Generalizable Policy Improvement via Reinforcement Sampling (Student Abstract)

Rui Kong, Chenyang Wu, Zongzhang Zhang*

National Key Laboratory for Novel Software Technology, Nanjing University, China
School of Artificial Intelligence, Nanjing University, China
{kongr, wucy}@lamda.nju.edu.cn, zzzhang@nju.edu.cn

Abstract

Current policy gradient techniques excel in refining policies over sampled states but falter when generalizing to unseen states. To address this, we introduce Reinforcement Sampling (RS), a novel method leveraging a generalizable action value function to sample improved decisions. RS is able to improve the decision quality whenever the action value estimation is accurate. It works by improving the agent's decision on the fly on the states the agent is visiting. Compared with the historically experienced states in which conventional policy gradient methods improve the policy, the currently visited states are more relevant to the agent. Our method sufficiently exploits the generalizability of the value function on unseen states and sheds new light on the future development of generalizable reinforcement learning.

Introduction

Reinforcement Learning (RL) thrives as a means of decision-making in dynamic, unfamiliar environments. Notably, policy gradient methods iteratively update policy parameters via gradient ascent on sampled states. However, the generalization to unfamiliar states remains wanting. While much research has delved into representation learning for better generalization, the possibility of generalizable policy improvement has not been explored. Our proposal, Reinforcement Sampling (RS), steps in here. Instead of improving the policy on historically experienced states as conventional policy gradient does, it improves on the fly the action distribution predicted by the current policy and samples an action to execute from the improved distribution — thus the name reinforcement sampling. RS makes good use of the generalizability of the value function on unseen states, which contributes to sample-efficient RL.

Method

Reinforcement Sampling (RS)

The RS technique is simply improving the action distribution in some ways and sampling from the improved one. There are multiple choices for policy improvement. Any updated action distribution $\pi'(\cdot \mid s)$ with an expected advantage $\mathbb{E}_{a\sim\pi'(\cdot|s)}[A^\pi(s,a)]$ greater than 0 is guaranteed to improve the policy performance according to the performance difference lemma. Here, the advantage $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ is the difference between the expected value under policy π after executing action a at state s, i.e., $Q^\pi(s,a)$, and the expected value under policy π at state s, i.e., $V^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a)] = \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_t \mid s_0 = s]$, where r_t is the reward at time t and γ is the discount factor. We consider two algorithms that improve policies in different ways, namely Natural Policy Gradient Reinforcement Sampling (NPG-RS) and Sign Policy Gradient Reinforcement Sampling (SignPG-RS).

NPG-RS NPG-RS performs NPG (Kakade 2001) update before sampling. NPG uses the Fisher information matrix of the policy to make policy updates adaptive to the curvature of the policy space. Consider a policy with softmax parameterization which means that the policy is represented by unconstrained parameters $\phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and

$$\pi_\phi(a \mid s) = \frac{\exp(\phi_{s,a})}{\sum_{a'\in\mathcal{A}} \exp(\phi_{s,a'})}. \tag{1}$$

It is shown in (Agarwal et al. 2021) that the NPG update for policy with softmax parameterization is of the form $\phi' = \phi + \frac{\eta}{1-\gamma} A^\pi$ which corresponds to the closed-form update $\pi'(a \mid s) = \pi(a \mid s)\frac{\exp(\eta A^\pi(s,a)/(1-\gamma))}{Z(s)}$, where $Z(s) = \sum_{a\in\mathcal{A}} \pi(a \mid s)\exp(\eta A^\pi(s,a)/(1-\gamma))$. Notice that the improved action distribution can be calculated statewise easily. For policies parameterized by neural networks, we could treat the logits predicted by the policy network as the parameter of a policy with softmax parameterization and improve the policy by applying the NPG update.

SignPG-RS Aside from NPG-RS, we propose SignPG-RS which uses sign policy gradient to update the policy before sampling. For softmax parameterization, the closed-form policy gradient is $\frac{\partial J(\phi)}{\partial \phi_{s,a}} = \frac{1}{1-\gamma} d^\pi(s)\pi(a \mid s)A^\pi(s,a)$, where $J(\phi) = \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t r_t]$ is the expected return following π, and $d^\pi(\cdot)$ is the visitation probability distribution of π. The vanilla policy gradient suffers from problematic dependency on the visitation probability $d^\pi(s)$ of state s under the current policy π and the action selection probability $\pi(a \mid s)$ of action a. The policy is subject to near-zero gradient on

*Corresponding author: Zongzhang Zhang.

states and actions not currently visited and slow improvement. This problem can be resolved by taking the sign of the policy gradient. For policies with softmax parameterization, all reachable states have a visitation probability greater than 0. Therefore, the resulting update is $\phi' = \phi + \eta \text{sign}(A^\pi)$. This update is in a similar form to NPG and is found to be effective in practice. A possible explanation for this is that the sign policy gradient is insensitive to the scale of the advantage, which functions similarly to the well-known advantage normalization trick in PPO(Schulman et al. 2017).

Value Learning

The generalizability of a value function learned by temporal difference learning is undermined by the non-smoothness introduced by the initial non-smooth target value. To tackle this challenge, we propose to use feature regularization.

It is reported in previous work that a lower effective dimension of representation space is closely related to the generalization ability of deep RL (Lan et al. 2022). To reduce the effective dimension of the learned representation, we adopt an auxiliary loss from (Lan et al. 2022) for state representation regularization. Denote the representation network as f_ϕ, where ϕ is the parameter. The feature regularization loss is defined as the smooth maximum of the L2 norm of extracted features on sampled states:

$$\mathcal{L}_\phi = \log \sum_{i=1}^{N} \exp\left(\|f_\phi(s_i)\|_2^2\right), \tag{2}$$

where s_i are states from mini-batch, and N is the batch size.

Policy Learning

The policy is learned simply by cloning the improved policy which is achieved by minimizing the KL divergence:

$$\min_\phi \mathbb{E}_{s\sim d^{\pi'}(\cdot)}[\text{KL}(\pi'(\cdot \mid s)\|\pi_\phi(\cdot \mid s))]. \tag{3}$$

This minimization problem learns a policy π_ϕ that approximates the improved policy π' on the distribution $d^{\pi'}$. Compared to policy gradient methods, our policy learning is a sheer supervised learning task that is more stable to learn and easier to optimize. Besides, the behavior cloning objective makes updates to all actions which can be more efficient compared to policy gradient methods.

Experiments

As an on-policy algorithm, we compare the performance of our proposed method with the fine-tuned PPO algorithm on MinAtar environments. We show the results of SignPG-RS due to its efficiency and defer the results of NPG-RS to the appendix*. MinAtar provides analogous environments to the original ALE games where blocks with different colors are used to represent different objects in the control task. Both methods are trained for 10 million samples for each environment respectively and we report these methods on the metric of overall and last scores. Overall performance is the average undiscounted return during the whole training process and last performance is the average undiscounted return in the last 100 training epochs. These two metrics represent the overall training performance and the final converged performance of testing methods. The results are shown in Table 1, we can observe that SignPG-RS outperforms in all MinAtar environments on both metrics. The huge improvement in both metrics indicates that the generalized policy gradient improves the sample efficiency and converged performance.

Env	Metrics	Methods	
		PPO	RS
Asterix	Overall	7.4 ± 0.2	14.5 ± 0.9
	Last	22.1 ± 2.2	98.1 ± 8.7
Breakout	Overall	12.1 ± 0.2	28.6 ± 1.7
	Last	679.3 ± 15.0	2287.2 ± 473.5
Freeway	Overall	29.6 ± 1.3	51.3 ± 1.2
	Last	48.9 ± 1.2	62.2 ± 0.2
Seaquest	Overall	7.3 ± 1.2	13.6 ± 1.5
	Last	125.9 ± 28.5	372.2 ± 46.9
Space-Invaders	Overall	63.1 ± 2.2	264.2 ± 39.7
	Last	1053 ± 18.0	4601.2 ± 1359.0

Table 1: Results on MinAtar environments. We report the average scores and standard deviation across 5 random seeds.

Conclusion

This paper presents a generalizable policy improvement method named Reinforcement Sampling (RS) which performs policy updates in the sampling stage. To make policy update efficient, we propose a sign policy gradient that is not proportional to the state visitation probability and is insensitive to the advantage scale. To facilitate generalizable value learning, we propose to use feature regularization. On several MinAtar environments, RS is competitive with existing policy gradient methods and shows excellent potential for generalizable policy improvement.

Acknowledgments

This work is supported by the National Science Foundation of China (No. 62276126).

References

Agarwal, A.; Kakade, S. M.; Lee, J. D.; and Mahajan, G. 2021. On the Theory of Policy Gradient Methods: Optimality, Approximation, and Distribution shift. *Journal of Machine Learning Research*, 22(1): 4431–4506.

Kakade, S. M. 2001. A Natural Policy Gradient. In *NIPS*, 1531–1538.

Lan, C. L.; Tu, S.; Oberman, A.; Agarwal, R.; and Bellemare, M. G. 2022. On the Generalization of Representations in Reinforcement Learning. In *AISTATS*, 4132–4157.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

*https://www.lamda.nju.edu.cn/kongr/files/aaai24_rs_appendix.pdf