\mathcal{S} \times \mathcal{A}} \rho_E \sup_{t \in \text{dom}_{f^*}} \left(t \frac{\rho_\pi}{\rho_E} - f^*(t) \right) dsda + \log 2 \\
&\geq \frac{1}{2} \sup_{t \in \text{dom}_{f^*}} \left(\int_{\mathcal{S} \times \mathcal{A}} t\rho_\pi - \rho_E f^*(t) dsda \right) + \log 2 \\
&= \frac{1}{2} \sup_{T \in \mathcal{T}} \left(\mathbb{E}_{(s,a) \sim \rho_\pi}[T(s,a)] + \mathbb{E}_{(s,a) \sim \rho_E}[-f^*(T(s,a))] \right) + \log 2
\end{aligned}
\tag{6}
$$

Replacing t by $T(s,a)$, $\mathcal{T} = \{T(s,a) : \mathcal{S} \times \mathcal{A} \to \text{dom}_{f^*}\}$. $T(s,a)$ is valid if and only if $\text{range}_T = \text{dom}_{f^*}$. And $T(s,a)$ can be formed with $h(U(s,a))$. $U(s,a) \in \mathbb{R}, h(u) \in (-\infty, 0)$. $\text{range}_T \in (-\infty, 0)$. Thus, $\text{range}_T = \text{dom}_f^* \in (-\infty, 0)$.

And,

$$\begin{aligned}
-f^*(T(s,a)) &= -f^*(h(U(s,a))) \\
&= -\log \frac{1}{1-e^{\log \frac{1}{1+e^{-U(s,a)}}}} \\
&= \log\left(1 - \frac{1}{1+e^{-U(s,a)}}\right) \\
&= \log \frac{e^{-U(s,a)}}{1+e^{-U(s,a)}} \\
&= \bar{h}(U(s,a))
\end{aligned}$$

Then, (6) becomes

$$D_{JS}(\rho_\pi, \rho_E) \geq \frac{1}{2} \sup_U \left(\mathbb{E}_{(s,a)\sim\rho_\pi}[h(U(s,a))] + \mathbb{E}_{(s,a)\sim\rho_E}[\bar{h}(U(s,a))]\right) + \log 2$$

□

3.4 Error in the Paper

The following is a part of the original proof of the theorem.

$$\begin{aligned}
& D_{JS}(\rho_\pi, \rho_E) \\
&= \int_{\mathcal{S}\times\mathcal{A}} \rho_\pi \log \frac{2\rho_\pi}{\rho_\pi + \rho_E} + \rho_E \log \frac{2\rho_E}{\rho_\pi + \rho_E} dsda \\
&= \dots \\
&\geq \sup_U \left(\mathbb{E}_{(s,a)\sim\rho_\pi}[h(U(s,a))] + \mathbb{E}_{(s,a)\sim\rho_E}[\bar{h}(U(s,a))]\right) + \log 4
\end{aligned}$$

We can find out that $\frac{1}{2}$ is missing in the first equation of the original proof, and it leads to a wrong result. Therefore, the theorem given in the paper is incorrect.

4 Conclusion

Although POfD is quite powerful and only need few demonstrations, it does not guarantee monotonic policy improvement as we mentioned. Therefore, it is possible that POfD will stuck at some suboptimal policy. I think the policy will especially stuck at some local optimal policy around the expert policy. If the optimal policy is far from the expert policy, then it is highly likely that POfD cannot find optimal policy. Thus, I think this is the major limitation of POfD.

References

Bingyi Kang, Zequn Jie, and Jiashi Feng. Policy optimization with demonstrations. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2469–2478. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/kang18a.html`.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2015. URL `https://arxiv.org/abs/1502.05477`.