Truncated Horizon Policy Search : Combining Reinforcement Learning and Imitation Learning

Wen-Tao,Chu
Department of Computer Science
National Yang Ming Chiao Tung University
wentao.cs10@nycu.edu.tw

1 Introduction

In this paper, we propose a method combining reinforcement and imitation learning. We focus on the effectiveness of the near-optimal cost-to-go oracle on the planning horizon and show that it shortens the learner's planning horizon as the function of its accuracy. Interestingly, some advanced technical approaches are used, such as *Reward Shaping* and *truncating the planning horizon*. In particularly, *Reward Shaping* is most important technique, which is the core and insights for this paper. We will briefly introduce them later.

Imitation learning can speed up the learning process in RL tasks, but it has an intractable problem which the performance is limited to the expert. Previous works such as Data Aggregation (DAgger)(Ross et al. [2011]) and Aggregation with Values (AGGREVATE)(Ross and Bagnell [2014]) give us some theoretical guarantees and show that a policy performing as well as the expert policy or a one-step improvement over the expert policy. Based on the above, it's unfortunate that if the expert is sub-optimal, than the learned policy will return a sub-optimal policy. Therefore, we combine the RL and IL, try to overcome this problem and exploring more how to improve upon the expert with RL.

With the above reasons, we propose the algorithm named ***Truncated HORizion policy search with cost-to-go oracle*** (THOR), which shapes the cost using $\hat{V}^e$ (the *imperfect estimator* of the cost-to-go of some expert π^e, and search for a policy by optimizing the truncated planning horizon of the new MDP.

2 Problem Formulation

2.1 Preliminaries and Notation

First, we consider the basic problem of optimizing Marcov Decision Process (MDP) defined as $M = (S, A, P, C, \gamma)$. In this MDP, S is a set of states, A is a set of actions, P is a transition probability which usually denoted by $P_{sa} \doteq P(s'|s, a)$ for any $s \in S$, $s' \in S$, $a \in A$. Instead of using Reward function R as usual, we consider the Cost function $c(s, a) \in C$, and a discount factor $\gamma \in [0, 1]$.

Second, we define the value function V^π_M and Q^π_M, and advantage function A^π_M and *k-step version* as the following.

- $V^\pi_M(s) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t)]$
- $Q^\pi_M(s, a) = \mathbb{E}[c(s, a) + \gamma \mathbb{E}_{s' \sim P_{sa}}[V^\pi_M(s')]]$
- $A^\pi_M(s, a) = Q^\pi_M(s, a) - V^\pi_M(s)$
- $V^{\pi,k}_{M'}(s) = \mathbb{E}[\sum_{t=1}^{k} \gamma^{t-1} c'(s_t, a_t)]$
- $Q^{\pi,k}_{M'}(s, a) = \mathbb{E}[c'(s, a) + \gamma \mathbb{E}_{s' \sim P_{sa}}[V^{\pi,k}_{M'}(s')]]$

Note that the optimal policy π^* defined by $\pi^* = \arg\min_{\pi} V^{\pi}(s), \forall s \in S$

Assumption 1 *Suppose we access to an cost-to-go oracle $\hat{V}^e(s) : S \to \mathbb{R}$, and note that it is not necessary that $\hat{V}^e(s)$ equal to V^*_M.*

2.2 Cost Shaping

The *Reward Shaping* is a very technique method to find the optimal policy between M and $M^{'}$, where $M^{'}$ is new MDP defined by $M^{'} = (S, A, P, C', \gamma)$. Ng et al. [1999] showed that $\pi^*_{M'}(s) = \pi^*_M(s)$, $\forall s \in S$, i.e., if we can find $\pi^*_{M'}(s)$ on $M^{'}$, then we also find $\pi^*_M(s)$. Again, We use the cost instead of the reward for this problem (but the reward is used in the experiment because the environment's basic setting).

Consider M and any potential function $\Phi : S \to \mathbb{R}$, we reshape $c(s, a)$ sampled from $C(s, a)$ as the following :

$$c'(s, a) = c(s, a) + \gamma\Phi(s') - \Phi(s) \tag{1}$$

2.3 Truncating the planning horizon

By previous works, we briefly say the advantages of *Truncating the planning horizon*. It is resulting in a trade-off not only between accuracy and computation complexity, but also bias and variance.

3 Theoretical Analysis

In this section we study the relationship of planning horizon with imperfect estimator $\hat{V}^e(s)$. Here we define the expected total cost as $J(\pi) = \mathbb{E}_{s_0 \sim v}[V^{\pi}_M(s_0)]$,for some initial state distribution v and for all π.

Theorem 1 *Given M and imperfect estimator $\hat{V}^e(s)$ with $|\hat{V}^e(s) - V^*_M(s)| = \epsilon$, for some $\epsilon > 0$, such that the lower bound between the performance of the induced policy from the cost-to-go oracle $\hat{\pi}^* = \arg\min_{a} Q^{\pi}_M(s, a)$ and the optimal policy π^*.*

$$J(\hat{\pi}^*) - J(\pi^*) \geq \Omega(\frac{\gamma}{1-\gamma}\epsilon) \tag{2}$$

Now, we look at the outperforming expert, and consider the *k-steps* total cost of a policy :

$$\mathbb{E}[\sum_{i=1}^{k} \gamma^{i-1} c'(s_i, a_i)] = \mathbb{E}[\sum_{i=1}^{k} \gamma^{i-1} c(s_i, a_i) + \gamma^k \hat{V}^e(s_{k+1}) - \hat{V}^e(s_1)], \forall s \tag{3}$$

The above equation is easy to get when we use the definition of cost shaping and $\hat{V}^e$ as a potential function.

Theorem 2 *Assume that $\hat{\pi}^*$ minimizes Eq.3 with $k > 1$ and $|\hat{V}^e(s) - V^*_M(s)| = \Theta(\epsilon)$, $\forall s \in S$:*

$$J(\hat{\pi}^*) - J(\pi^*) \leq O(\frac{\gamma^k}{1-\gamma^k}\epsilon) \tag{4}$$

Note that we do not consider $k = 1$ (one-step), since it is not sufficient to guarantee near-optimal performance. We will provide the proof for Theorem 2 later.

4 Algorithm

In this section, we focus on the updated policy gradient of THOR. Denote $\pi_{\theta,n}$ as the current policy, $P_{\bar{\pi}_{\theta,n}}(\cdot)$ as the average state distribution and H is the original planning horizon. The policy gradient is :

$$\begin{aligned}
&\mathbb{E}_{s\sim P_{\bar{\pi}_{\theta,n}}}\mathbb{E}_{\tau_k\sim\pi_{\theta,n}}[\sum_{t=1}^{k}\nabla_\theta \ln \pi_\theta(a_t|s_t)(\sum_{j=i}^{k+i}\gamma^{j-i}c'(s_j,a_j))] \\
&\approx \mathbb{E}_{s\sim P_{\bar{\pi}_{\theta,n}}}\mathbb{E}_{\tau_k\sim\pi_{\theta,n}}[\sum_{t=1}^{k}\nabla_\theta \ln \pi_\theta(a_t|s_t)Q_M^{\pi,k}(s_i,a_i)] \\
&\approx k\sum_{\tau}\sum_{t=1}^{|\tau|}\nabla_\theta \ln \pi_\theta(a_t|s_t)\hat{A}_M^{\pi,k}(s_i,a_i)/H
\end{aligned}$$

Note that the last approximation is by Generalized Advantage Estimator (GAE).

In particular,when $k=1$ and $\hat{V}^e = V^*_{M_0}$, we have the following result :

$$\begin{aligned}
&\mathbb{E}_\tau[\sum_{t=1}^{|\tau|}\nabla_\theta \ln \pi_\theta(a_t|s_t)A_M^{\pi^{e,1}}(s_i,a_i)] \\
&= \mathbb{E}_\tau[\sum_{t=1}^{|\tau|}\nabla_\theta \ln \pi_\theta(a_t|s_t)Q_M^{\pi^{e,1}}(s_i,a_i)] \\
&= \mathbb{E}_\tau[\sum_{t=1}^{|\tau|}\nabla_\theta \ln \pi_\theta(a_t|s_t)\mathbb{E}[c'(s_t,a_t)]] \\
&= \mathbb{E}_\tau[\sum_{t=1}^{|\tau|}\nabla_\theta \ln \pi_\theta(a_t|s_t)A_M^{\pi^*}(s_i,a_i)]
\end{aligned}$$

5 Conclusion

This is an interesting method combining imitation and reinforcement learning with the concept of reward shaping. In addition, we truncate the planning horizon to shorten the length of trajectory to attain the smaller variance. finally, we also discuss the special case($k=1$), where is same goal to the previous work AGGREVATE, and give some theoretical guarantee for imperfect estimator.

The following will discuss some issue and the potential future work for this paper in my perspective.

- Is *Assumption 1* necessary? i.e.,$\hat{V}^e$ can equal to V^*_M. It has not an effect on this paper, since we have discussed all the case for it and it's more common that $\hat{V}^e$ is imperfect.
- In this paper, we truncated the infinite long horizon to finite horizon. But can we truncated the finite long horizon to finite short horizon? Even though the finite long horizon will be truncated naturally.
- Can online update $\hat{V}^e$ by expert's information during training? This maybe an efficient way to train the policy, since it can update simultaneously and might convergent to the optimal policy faster.

6 Proof for Theorem 2

Consider $V^{\hat{\pi}^*}(s) - V^*(s), \forall s \in S$, we have the following result :

$$
\begin{aligned}
V^{\hat{\pi}^*}(s) - V^*(s) &= \mathbb{E}_{\hat{\pi}^*}[\sum_{i=1}^{k} \gamma^{i-1} c(s_i, a_i) + \gamma^k V^{\hat{\pi}^*}(s_{k+1})] - \mathbb{E}_{\pi^*}[\sum_{i=1}^{k} \gamma^{i-1} c(s_i, a_i) + \gamma^k V^*(s_{k+1})] \\
&= \gamma^k \mathbb{E}_{\hat{\pi}^*}[V^{\hat{\pi}^*}(s_{k+1}) - V^*(s_{k+1})] + \mathbb{E}_{\hat{\pi}^*}[\sum_{i=1}^{k} \gamma^{i-1} c(s_i, a_i) + \gamma^k V^*(s_{k+1})] \qquad (5) \\
&+ \mathbb{E}_{\pi^*}[\sum_{i=1}^{k} \gamma^{i-1} c(s_i, a_i) + \gamma^k V^*(s_{k+1})] \\
&\leq \gamma^k \mathbb{E}_{\hat{\pi}^*}[V^{\hat{\pi}^*}(s_{k+1}) - V^*(s_{k+1})] + (\mathbb{E}_{\hat{\pi}^*}[\sum_{i=1}^{k} \gamma^{i-1} c(s_i, a_i) + \gamma^k \hat{V}^e(s_{k+1})] + \gamma^k \epsilon) \qquad (6) \\
&+ (-\mathbb{E}_{\pi^*}[\sum_{i=1}^{k} \gamma^{i-1} c(s_i, a_i) + \gamma^k \hat{V}^e(s_{k+1})] + \gamma^k \epsilon) \\
&\leq \gamma^k \mathbb{E}_{\hat{\pi}^*}[V^{\hat{\pi}^*}(s_{k+1}) - V^*(s_{k+1})] + 2\gamma^k \epsilon \\
&+ (\mathbb{E}_{\hat{\pi}^*}[\sum_{i=1}^{k} \gamma^{i-1} c(s_i, a_i) + \gamma^k \hat{V}^e(s_{k+1})] - \mathbb{E}_{\pi^*}[\sum_{i=1}^{k} \gamma^{i-1} c(s_i, a_i) + \gamma^k \hat{V}^e(s_{k+1})]) \qquad (7) \\
&\leq \gamma^k \mathbb{E}_{\hat{\pi}^*}[V^{\hat{\pi}^*}(s_{k+1}) - V^*(s_{k+1})] + 2\gamma^k \epsilon \qquad (8)
\end{aligned}
$$

Remark

- In Eq.(5), we add and minus the new term : $\mathbb{E}_{\hat{\pi}^*}[\sum_{i=1}^{k} \gamma^{i-1} c(s_i, a_i) + \gamma^k V^*(s_{k+1})]$
- In Eq.(6), we use the property $|\hat{V}^e(s) - V^*(s)| = \Theta(\epsilon)$.
- In Eq.(7), considering the $\hat{\pi}^*$ and π^*.

By recursion for $V^{\hat{\pi}^*}(s_{k+1}) - V^*(s_{k+1})$, we have :

$$
V^{\hat{\pi}^*}(s) - V^*(s) \leq 2\gamma^k \epsilon (1 + \gamma^k + \gamma^{2k} + ...) = \frac{2\gamma^k}{1 - \gamma^k} \epsilon = O(\frac{\gamma^k}{1 - \gamma^k} \epsilon), \forall s
$$

Hence, we get $J(\hat{\pi}^*) - J(\pi^*) \leq O(\frac{\gamma^k}{1-\gamma^k} \epsilon)$.

□

References

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

Stephane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.

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.