Policy Gradient has No Spurious Local Optima

Tengyang Xie*† Yu Bai *‡ Philip S. Thomas§

Abstract

We study the landscape of policy gradient methods in reinforcement learning. Typically, policy gradient methods are known to have a significant drawback compared with value-based methods: policy gradient methods can become stuck in local optima, while value-based methods like Q-learning are guaranteed to converge to globally optimal policies for problems with finite states and actions. We show that, in the case where value-based methods are guaranteed to converge to global optima, policy gradient methods have no local optima.

1 Introduction

Policy gradient methods and value-based methods are two popular classes of *reinforcement learning* (RL) algorithms. For *Markov decision processes* (MDPs) with finite state and action sets, value-based algorithms like Q-learning and Sarsa converge to globally optimal policies, under additional mild techical assumptions [Jaakkola et al., 1994; Singh et al., 2000]. However, there has been no similar results for the policy gradient methods. We extend and further formalize similar result for policy gradient methods [Thomas, 2014] by showing that the objective function that they optimize has no spurious local optima for MDPs with finite state sets.[1]

2 Main Results

We assume that the reader is familiar with reinforcement learning [Sutton and Barto, 2018] and adopt notational standard MDPNv1 [Thomas and Okal, 2015]. We first consider a tabular representation for the policy wherein the policy, π, is itself a vector in $[0,1]^{|\mathcal{S}\times\mathcal{A}|}$, and

$$\pi_{s,a} \coloneqq \Pr(A_t = a | S_t = s),$$

for all states s, actions a, and times t. For notational simplicity, we assume that the set of possible actions is: $\mathcal{A} = \{1, 2, \dots, |\mathcal{A}|\}$. We write $\pi(s)$ to denote the vector

$$\left[\pi_{s,1}, \pi_{s,2}, \dots, \pi_{s,|\mathcal{A}|}\right]^\intercal,$$

for all $s \in \mathcal{S}$. We write q^π and v^π to denote the discounted action-value and state-value functions associated with policy π, respectively, and further define:

$$q^\pi_s \coloneqq \left[q^\pi(s,1), q^\pi(s,2), \dots, q^\pi(s,|\mathcal{A}|)\right]^\intercal.$$

*Equal contribution.

†College of Information and Computer Sciences, University of Massachusetts Amherst, txie@cs.umass.edu

‡Department of Statistics, Stanford University. yub@stanford.edu

§College of Information and Computer Sciences, University of Massachusetts Amherst, pthomas@cs.umass.edu

[1] Thomas [2014] showed that no local optima exist, but his proof was informal. Furthermore, he did not show convergence to a globally optimal policy, only that no local optima exist (in general, *no* optima exist in policy parameter space).

The objective function optimized by standard policy gradient algorithms is [Sutton et al., 2000]:

$$J(\pi) = \sum_{s\in\mathcal{S}} d^0(s) v^\pi(s) = \sum_{s\in\mathcal{S}} d^0(s) \sum_{a\in\mathcal{A}} \pi_{s,a} q^\pi(s,a), \tag{1}$$

where Π is the set of all possible policies, π. In this paper, we also use $\nabla_x \coloneqq \partial/\partial x$ to denote the partial gradient (not the directional derivative—the typical meaning of this symbol), where x is a vector or scalar. We also use $\langle\cdot\rangle$ to denote the inner product. We now have the following lemma for any suboptimal policies:

Lemma 1. *Let π be a suboptimal policy and π^* be an optimal policy. If $J(\pi^*) - J(\pi) > \varepsilon$, then there must exist at least one state $s \in \mathcal{S}$, such that $\sum_{a\in\mathcal{A}} \pi^*(a|s) q^\pi(s,a) - v^\pi(s) > (1-\gamma)\varepsilon$.*

Proof. We prove Lemma 1 by contradiction. Let $\mathcal{B} \coloneqq \{s \in \mathcal{S} : d^0(s) > 0\}$. Given the definition of J in (1), we have

$$\begin{aligned}\max_{s\in\mathcal{B}} \left(v^{\pi^*}(s) - v^\pi(s)\right) \geq & \sum_{s\in\mathcal{S}} d^0(s) \left(v^{\pi^*}(s) - v^\pi(s)\right)\\ =& J(\pi^*) - J(\pi)\\ >& \varepsilon.\end{aligned}$$

Let s^* be any element of $\arg\max_{s\in\mathcal{B}} \left(v^{\pi^*}(s) - v^\pi(s)\right)$. If for all $s \in \mathcal{S}$,

$$\sum_{a\in\mathcal{A}} \pi^*_{s,a} q^\pi(s,a) - v^\pi(s) \leq (1-\gamma)\varepsilon$$

then

$$v^\pi(s) \geq \sum_{a\in\mathcal{A}} \pi^*_{s,a} q^\pi(s,a) - (1-\gamma)\epsilon, \tag{2}$$

and thus

$$\begin{aligned}
v^\pi(s^*) \overset{\text{(a)}}{\geq} & \sum_{a\in\mathcal{A}} \pi^*_{s^*,a} q^\pi(s^*,a) - (1-\gamma)\varepsilon\\
=& \mathbf{E}\left[R_t + \gamma v^\pi(S_{t+1}) \middle| S_t = s^*, \pi^*\right] - (1-\gamma)\varepsilon\\
\overset{\text{(b)}}{\geq} & \mathbf{E}\left[R_t + \gamma\left(\sum_{a\in\mathcal{A}} \pi^*_{S_{t+1},a} q^\pi(S_{t+1},a) - (1-\gamma)\epsilon\right) \middle| S_t = s^*, \pi^*\right] - (1-\gamma)\varepsilon\\
=& \mathbf{E}\left[R_t + \gamma \sum_{a\in\mathcal{A}} \pi^*_{S_{t+1},a} q^\pi(S_{t+1},a) \middle| S_t = s^*, \pi^*\right] - (1-\gamma)(1+\gamma)\varepsilon\\
\overset{\text{(c)}}{=} & \mathbf{E}\left[R_t + \gamma R_{t+1} + \gamma^2 v^\pi(S_{t+2}) \middle| S_t = s^*, \pi^*\right] - (1-\gamma)(1+\gamma)\varepsilon\\
\geq & \mathbf{E}\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \gamma^3 v^\pi(S_{t+3}) \middle| S_t = s^*, \pi^*\right] - (1-\gamma)\left(1+\gamma+\gamma^2\right)\varepsilon\\
& \vdots\\
\geq & \mathbf{E}\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \gamma^3 R_{t+3} + \cdots \middle| S_t = s^*, \pi^*\right] - (1-\gamma)\left(1+\gamma+\gamma^2+\gamma^3+\cdots\right)\varepsilon\\
=& \mathbf{E}\left[\sum_{i=0}^{\infty} \gamma^i R_{t+i} \middle| S_t = s^*, \pi^*\right] - (1-\gamma)\left(\sum_{i=0}^{\infty}\gamma^i\right)\varepsilon\\
=& v^{\pi^*}(s^*) - \varepsilon,
\end{aligned} \tag{3}$$

where inequalities (a) and (b) follow from (2), equation (c) follows from

$$
\begin{aligned}
&\mathbf{E}\left[\sum_{a\in\mathcal{A}} \pi^*_{S_{t+1},a} q^\pi(S_{t+1},a) \middle| S_t = s^*, \pi^*\right] \\
=&\mathbf{E}\left[\mathbf{E}\left[\sum_{a\in\mathcal{A}} \pi^*_{S_{t+1},a} q^\pi(S_{t+1},a) \middle| S_t = s^*, S_{t+1}, \pi^*\right] \middle| S_t = s^*, \pi^*\right] \\
=&\mathbf{E}\left[\mathbf{E}\left[\sum_{a\in\mathcal{A}} \pi^*_{S_{t+1},a} \left(R_{t+1} + \gamma v^\pi(S_{t+2})\right) \middle| S_t = s^*, S_{t+1}, \pi^*\right] \middle| S_t = s^*, \pi^*\right] \\
=&\mathbf{E}\left[\mathbf{E}\left[R_{t+1} + \gamma v^\pi(S_{t+2}) | S_t = s^*, S_{t+1}, \pi^*\right] | S_t = s^*, \pi^*\right] \\
=&\mathbf{E}\left[R_{t+1} + \gamma v^\pi(S_{t+2}) | S_t = s^*, \pi^*\right], \qquad (4)
\end{aligned}
$$

and the inequalities after (c) follow from (2) and (4). However, (3) contradicts the fact $J(\pi^*) - J(\pi) > \varepsilon$. Thus, we must have $\sum_{a\in\mathcal{A}} \pi^*(a|s^*)q^\pi(s^*,a) - v^\pi(s^*) > (1-\gamma)\varepsilon$. This completes the proof. $\square$

Our main result is as follows:

Theorem 1. *The objective function, J, is only maximized at an optimal policy π^* when using a tabular policy parameterization, and has no other local optima or stationary points.*

Proof. For any given suboptimal policy π, there exists some $\varepsilon > 0$ such that $J(\pi^*) - J(\pi) > \varepsilon > 0$. By Lemma 1, there exists a state $s^* \in \mathcal{S}$ such that $\sum_{a\in\mathcal{A}} \pi^*(a|s^*)q^\pi(s^*,a) - v^\pi(s^*) > (1-\gamma)\varepsilon$. Phil: You still haven't defined $\langle\cdot\rangle$. Different branches of math use it differently. Say explicitly that it is an inner product.

$$
\begin{aligned}
\langle \pi^*(s^*) - \pi(s^*), \nabla_{\pi(s^*)} J(\pi)\rangle =& \langle \pi^*(s^*) - \pi(s^*), d^\pi(s^*) q^\pi_{s^*}\rangle \\
=& d^\pi(s^*) \sum_{a\in\mathcal{A}} \pi^*(a|s^*) q^\pi(s^*,a) - v^\pi(s^*) \\
\geq& d^\pi(s^*)(1-\gamma)\varepsilon > 0.
\end{aligned}
$$

By convexity of the probability simplex, π has to be non-stationary: moving $\pi(s^*)$ towards the direction $\pi^*(s^*) - \pi(s^*)$ (and keeping the policy at other states fixed) will give a first-order increase in the objective value. $\square$

We now consider the case of a softmax policy, $\pi_\theta(a|s) = \exp(\theta_{s,a})/\sum_{a'}\exp(\theta_{s,a'})$. Let still consider the partial gradient $\nabla_{\theta_s} J(\pi_\theta) = [\pi_\theta(1|s)(Q(s,1) - V(s)), \ldots, \pi_\theta(|\mathcal{A}||s)(Q(s,|\mathcal{A}|) - V(s))]^\intercal$ obtained by the policy gradient theorem [Sutton et al., 2000], where $\theta_s = [\theta_{s,1}, \ldots, \theta_{s,|\mathcal{A}|}]^\intercal$. To get this gradient expression, let $M(s) = \frac{\partial \pi_\theta(s)}{\partial \theta_s}$ be the Jacobian on the state s, then we have:

$$
\begin{aligned}
M =& \mathrm{diag}(\pi(s)) - \pi(s)\pi(s)^\intercal \\
M_{ii} =& \pi(i|s)(1 - \pi(i|s)) \\
M_{ij} =& -\pi(i|s)\pi(j|s),
\end{aligned}
$$

where $\pi(s) = [\pi(1|s), \pi(2|s), \ldots, \pi(|\mathcal{A}||s)]^\intercal$. Thus, we can obtain that

$$
\begin{aligned}
&\nabla_{\theta_s} J(\pi_\theta) = M(s) \nabla_{\pi_\theta(s)} J(\pi_\theta) \\
=& (\mathrm{diag}(\pi(s)) - \pi(s)\pi(s)^\intercal) q^\pi_s = [\pi_\theta(1|s)(Q(s,1) - V(s)), \ldots, \pi_\theta(|\mathcal{A}||s)(Q(s,|\mathcal{A}|) - V(s))]^\intercal.
\end{aligned}
$$

holds for all state s. For any $u \in \mathbb{R}^{|\mathcal{A}|}$, we have

$$\langle u, \nabla_{\theta_s} J(\pi_\theta) \rangle = \left\langle (\mathrm{diag}(\pi_\theta(s)) - \pi_\theta(s)\pi_\theta(s)^\intercal)u, \nabla_{\pi_\theta(s)} J(\pi_\theta) \right\rangle .$$

Taking u such that $(\mathrm{diag}(\pi_\theta(s)) - \pi_\theta(s)\pi_\theta(s)^\intercal)u = \pi^*(s) - \pi(s)$ [2] and performing the same argument as Theorem 1, we know that π_θ is not a fixed-point of the gradient ascent update if it is sub-optimal. Therefore there is also no local optima in the parameter space of θ.

References

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710.

Singh, S., Jaakkola, T., Littman, M. L., and Szepesvári, C. (2000). Convergence results for single-step on-policy reinforcement-learning algorithms. *Machine learning*, 38(3):287–308.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063.

Thomas, P. S. (2014). Bias in natural actor-critic algorithms. In *Proceedings of the Thirty-First International Conference on Machine Learning*.

Thomas, P. S. and Okal, B. (2015). A notation for markov decision processes. *arXiv preprint arXiv:1512.09075*.

[2] As $\mathbf{1}^\intercal(\pi^*(s) - \pi(s)) = 0$, there exists such u.