CSCI 670 - Theoretical Thinking Assignment 1

Varun Bhatt
8461808238
vsbhatt@usc.edu

September 24, 2021

In this report, I will be writing about the paper "On the Global Convergence Rates of Softmax Policy Gradient Methods" by Mei et al. [1]. Policy gradient methods are a class of policy search algorithms used to solve Reinforcement Learning (RL) problems. In this paper, the authors consider a specific policy gradient algorithm and give a bound of $O(1/t)$ for its convergence rate. Furthermore, they show that action entropy regularization, a technique commonly used in practice for faster training, has theoretical benefits and improves the convergence rate of the algorithm to $O(e^{-t})$.

1. **How did you find this paper?**

   During my master's degree, I worked on practical RL algorithms. One of the professors in that university worked on the theoretical side of these algorithms. I found this paper while looking at some of his recent works. Since this paper contained an analysis of a popular RL algorithm, it was interesting to read and also a good fit for the assignment.

2. **Give a brief, but clear, discussion of the problem as well as the main result of the paper.**

   RL problems are generally defined as a Markov Decision Process (MDP). The problem consists of an environment with a set of states (S) and an agent with a set of possible actions it can take (A). At each time step, the agent takes an action a that results in receiving a reward $r(s, a)$ and the environment transitioning from state s into a new state s' based on the probability distribution $P(s'|s, a)$. With an initial state distribution ρ and a discount factor γ, the agent's goal is to find a policy $\pi(a|s)$ that maximizes the discounted return

$$J = \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, a_t \sim \pi(\cdot|s_t), \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right]$$

   The simplest RL setting is the tabular setting in which the state and action spaces are finite, and the policy is defined independently for each state-action pair. Most of the theoretical results on RL algorithms, including those in this paper, are restricted to tabular settings. Additionally, a common assumption made for theoretical analysis is that the state distribution induced by a policy has non-zero probabilities for all states.

   Policy gradient methods parametrize the policy using parameters θ and optimize them using the gradient $\nabla_\theta J$. In particular, the algorithm considered in the paper is called softmax policy gradient and uses softmax parametrization $\pi_\theta(\cdot|s) = \text{softmax}(\theta(s, \cdot))$. The algorithm initializes $\theta(s, a)$ for all (s, a) and iteratively updates it as $\theta_{t+1} \leftarrow \theta_t + \eta \frac{\partial J}{\partial \theta_t}$.

   Previous work by Agarwal et al. [2] gave the closed form of the gradient and proved that softmax policy gradient converges to the optimal policy as $t \to \infty$. This paper shows that the rate of convergence is $O(1/t)$, i.e., the sub-optimality of the objective obtained using the current policy reduces at a $O(1/t)$ rate. The authors prove this in two steps. First, they show that the norm of the gradient is always greater than the sub-optimality of the objective if the probability of taking the optimal action is always non-trivial (non-uniform version of Łojasiewicz inequality). Then, using the previously proved convergence guarantee, they show that the probability of taking the optimal action has a non-zero infimum. Consequently, they show that the sub-optimality is upper bounded by $1/t$. They also show that the constant in the bound depends on the specific problem and the initialization.

In the second part of the paper, the authors analyze action entropy regularization, which is commonly used in practice with policy gradient methods. Entropy regularization modifies the objective by adding a penalty to near-deterministic policies. The convergence rate with this regularization was shown to be $O(1/t)$ by Shani et al. [3], which can also be obtained by extending the previous result. The authors improve this bound by first showing that with entropy regularization, the infimum of probabilities of all actions is greater than zero (as opposed to the previous case in which this applied only to the optimal action). Then, they rewrite the Łojasiewicz inequality and show that the gradients are stronger w.r.t sub-optimality than in the previous case. This results in a better convergence rate of $O(e^{-t})$ to the regularized objective.

Since the regularized objective is biased, the authors also provide two possible methods to remove it. The first method removes the regularization after a certain amount of training, while the second method decays the regularization scale. But a convergence rate faster than $O(1/t)$ is not proved for either.

Finally, using a reverse form of Łojasiewicz inequality, the authors prove that the convergence rate is lower bounded by $1/t$ for the vanilla algorithm. Hence, entropy regularization strictly improves the convergence rate.

3. **Briefly discuss the reason that you find this paper interesting and result(s) important.**

   Since I mostly worked on the practical side of RL algorithms, seeing theoretical results for them is always interesting. There are two main results that I think are important. First, the idea of non-uniform Łojasiewicz inequality seems applicable to other RL algorithms too, and hence, might be an important tool in future proofs.

   Second, showing the theoretical advantage of entropy regularization was interesting to me personally. I always considered entropy regularization as a way of helping exploration, but the results of this paper show that it leads to better gradients that speed up training as well. The authors also conjecture that the convergence rate of the algorithm with a decaying regularization scale is strictly faster than that without regularization. If proved correct, it would provide a way to learn optimal policies while still enjoying the advantages of entropy regularization.

4. **State one major open question posted or inspired by the paper. What might be your initial idea to address this problem?**

   An open question posted by the paper is to extend the results to other policy parameterizations. A major result used in the proofs is the non-uniform Łojasiewicz inequality, which directly depends on the form of the policy gradient. Hence, my first idea would be to find a similar inequality for the new parametrization. The second result used in this paper was that the infimum of the probability of taking the optimal action is greater than zero. As long as there is proof of asymptotic convergence using the new parametrization, this result will follow. These two results can then be used to prove the convergence rate, similar to how it is done in this paper.

References

[1] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans, "On the global convergence rates of softmax policy gradient methods," in *Proceedings of the International Conference on Machine Learning*, 2020.

[2] A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan, "Optimality and approximation with policy gradient methods in markov decision processes," in *Proceedings of the Conference on Learning Theory*, 2020.

[3] L. Shani, Y. Efroni, and S. Mannor, "Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.