POLICY GRADIENT WITH TREE SEARCH (PGTS) IN REINFORCEMENT LEARNING EVADES LOCAL MAXIMA

Navdeep Kumar
Technion
navdeepkumar@alum.iisc.ac.in

Priyank Agrawal
Columbia University
pa2608@columbia.edu

Kfir Levy
Technion
kfirylevy@technion.ac.il

Shie Mannor
Technion
shie@ee.technion.ac.il

ABSTRACT

The policy gradient (PG) methods are being extensively used in practice. However, their theoretical convergence guarantees require strict regularity conditions. Such conditions are unnatural and generally not satisfied in practice, causing such techniques to get stuck in a sub-optimal local maximum (rewards). Tree search (TS) methods, have been recently shown to enjoy strong empirical performance in related planning tasks. In this work, we attempt at first theoretical analysis of Tree search-based policy gradient and its convergence properties. Specifically, we show that for a large tree length, the number of local maxima decreases, and therefore in the limiting case, PG converges to a global optimal solution.

1 INTRODUCTION

In reinforcement learning, the agent learns to maximize the return by interacting with the environment. Policy gradient methods in reinforcement learning have been widely studied in multiple variants (Sutton et al., 1999; Schulman et al., 2015b;a; Sutton et al., 2000; Puterman, 2014). Policy gradient (PG) is the first-order gradient of the cumulative return w.r.t. the policy. Hence, at every step, the policy gradient is directed towards the direction of greedy one-step improvement. The return function is highly non-concave in the policy space. Therefore, PG is bound to get stuck in a local maximum. Dynamic programming approach, on the other hand, mitigates this issue (Sutton & Barto, 2018). However, it requires the updating value function at all states at each time, which gets computationally intractable in practice (curse of dimensionality).

Tree Search methods have been studied in the context of value iteration (Efroni et al., 2019) and with PG (Silver et al., 2017a;b). In the context of PG, while the empirical performance of the tree-search methods has been impressive, there are still gaps in the theoretical convergence guarantees. Moreover, Dalal et al. (2023) proposes a soft-tree-max policy, a variant of combining tree search and policy gradient. It established that the soft-tree-max policy has a lower variance than the standard policy gradient. Hence it may lead to better stability.

The convergence of the tree search method with policy gradient has been an open question. This work attempts to theoretically explain the convergence of Policy Gradient with Tree Search (PGTS). We show that the quality of the convergence point improves with the tree search length. That is, as the tree search length increases, the number of local maxima decreases. To the best of our knowledge, this is the first work that establishes this connection.

2 METHOD

A Markov Decision Process is a tuple $(\mathcal{S},\mathcal{A},P,R,\gamma,\mu)$ where the notations are as follows $\mathcal{S},\mathcal{A}$ state and action spaces; $,P\in(\Delta_{\mathcal{S}})^{\mathcal{S}\times\mathcal{A}}$: transition kernel; $R\in\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$: reward function;

$\gamma \in [0, 1)$: discount factor; $\mu \in \Delta_{\mathcal{S}}$: initial distribution; and $\Delta_{\mathcal{X}}$: probability simplex over the set $\mathcal{X}$(Puterman, 2014; Sutton & Barto, 2018). Π is the set of policies $\pi \in \Pi$, where $\pi(a|s)$ denotes the probability of action a at state s. Furthermore, $P(s'|s, a)$ denotes the probability of transition to state s' from the state s under action a, and $P^\pi(s'|s) = \sum_a \pi(a|s)P(s'|s, a), R^\pi(s) = \sum_a \pi(a|s)R(s, a)$ are shorthand. The objective is to find the policy $\pi \in \Pi$, that maximizes the return $\rho^\pi := \mu^T(I - \gamma P^\pi)^{-1}R^\pi$. The first-order (one step gradient) is given by is given by $\frac{\partial \rho^\pi}{\partial \pi(a|s)} = d^\pi(s)Q^\pi(s, a)$, where $d^\pi := \mu^T(I - \gamma P^\pi)^{-1}$ is occupancy measure, $Q^\pi = R + \gamma P v^\pi$ is Q-value function and $v^\pi := (I - \gamma P^\pi)^{-1}R^\pi$ is value function (Sutton et al., 1999). Thus the first-order policy gradient update rule is given by:

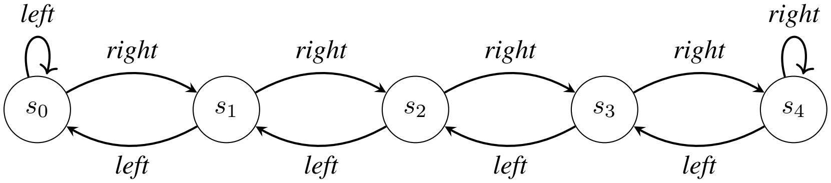
$$\pi_{t+1} = \text{proj}_\Pi \left(\pi_t + \eta \frac{\partial \rho^{\pi_t}}{\partial \pi} \right), \tag{1}$$

where

$$\frac{\partial \rho^\pi}{\partial \pi(a|s)} = d^\pi(s)Q^\pi(s, a).$$

This rule has been proven to converge to a globally optimal solution, given $\min_s \mu(s) > 0$ and for appropriate step size η (Agarwal et al., 2021; Xiao, 2022). The assumption on the initial distribution is difficult to satisfy in practice, especially for large state spaces, unless we are able to visit all states by specifically designed exploration policy. The following simple example illustrates that the above policy gradient update rule may get stuck into a local maxima without the assumption on the initial distribution.

Example 1. *The ladder MDP as illustrated below has reward zero in all states except unit reward at the state s_4. There are two possible actions, 'left' and 'right'.*



Now let the initial state be s_0 (that is $\mu(s_0) = 1$), and the policy π_0 be the initial policy that always plays the action 'left'. It is easy to see that this policy is the local maximum, that is, $\frac{\partial \rho^{\pi_0}}{\partial \pi} = 0$. This implies standard policy gradient will get stuck at π_0, which is a local minimum, while the global optimal policy is always playing the action 'right'.

The reason the PG is zero at π_0 is because the one-step gradient, that is, PG looks only one step ahead in search of better action, and there are better policy is not visible from π_0 by one-step look ahead.

We address this issue by proposing a tree search inspired $m-$step look-ahead gradient:

$$\nabla^m_\pi(s, a) = d^\pi(s) \max_{a_1,\cdots,a_m} \mathbb{E} \left[\sum_{n=0}^{m-1} \gamma^n R(s_n, a_n) + \gamma^m Q^\pi(s_m, a_m) \mid s_0 = s, a_0 = a, P \right]. \tag{2}$$

Let T be optimal Bellman operator defined as $(TQ)(s, a) = R(s, a) + \gamma \sum_{s'} P(s'|s, a) \max_{a'} Q(s', a')$.

Lemma 1. *The above can be written compactly as,*

$$\nabla^m_\pi(s, a) = d^\pi(s)(T^m Q^\pi)(s, a).$$

Theorem 1. *(Limiting case Convergence) For the limiting case $m = \infty$ and $\eta = \infty$ in the update rule $\pi_{t+1} = Proj[\pi_t + \eta \nabla^m_{\pi_t}]$, we have global convergence in S steps, i.e. $\rho^{\pi_S} = \rho^* := \max_\pi \rho^\pi$.*

The result can be easily extended for finite learning rate η. However, its extension to finite tree search length is left for future work.

Example 1 together with Theorem 1 establishes that the number of local maxima decreases to zero as the tree search length increases in PG (since with $m = 0$ as in Example 1 we have multiple maxima, whereas, with $m = \infty$, the number of non-global local maxima is exactly zero)

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

Gal Dalal, Assaf Hallak, Gugan Thoppe, Shie Mannor, and Gal Chechik. Softtreemax: Exponential variance reduction in policy gradient via tree search, 2023.

Yonathan Efroni, Gal Dalal, Bruno Scherrer, and Shie Mannor. How to combine tree-search methods in reinforcement learning, 2019.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International conference on machine learning*, pp. 1889–1897. PMLR, 2015a.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization, 2015b.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017a.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy P. Lillicrap, Fan Hui, L. Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550:354–359, 2017b. URL `https://api.semanticscholar.org/CorpusID:205261034`.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, second edition, 2018. URL `http://incompleteideas.net/book/the-book-2nd.html`.

Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 99, pp. 1057–1063. Citeseer, 1999.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 2000. URL `https://proceedings.neurips.cc/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf`.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

A Appendix

Lemma. *The above can be written compactly as,*

$$\nabla_\pi^m(s,a) = d^\pi(s)(T^m Q^\pi)(s,a).$$

Proof. From definition, we have

$$
\begin{aligned}
\nabla_{\pi}^{m}(s,a) =& d^{\pi}(s) \max_{a_1,\cdots,a_m} \mathbb{E} \Big[\sum_{n=0}^{m-1} \gamma^n R(s_n,a_n) + \gamma^m Q^{\pi}(s_m,a_m) \mid s_0 = s, a_0 = a, P \Big], \\
=& d^{\pi}(s) \max_{a_1,\cdots,a_m} \mathbb{E} \Big[\sum_{n=0}^{m-2} \gamma^n R(s_n,a_n) + \gamma^{m-1} \big[R(s_{m-1},a_{m-1}) + \\
& \qquad \gamma \sum_{s_m} P(s_m|s_{m-1},a_{m-1}) Q^{\pi}(s_m,a_m) \big] \Big] \\
=& d^{\pi}(s) \max_{a_1,\cdots,a_{m-1}} \mathbb{E} \Big[\sum_{n=0}^{m-2} \gamma^n R(s_n,a_n) + \gamma^{m-1} \big[R(s_{m-1},a_{m-1}) + \\
& \qquad \gamma \sum_{s_m} P(s_m|s_{m-1},a_{m-1}) \max_{a_m} Q^{\pi}(s_m,a_m) \big] \Big] \\
=& d^{\pi}(s) \max_{a_1,\cdots,a_{m-1}} \mathbb{E} \Big[\sum_{n=0}^{m-2} \gamma^n R(s_n,a_n) + \gamma^{m-1} \big(TQ^{\pi} \big) (s_{m-1},a_{m-1}) \big] \Big] .
\end{aligned}
$$

Proceeding recursively, we get the desired result. □

Theorem. *(Limiting case Convergence) For the limiting case $d = \infty$ and $\eta = \infty$ in the update rule $\pi_{t+1} = Proj[\pi_t + \eta_t \nabla_{\pi_t}^{m}]$, then $\rho^{\pi_S} = \rho^*$.*

Proof. At $d = \infty$, we have

$$
\begin{aligned}
\nabla_{\theta}^{\infty} &= \sum_{s,a} d^{\pi_\theta}(s) \nabla_\theta \pi(a|s) (T^{\infty} Q^{\pi_\theta})(s,a) && (3) \\
&= \sum_{s,a} d^{\pi_\theta}(s) \nabla_\theta \pi(a|s) Q^{*}(s,a). && (4)
\end{aligned}
$$

We note that if $d^{\pi_t}(s) > 0$ then

$$\pi_{t+1}(\cdot|s) = \pi^*(\cdot|s).$$

Now, let $\mathcal{S}_t = \{s \in \mathcal{S} \mid d^{\pi_t}(s) > 0\}$ and $S_t = |\mathcal{S}_t|$. Then it is easy to see: $S_{t+1} \geq S_t + 1$ or $\mathcal{S}_t = \{s \in \mathcal{S} \mid d^{\pi^*}(s) > 0\}$. If the later is true then, $\pi_{t+1}(\cdot|s) = \pi^*(\cdot|s), \forall s \in \{s \in \mathcal{S} \mid d^{\pi^*}(s) > 0\}$ that implies $\rho^{\pi_{t+1}} = \rho^*$. If the former is true, then this iterates has to terminate at max at S iterates. □