Report of "When to Trust Your Model: Model-Based Policy Optimization"

Ru-Jun Wang
Department of Computer Science
National Tsing Hua University, Hsinchu, Taiwan
s107062649@m107.nthu.edu.tw

1 Introduction

Reinforcement learning algorithms principally belong to two categories, model-based and model-free methods. Model-based reinforcement learning methods explicitly build a predictive transition model of the environment to derive the synthetic sample data points. Conversely, model-free methods aim to directly map a value to state or state-action pair, which abstain from training a transition model of environments. When dealing with real-world physical systems, data collection can be an arduous process, model-based methods achieve better sample efficiency. However, model-based methods often struggle to achieve the same asymptotic performance as their model-free counterparts due to the inevitable errors between the learned model and the real dynamics. Thus, how to effectively use a predictive model for policy optimization while keeping the sample efficiency is research focus of this paper.

Based on the previous model-based methods, Janner et al. [2019] find out full rollouts through model will result in failure on long-horizon tasks. However, sufficiently short rollouts suffer from worse performance of model exploitation. Thus, this paper incorporates model-free and model-based methods by modifying the lengths of rollouts through the model. Specifically, the authors prove the bound of monotonic model-based improvement with generalization error and distribution shift. Motivated by the limited use of truncated, but nonzero-length, model rollouts, the authors introduce Model-based policy optimization (MBPO), which disentangles the task horizon and model horizon by querying the model only for short rollouts. The empirical analysis shows that MBPO achieves asymptotic performance rivaling the best model-free algorithms, faster learning speed than prior model-free or model-based methods, and better generalization to long-horizon tasks.

2 Preliminaries

Markov decision process (MDP):
Defined by the tuple $(\mathcal{S}, \mathcal{A}, p, r, \gamma, \rho_0)$. $\mathcal{S}$ and $\mathcal{A}$ are the state and action spaces, respectively, and $\gamma \in (0, 1)$ is the discount factor. The dynamics or transition distribution are denoted as $p(s'|s; a)$, the initial state distribution as $\rho_0(s)$, and the reward function as $r(s, a)$. The goal of reinforcement learning is to find the optimal policy π^* that maximizes the expected sum of discounted rewards, denoted by η:

$$\pi^* = \arg\max_{\pi} \eta[\pi] = \arg\max_{\pi} E_{\pi}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)].$$

The dynamics $p(s'|s, a)$ are assumed to be unknown. Model-based reinforcement learning methods aim to construct a model of the transition distribution, $p_\theta(s'|s, a)$, using data collected from interaction with the MDP, typically using supervised learning. We additionally assume that the reward function has unknown form, and predict r as a learned function of s and a.

3 Theoretical Analysis

In Lemma B.4, the notation of "pre" and "post" in the proof define the time step before the branch point k and the time step after the point k, respectively. Thus, we should modify the notations of "pre" and "post" in the proof as the following:
For $t \leq k$:

$$D_{TV}(d_1^t(s,a)||d_2^t(s,a)) \leq t(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \epsilon_\pi^{pre} \leq k(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \epsilon_\pi^{pre}$$

For $t \geq k$:

$$D_{TV}(d_1^t(s,a)||d_2^t(s,a)) \leq k(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + (t-k)(\epsilon_m^{post} + \epsilon_\pi^{post}) + \epsilon_\pi^{pre} + \epsilon_\pi^{post}$$

The modified bound in proof of Lemma B.4 will be:

$$\begin{aligned}
D_{TV}(d_1^t(s,a)||d_2^t(s,a)) &\leq (1-\gamma)\sum_{t=0}^{\infty}\gamma^t D_{TV}(d_1^t(s,a)||d_2^t(s,a)) \\
&\leq (1-\gamma)\sum_{t=0}^{k}\gamma^t(k(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \epsilon_\pi^{pre}) \\
&+ (1-\gamma)\sum_{t=k+1}^{\infty}\gamma^t(k(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + (t-k)(\epsilon_m^{post} + \epsilon_\pi^{post}) + \epsilon_\pi^{pre} + \epsilon_\pi^{post}) \\
&= k(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \epsilon_\pi^{pre} + \frac{\gamma^{k+1}}{1-\gamma}(\epsilon_m^{post} + \epsilon_\pi^{post}) + \gamma^k\epsilon_\pi^{post}
\end{aligned}$$

Thus, the new bound induced by above inequalities should be:

$$|\eta_1 - \eta_2| \leq 2r_{max}[\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_m^{post} + \epsilon_\pi^{post}) + \frac{k}{1-\gamma}(\epsilon_m^{pre} + \epsilon_\pi^{pre}) + \frac{\gamma^{k+1}}{1-\gamma}\epsilon_\pi^{post} + \frac{1}{1-\gamma}\epsilon_\pi^{pre}]$$

Since the authors use Lemma B.4 to derive Theorem 4.2 (i.e., Theorem A.3) and Theorem 4.3 (i.e., Theorem A.2). By applying the modified Lemma B.4 to Theorem 4.2 and Theorem 4.3, following the similar proof, we have new bounds for Theorem 4.2 and Theorem 4.3. That is:
(New Theorem 4.2)

$$\eta[\pi] \geq \eta^{branch}[\pi] - 2r_{max}[\frac{\gamma^{k+1}}{(1-\gamma)^2}(\epsilon_m + 2\epsilon_\pi) + \frac{2\gamma^{k+1} + k + 1}{1-\gamma}\epsilon_\pi].$$

(New Theorem 4.3)

$$\eta[\pi] \geq \eta^{branch}[\pi] - 2r_{max}[\frac{k+1}{1-\gamma}\epsilon_\pi + \frac{\gamma^{k+1}}{(1-\gamma)^2}\epsilon'_m].$$

Moreover, there are some typos and errors in the proof.

- In Theorem A.2, the terms L_2 should be $|\eta[\pi] - \eta^{branch}|$ instead of $|\eta[\pi] - \eta^{\pi_D,\pi}|$.
- The author uses a bounded TV-distance of distributions in the proof. Thus, the condition should instead be "the expected TV-distance between two transition distributions is bounded as $max_t E_{s\sim p_1^t(s)} D_{TV}(p_1(s'|s)||p_2(s'|s)) \leq \delta$" in Lemma B.2
- In Lemma B.2, page 16 line 5 and line 6 should be $\sum_{i=1}^{t}\delta_i$.
- In Lemma B.3, the condition should be "the expected TV-distance between two dynamics distributions is bounded as $max_t E_{s\sim p_1^t(s)} D_{TV}(p_1(s'|s,a)||p_2(s'|s,a)) \leq \epsilon_m$"
- In Lemma B.4, the condition should be "...the dynamics distributions are bounded as $max_t E_{s\sim p_1^t(s)} D_{TV}(p_1^{pre}(s'|s,a)||p_2^{pre}(s'|s,a)) \leq \epsilon_m^{pre}$ and after the branch as $max_t E_{s\sim p_1^t(s)} D_{TV}(p_1^{post}(s'|s,a)||p_2^{post}(s'|s,a)) \leq \epsilon_m^{post}$..."

4 Conclusion

Theorem 4.3 requires the error to be bounded on the distribution of the current policy π_t. The authors state that empirically we can estimate model error on the distribution of the current policy with a linear function of the policy. This estimation lacks rigorous theoretical analysis and proof. Moreover, the authors design the algorithm by intuitive of the branch rollout. I will suggest potential future research directions to address more complicated designing by considering length of branch rollout and other hyper parameters, e.g., approximates of model error on the distribution of the current policy and size of the reply buffer.

References

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/5faf461eff3099671ad63c6f3f094f7f-Paper.pdf`.