2023 Ninth Indian Control Conference (ICC)
18 - 20 December, 2023
GITAM University, Visakhapatnam, India

2023 Ninth Indian Control Conference (ICC) | 979-8-3503-9525-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICC61519.2023.10442123

The Reinforce Policy Gradient Algorithm Revisited

Shalabh Bhatnagar

Abstract— **We revisit the Reinforce policy gradient algorithm that works with full cost returns obtained over random length episodes. We propose a new Reinforce type algorithm that estimates the policy gradient using a function measurement over a perturbed parameter using a smoothed functional based gradient estimator. We observe that even though we estimate the gradient of the performance objective using sample performance (and not the sample gradient), the algorithm converges to a neighborhood of a local minimum. We further describe the main convergence result.**

I. Introduction

Policy gradient methods [5] form a popular class of approaches in reinforcement learning where the policy is considered parameterized and the policy parameter is updated along a gradient search direction where the gradient is normally of the value function. The policy gradient theorem [5] is a fundamental result in these approaches and relies on an interchange of the gradient and expectation operators which is a straightforward operation when the state-action space is finite. When this is not so, one would need extra regularity conditions to interchange the two operators much like the previously studied perturbation analysis based sensitivity approaches for optimization via simulation [3].

The Reinforce algorithm [6] is a noisy gradient scheme that updates the policy parameter once after the full return on an episode and is based on the gradient of the performance function. In this paper, we revisit the Reinforce algorithm and present a new algorithm for the case of episodic tasks or the stochastic shortest path setting. In this setting, updates are performed only at instants of visit to a prescribed recurrent state. This algorithm is based on a single function measurement or simulation at a perturbed parameter value where the perturbations are obtained using independent Gaussian random variates.

II. The SF Reinforce Algorithm

We assume here that all stationary policies are proper [1]. Let $C \subset \mathcal{R}^d$ denote a compact and convex projection set and $\Gamma : \mathcal{R}^d \to C$ denote a projection operator that projects any $x \in \mathcal{R}^d$ to its nearest point in C.

Let $\theta(n)$ denote the parameter value obtained after the nth update of this procedure obtained after the $(n-1)$st episode and which is run using the policy parameter $\Gamma(\theta(n)+\delta_n\Delta(n))$, $n \geq 0$, where $\theta(n) = (\theta_1(n), \ldots, \theta_d(n))^T \in \mathcal{R}^d$, $\delta_n > 0$ $\forall n$ with $\delta_n \to 0$ as $n \to \infty$ and $\Delta(n) = (\Delta_1(n), \ldots, \Delta_d(n))^T, n \geq 0$, where $\Delta_i(n), i = 1, \ldots, d, n \geq 0$ are independent random variables distributed according to the $N(0,1)$ distribution.

Let χ^n denote the nth state-action trajectory $\chi^n = \{s_0^n, a_0^n, s_1^n, a_1^n, \ldots, s_{T-1}^n, a_{T-1}^n, s_T^n\}$, $n \geq 0$ where the actions $a_0^n, \ldots, a_{T-1}^n$ in χ^n are obtained using the policy parameter $\theta(n) + \delta_n\Delta(n)$. The instant T denotes the termination instant in the trajectory χ^n when the goal state t is reached. The various actions in χ^n are chosen according to the policy $\phi_{(\theta(n)+\delta_n\Delta(n))}$. The initial state is assumed to be sampled from an initial distribution $\nu = (\nu(i), i \in S)$.

Let $G^n = \sum_{k=0}^{T-1} g_k^n$ be obtained from χ^n, with $g_k^n \equiv g(X_k^n, Z_k^n, X_{k+1}^n)$. The update rule that we consider here is the following: For $n \geq 0, i = 1, \ldots, d$,

$$\theta_i(n+1) = \Gamma_i\left(\theta_i(n) - a(n)\left(\Delta_i(n)\frac{G^n}{\delta_n}\right)\right). \quad (1)$$

The step-sizes $a(n), n \geq 0$ are assumed to satisfy the Robbins-Monro conditions.

Consider now the ODE $\dot{\theta}(t) = \bar{\Gamma}(-\sum_s \nu(s)\nabla V_\theta(s))$, where $\bar{\Gamma} : \mathcal{C}(C) \to \mathcal{C}(\mathcal{R}^d)$ is as in [4] (Chapter 5).

Let $H \overset{\triangle}{=} \{\theta \mid \bar{\Gamma}(-\sum_s \nu(s)\nabla V_\theta(s)) = 0\}$ denote the set of all equilibria of the ODE. By Lemma 11.1 of [2], the only possible ω-limit sets that can occur as invariant sets for the ODE above are subsets of H. Let $\bar{H} \subset H$ be the set of all internally chain recurrent points of this ODE. Our main result below is based on Theorem 5.3.1 of [4] for projected stochastic approximation algorithms and is stated below.

Theorem 1: The iterates $\theta(n), n \geq 0$ governed by (1) converge almost surely to $\bar{H}$.

III. Conclusions

We presented a version of Reinforce that incorporates a one-simulation SF for the episodic task setting and stated a convergence result. In a longer version of this paper, we shall present the analysis and experiments with this algorithm.

References

[1] D. P. Bertsekas. *Dynamic Programming and Optimal Control, Vol.II*. Athena Scientific, 2012.

[2] V. S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint, 2'nd Edition*. Cambridge University Press, 2022.

[3] Y. C. Ho and X. R. Cao. *Perturbation Analysis of Discrete Event Dynamical Systems*. Kluwer, Boston, 1991.

[4] H. J. Kushner and D. S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer Verlag, 1978.

[5] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, volume 99, pages 1057–1063, 1999.

[6] R.J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, pages 5–32, 1992.

The author is with the Department of Computer Science and Automation, Indian Institute of Science, Bangalore. Email: shalabh@iisc.ac.in. He was supported by a J. C. Bose Fellowship, Project No. DFTM/ 02/ 3125/M/04/AIR-04 from DRDO under DIA-RCOE, a project from DST-ICPS, and the RBCCPS, IISc.

979-8-3503-9525-9/23/$31.00 ©2023 IEEE