Improved Sample Complexity Analysis of Natural Policy Gradient Algorithm with General Parameterization for Infinite Horizon Discounted Reward Markov Decision Processes

Washim Uddin Mondal [1] Vaneet Aggarwal [2]

[1]Indian Institute of Technology Kanpur, India [2]Purdue University, USA

Introduction and Background

- The framework of Reinforcement Learning (RL) has a wide array of applications: from epidemic control to transportation to wireless communication.
- An agent aims to learn the best 'policy' by repeatedly interacting with an environment.
- Environment consists of a state that changes following an unknown probability law when the agent executes an action.
- The agent immediately receives a reward value as feedback.
- The goal is to maximize the discounted sum of rewards over an infinite horizon.
- We consider general parameterization where policies are indexed by some d dimensional parameter, θ (e.g., the weights of a neural network). It allows infinite state space.
- The number of state transition samples needed by a learning algorithm to reach within ϵ distance of optimality is known as its sample complexity.
- The number of times it updates the policy parameters is known as its iteration complexity.

Research Gap

Algorithm	Sample Complexity	Iteration Complexity	Hessian-free	IS-free
Vanilla-PG [6]	$\tilde{\mathcal{O}}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	Yes	Yes
STORM-PG-F [1]	$\tilde{\mathcal{O}}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-3})$	Yes	No
SCRN [5]	$\tilde{\mathcal{O}}(\epsilon^{-2.5})$	$\mathcal{O}(\epsilon^{-0.5})$	No	Yes
VR-SCRN [5]	$\mathcal{O}(\epsilon^{-2}\log\left(\frac{1}{\epsilon}\right))$	$\mathcal{O}(\epsilon^{-0.5})$	No	No
NPG [4]	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-1})$	Yes	Yes
SRVR-NPG [4]	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-1})$	Yes	No
SRVR-PG [4]	$\mathcal{O}(\epsilon^{-3})$	$\mathcal{O}(\epsilon^{-2})$	Yes	No
N-PG-IGT [2]	$\tilde{\mathcal{O}}(\epsilon^{-2.5})$	$\mathcal{O}(\epsilon^{-2.5})$	Yes	Yes
HARPG [2]	$\mathcal{O}(\epsilon^{-2}\log\left(\frac{1}{\epsilon}\right))$	$\mathcal{O}(\epsilon^{-2})$	No	Yes

Table 1. Sample and iteration complexities of the existing algorithms for general parameterization.

- As seen from the above table, many existing algorithms use either importance sampling (IS), which requires unreasonable assumptions for the analysis, or second-order (Hessian-related) information, which demands larger memory than first-order algorithms.
- The best known sample complexity is $\mathcal{O}(\epsilon^{-2}\log\left(\frac{1}{\epsilon}\right))$ while the lower bound is $\mathcal{O}(\epsilon^{-2})$.
- Two algorithms achieve the best-known sample complexity: VR-SCRN and HARPG.
- The first one is neither first-order nor IS-free.
- The second one has rather large iteration complexity and is Hessian-based.

Research Question

Does there exist an IS-free and Hessian-free algorithm that either achieves or improves the SOTA $\mathcal{O}(\epsilon^{-2}\log\left(\frac{1}{\epsilon}\right))$ sample complexity?

Our Contributions

- We propose an accelerated natural policy gradient (ANPG) algorithm.
- The proposed algorithm is Hessian-free and IS-free.
- Its sample complexity is $\mathcal{O}(\epsilon^{-2})$ which improves the SOTA by a factor of $\mathcal{O}(\log\left(\frac{1}{\epsilon}\right))$.
- Its iteration complexity is $\mathcal{O}(\epsilon^{-1})$ which beats that of HARPG by a factor of $\mathcal{O}(\epsilon^{-1})$.

Algorithm Design: Key Ideas

The goal of the algorithm is to maximize the value function defined below.

$$J_\rho(\theta) = \mathbf{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\big|s_0\sim\rho,\pi_\theta\right]$$

where the symbols carry their usual meanings. An NPG update is the following.

$$\theta_{k+1} = \theta_k + \eta F_\rho(\theta_k)^{\dagger}\nabla_\theta J_\rho(\theta_k)$$

where η is the learning rate. Note that the update is similar to a PG update except η is modulated by the Moore-Penrose inverse of the Fisher information matrix defined as,

$$F_\rho(\theta) \triangleq \mathbf{E}_{(s,a)\sim\nu_\rho^{\pi_\theta}}\left[\nabla_\theta\log\pi_\theta(a|s)\otimes\nabla_\theta\log\pi_\theta(a|s)\right]$$

where $\nu_\rho^{\pi_\theta}$ is the occupation measure and $\otimes$ is the outer product. One can show that,

$$\omega_k^* \triangleq F_\rho(\theta_k)^{\dagger}\nabla_\theta J_\rho(\theta_k) \in \arg\min_{\omega\in\mathbb{R}^{\mathrm{d}}} L_{\nu_\rho^{\pi_\theta}}(\omega,\theta) \triangleq \frac{1}{2}\mathbf{E}_{(s,a)\sim\nu_\rho^{\pi_\theta}}\left[\frac{1}{1-\gamma}A^{\pi_\theta}(s,a) - \omega^{\mathrm{T}}\nabla_\theta\log\pi_\theta(a|s)\right]^2$$

Thus, the natural gradient ω_k^* can be obtained by iteratively applying gradient descent to $L_{\nu_\rho^{\pi_\theta}}(\cdot,\theta)$. In this paper, we use momentum-based accelerated gradient descent to estimate ω_k^*. Note that,

$$\nabla_\omega L_{\nu_\rho^{\pi_\theta}}(\omega,\theta) = F_\rho(\theta)\omega - \frac{1}{1-\gamma}H_\rho(\theta),\ \text{where } H_\rho(\theta) \triangleq \mathbf{E}_{(s,a)\sim\nu_\rho^{\pi_\theta}}\left[A^{\pi_\theta}(s,a)\nabla_\theta\log\pi_\theta(a|s)\right]$$

Since the transition probability and therefore, $\nu_\rho^{\pi_\theta}$ and $A^{\pi_\theta}(\cdot,\cdot)$ are unknown, we obtain sample-based unbiased estimates of $F_\rho(\theta)$ and $H_\rho(\theta)$ (Algorithm 1 in the paper) which leads to an unbiased estimate $\hat{\nabla}_\omega L_{\nu_\rho^{\pi_\theta}}(\omega,\theta)$.

Psuedo-Code

For $k\in\{0,\cdots,K-1\}$ ▷ Outer Loop

  $\mathbf{x}_0,\mathbf{v}_0 \leftarrow \mathbf{0}$

  For $h\in\{0,\cdots,H-1\}$ ▷ Inner Loop: Accelerated Gradient Descent

$$\mathbf{y}_h \leftarrow \alpha\mathbf{x}_h + (1-\alpha)\mathbf{v}_h \tag{1}$$

$$\mathbf{x}_{h+1} \leftarrow \mathbf{y}_h - \delta\hat{\nabla}_\omega L_{\nu_\rho^{\pi_\theta}}(\omega,\theta_k)\big|_{\omega=\mathbf{y}_h} \tag{2}$$

$$\mathbf{z}_h \leftarrow \beta\mathbf{y}_h + (1-\beta)\mathbf{v}_h \tag{3}$$

$$\mathbf{v}_{h+1} \leftarrow \mathbf{z}_h - \xi\hat{\nabla}_\omega L_{\nu_\rho^{\pi_\theta}}(\omega,\theta_k)\big|_{\omega=\mathbf{y}_h} \tag{4}$$

  $\omega_k \leftarrow \frac{2}{H}\sum_{\frac{H}{2}<h\leq H}\mathbf{x}_h$ ▷ Tail Averaging

$\theta_{k+1} \leftarrow \theta_k + \eta\omega_k$ ▷ Policy Parameter Update

- α,β,δ,ξ are appropriately chosen learning parameters.

Some Important Lemmas: Key Proof Ideas

It can be shown (Corollary 1 in the paper) that the global optimality error can be bounded by the natural gradient estimation error in the inner loop as follows for certain parameter choices.

$$J_\rho^* - \frac{1}{K}\sum_{k=0}^{K-1}\mathbf{E}[J_\rho(\theta_k)] \leq \sqrt{\epsilon_{\text{bias}}} + \frac{G}{K}\sum_{k=0}^{K-1}\mathbf{E}\|(\mathbf{E}\left[\omega_k|\theta_k\right] - \omega_k^*)\|$$
$$+\frac{B}{4L}\left(\frac{\mu_F^2}{G^2}+G^2\right)\left(\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{E}\|\omega_k-\omega_k^*\|^2\right) + \frac{G^2}{\mu_F^2 K}\left(\frac{B}{1-\gamma} + 4L\mathbf{E}_{s\sim d_\rho^{\pi^*}}[KL(\pi^*(\cdot|s)\|\pi_{\theta_0}(\cdot|s))]\right)$$

where B,L,G,μ_F are appropriately defined constants and ϵ_{bias} denotes the expressivity power of the policy parameterization. This result is similar to the result given in [4] except here the first order term is modified to $\mathbf{E}\|\mathbf{E}[\omega_k|\theta_k]-\omega_k^*\|$. Following [3], we can show that,

$$\mathbf{E}\|\omega_k-\omega_k^*\|^2 \leq 22\frac{\sigma^2\mathrm{d}}{\mu_F H} + C\exp\left(-\frac{\mu_F}{20G^2}H\right)\left[\frac{1}{\mu_F(1-\gamma)^4}\right] = \mathcal{O}\left(\frac{1}{H}\right) \tag{5}$$

where H is sufficiently large and the appropriately defined constant σ^2 denotes the (scaled) variance of the gradient estimate $\hat{\nabla}_\omega L_{\nu_\rho^{\pi_\theta}}(\omega_\theta^*,\theta)$, ω_θ^* being the exact minimizer of $L_{\nu_\rho^{\pi_\theta}}(\cdot,\theta)$. To bound the first-order term, observe that if $\bar{\mathbf{x}}_h \triangleq \mathbf{E}\left[\mathbf{x}_h|\theta_k\right]$, $\bar{\mathbf{y}}_h \triangleq \mathbf{E}\left[\mathbf{y}_h|\theta_k\right]$, $\bar{\mathbf{v}}_h \triangleq \mathbf{E}\left[\mathbf{v}_h|\theta_k\right]$, $\bar{\mathbf{z}}_h \triangleq \mathbf{E}\left[\mathbf{z}_h|\theta_k\right]$, $\forall h\in\{0,\cdots,H\}$, then it follows from (1)$-$(4) and the unbiasedness of the gradient estimate that,

$$\bar{\mathbf{x}}_0 = \mathbf{0}, \bar{\mathbf{v}}_0 = \mathbf{0} \tag{6}$$

$$\bar{\mathbf{y}}_h = \alpha\bar{\mathbf{x}}_h + (1-\alpha)\bar{\mathbf{v}}_h \tag{7}$$

$$\bar{\mathbf{x}}_{h+1} = \bar{\mathbf{y}}_h - \delta\nabla_\omega L_{\nu_\rho^{\pi_\theta}}(\omega,\theta_k)\big|_{\omega=\bar{\mathbf{y}}_h} \tag{8}$$

$$\bar{\mathbf{z}}_h = \beta\bar{\mathbf{y}}_h + (1-\beta)\bar{\mathbf{v}}_h \tag{9}$$

$$\bar{\mathbf{v}}_{h+1} = \bar{\mathbf{z}}_h - \xi\nabla_\omega L_{\nu_\rho^{\pi_\theta}}(\omega,\theta_k)\big|_{\omega=\bar{\mathbf{y}}_h} \tag{10}$$

Note that $\mathbf{E}[\omega_k|\theta_k] = \frac{2}{H}\sum_{\frac{H}{2}<h\leq H}\bar{\mathbf{x}}_h$. Therefore, $\mathbf{E}[\omega_k|\theta_k]$ can be thought of as an estimate of ω_k^* when exact gradients $\nabla_\omega L_{\nu_\rho^{\pi_\theta}}(\omega,\theta)$ are available (no noise or deterministic scenario). We have,

$$\mathbf{E}\|(\mathbf{E}[\omega_k|\theta_k]-\omega_k^*)\| \leq \sqrt{C}\exp\left(-\frac{\mu_F}{40G^2}H\right)\left(\frac{1}{\sqrt{\mu_F}(1-\gamma)^2}\right) = \mathcal{O}\left(\frac{1}{H}\right) \tag{11}$$

Using (5) and (11), the global error can be bounded as $\sqrt{\epsilon_{\text{bias}}} + \mathcal{O}\left(\frac{1}{H}+\frac{1}{K}\right)$. To make the second term ϵ, we have to take $H=\mathcal{O}(\epsilon^{-1})$ and $K=\mathcal{O}(\epsilon^{-1})$. This results in $\mathcal{O}(\epsilon^{-2})$ sample complexity and $\mathcal{O}(\epsilon^{-1})$ iteration complexity.

Remark: Note the importance of the first-order term. Without our modification, this term will be $\mathbf{E}\|\omega_k-\omega_k^*\|$ (as in [4]) which would lead to a global optimality error of $\sqrt{\epsilon_{\text{bias}}} + \mathcal{O}\left(\frac{1}{\sqrt{H}}+\frac{1}{K}\right)$ leading to a sample complexity of $\mathcal{O}(\epsilon^{-3})$.

References

[1] Ding et al., On the global optimum convergence of momentum-based policy gradient, AISTATS, 2022.

[2] Fatkhullin et al., Stochastic policy gradient methods: Improved sample complexity for Fisher-non-degenerate policies, ICML, 2023.

[3] Jain et al., Accelerating stochastic gradient descent for least squares regression, COLT, 2018.

[4] Liu et al., An improved analysis of (variance-reduced) policy gradient and natural policy gradient methods, NeurIPS, 2020.

[5] Masiha et al., Stochastic second-order methods improve best-known sample complexity of SGD for gradient-dominated functions, NeurIPS, 2022.

[6] Yuan et al., A general sample complexity analysis of vanilla policy gradient, AISTATS, 2022.

wmondal@iitk.ac.in, vaneet@purdue.edu