Reusing Historical Trajectories in Natural Policy Gradient via Importance Sampling: Convergence and Convergence Rate

Yifan Lin*, **Yuhao Wang***, **Enlu Zhou**
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta
yifan.lin@c3.ai, yuhaowang@gatech.edu,enlu.zhou@isye.gatech.edu

Abstract

We study trajectory reuse in natural policy gradient methods. Classical policy gradient algorithms require large amounts of fresh data, which limits their sample efficiency. We propose RNPG, a reuse-based natural policy gradient algorithm that incorporates past trajectories through importance weighting of both the gradient and the Fisher information matrix estimators. We establish asymptotic convergence and a weak convergence rate for RNPG, showing that reuse improves efficiency without altering the limiting behavior. Experiments on the Cartpole benchmark demonstrate that RNPG achieves faster convergence and smoother performance than vanilla policy gradient (VPG) and vanilla natural policy gradient (VNPG), with additional gains from larger reuse sizes. Our results highlight the theoretical and empirical benefits of reusing trajectories in policy optimization.

1 Introduction

Reinforcement learning (RL) has achieved remarkable success in complex decision-making tasks. However, a key challenge remains: how to efficiently use the large amounts of data required for effective training. Policy gradient (PG) methods are among the most widely adopted algorithms, but their reliance on freshly collected trajectories limits their scalability.

Recent studies have considered importance sampling to reuse past data, though these approaches often neglect the statistical dependence among trajectories and provide little theoretical grounding. This raises a central question: how can we systematically incorporate past trajectories to improve efficiency while maintaining stability?

Contributions. We make the following contributions: (i) introduce a reuse-based PG framework that accounts for trajectory dependence across iterations; (ii) demonstrate improved sample efficiency and convergence properties; and (iii) validate performance through experiments on standard control benchmarks.

2 Methodology

We study an infinite-horizon MDP $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma, \rho_0)$ with discount factor $\gamma \in (0, 1)$ and initial distribution ρ_0. A stochastic policy $\pi_\theta : \mathcal{S} \to \Delta(\mathcal{A})$ is parameterized by $\theta \in \mathbb{R}^d$, and we write

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

$\pi_\theta(a \mid s)$ for its action probability (density). The goal is to find

$$\theta^\star \in \arg\max_{\theta\in\Theta} \ \eta(\theta) := \mathbb{E}_{s_0,a_0,\ldots}\left[\sum_{t=0}^{\infty}\gamma^t \mathcal{R}(s_t,a_t)\right],$$

where $s_0 \sim \rho_0$, $a_t \sim \pi_\theta(\cdot \mid s_t)$, and $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, a_t)$. For a trajectory $\tau = (s_0, a_0, r_0, \ldots, s_H)$, we also use the discounted return $R(\tau) = \sum_{t=0}^{H-1}\gamma^t \mathcal{R}(s_t,a_t)$.

Discounted occupancy measure and advantage. Let $d^{\pi_\theta}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathcal{P}(s_t = s \mid \pi_\theta)$ be the discounted state visitation distribution and $d^{\pi_\theta}(s,a) = d^{\pi_\theta}(s)\pi_\theta(a \mid s)$ the discounted occupancy measure. Then $\eta(\theta) = \mathbb{E}_{(s,a)\sim d^{\pi_\theta}}[\mathcal{R}(s,a)]$. Define the value, action-value, and advantage functions by $V^{\pi_\theta}(s_t) = \mathbb{E}\big[\sum_{l=0}^{\infty}\gamma^l R(s_{t+l},a_{t+l})\big]$, $Q^{\pi_\theta}(s_t,a_t) = \mathbb{E}\big[\sum_{l=0}^{\infty}\gamma^l R(s_{t+l},a_{t+l})\big]$, $A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)$.

2.1 Natural Policy Gradient (NPG)

The policy is updated via

$$\theta_{n+1} = \mathrm{Proj}_\Theta\big(\theta_n + \alpha_n F^{-1}(\theta_n)\nabla\eta(\theta_n)\big), \tag{1}$$

where $F(\theta)$ is the Fisher information matrix (FIM). The policy gradient (e.g., [1]) is

$$\nabla\eta(\theta) = \frac{1}{1-\gamma}\,\mathbb{E}_{(s,a)\sim d^{\pi_\theta}}\big[A^{\pi_\theta}(s,a)\,\nabla_\theta \log\pi_\theta(a \mid s)\big], \tag{2}$$

and the FIM is

$$F(\theta) = \mathbb{E}_{(s,a)\sim d^{\pi_\theta}}\big[\nabla_\theta \log\pi_\theta(a \mid s)\,\nabla_\theta \log\pi_\theta(a \mid s)^\top\big]. \tag{3}$$

In practice, both are estimated from samples $\{\xi_n^i = (s_n^i, a_n^i)\}_{i=1}^B$ drawn i.i.d. from $d^{\pi_{\theta_n}}(s,a)$:

$$\widetilde{\nabla\eta}(\theta_n) = \frac{1}{B}\sum_{i=1}^{B} G(\xi_n^i,\theta_n), \qquad \widetilde{F}(\theta_n) = \frac{1}{B}\sum_{i=1}^{B} S(\xi_n^i,\theta_n), \tag{4}$$

with

$$G(\xi,\theta) = \frac{1}{1-\gamma}\,A^{\pi_\theta}(s,a)\,\nabla_\theta\log\pi_\theta(a \mid s), \qquad S(\xi,\theta) = \nabla_\theta\log\pi_\theta(a \mid s)\,\nabla_\theta\log\pi_\theta(a \mid s)^\top. \tag{5}$$

2.2 RNPG: Reusing Historical Trajectories

To reduce estimator variance when online interaction is limited, we reuse samples from the previous K_1 iterations for the gradient and the previous K_2 iterations for the FIM. Let $\{\xi_m^i\}_{i=1}^B$ be i.i.d. from $d^{\pi_{\theta_m}}$ collected at iteration $m < n$. Define the importance ratio

$$\omega(\xi_m^i, \theta_n \mid \theta_m) = \frac{d^{\pi_{\theta_n}}(\xi_m^i)}{d^{\pi_{\theta_m}}(\xi_m^i)}. \tag{6}$$

The reuse-based estimators are

$$\widehat{\nabla\eta}(\theta_n) = \frac{1}{K_1 B}\sum_{m=n-K_1+1}^{n}\sum_{i=1}^{B}\omega(\xi_m^i,\theta_n \mid \theta_m)\,G(\xi_m^i,\theta_n), \tag{7}$$

$$\widehat{F}(\theta_n) = \epsilon I_d + \frac{1}{K_2 B}\sum_{m=n-K_2+1}^{n}\sum_{i=1}^{B}\omega(\xi_m^i,\theta_n \mid \theta_m)\,S(\xi_m^i,\theta_n), \tag{8}$$

where $\epsilon > 0$ ensures numerical stability and invertibility of $\widehat{F}(\theta_n)$. The RNPG update is then

$$\theta_{n+1} = \mathrm{Proj}_\Theta\Big(\theta_n + \alpha_n\,\widehat{F}^{-1}(\theta_n)\,\widehat{\nabla\eta}(\theta_n)\Big). \tag{9}$$

Practical note on ω. The importance ratio in (6) generally lacks a closed form; in practice, it can be approximated (e.g., via self-normalization or by using per-step likelihood ratios) while retaining the variance-reduction benefits of reuse.

3 Theoretical Results

We present two theoretical guarantees for RNPG. For clarity, both results are stated informally and detailed assumptions are omitted.

Theorem 1 (Asymptotic Convergence) *The iterates $\{\theta_n\}$ generated by Algorithm 9 converge to the limiting point of the ordinary differential equation*

$$\dot{\theta} = \bar{F}^{-1}(\theta)\,\nabla\eta(\theta) + z,$$

where

$$\bar{F}^{-1}(\theta) = \mathbb{E}\left[\left(\epsilon I_d + \tfrac{1}{B}\sum_{i=1}^{B} S(\xi_i,\theta)\right)^{-1}\right],$$

$\xi_1,\ldots,\xi_B$ are i.i.d. samples from the occupancy measure d^{π_θ}, and z denotes the projection force needed to keep the ODE trajectory within the feasible set Θ.

Discussion. If no local minimum lies on the boundary of Θ, then $z = 0$ and $\nabla\eta(\theta_n) \to 0$ almost surely. Importantly, both the reuse of historical samples and the regularization term ϵ preserve convergence, altering only the effective Fisher information matrix $\bar{F}^{-1}(\theta)$ but not the limiting behavior.

Theorem 2 (Weak Convergence and Asymptotic Rate) *Let $\bar{\theta}$ denote a limiting point of the ODE in Theorem 1. As $n \to \infty$,*

$$\frac{\theta_n - \bar{\theta}}{\sqrt{\alpha_n}} \;\Rightarrow\; \mathcal{N}(0,\Sigma_\infty)\,,$$

where $\Rightarrow$ denotes convergence in distribution and

$$\mathrm{vec}(\Sigma_\infty) = -(\mathcal{G}\oplus\mathcal{G})^{-1}\,\mathrm{vec}\big(\widehat{\Sigma}(\bar{\theta})\big), \qquad \widehat{\Sigma}(\bar{\theta}) = \tfrac{1}{B}\Sigma_1(\bar{\theta}) + \tfrac{1}{KB}\Sigma_2(\bar{\theta}).$$

Here $\oplus$ is the Kronecker sum and $\mathrm{vec}(\cdot)$ the vectorization operator.

Discussion. The covariance structure highlights two sources of variation: $\Sigma_1(\bar{\theta})$ captures randomness in the gradient estimator, while $\Sigma_2(\bar{\theta})$ accounts for the joint effect of randomness in both the gradient and the inverse FIM estimator. The asymptotic bound

$$\theta_n - \bar{\theta} = O\left(\alpha_n^{1/2}\sqrt{\tfrac{1}{K} + O(1)}\right)$$

makes explicit that increasing the reuse parameter K reduces the variance of the limiting distribution, thereby improving convergence efficiency.

4 Numerical Experiments

We demonstrate the performance improvement of RNPG over VNPG on the Cartpole benchmark. Unless otherwise noted, the policy is a two-layer fully connected network (32 units per layer, ReLU activations) with a softmax output; optimization uses Adam; discount factor $\gamma = 0.99$; minibatch size $B = 4$ trajectories per iteration; and we add $\epsilon = 10^{-3}$ to the FIM for numerical stability. Reported curves are averaged over 50 macro replications.

4.1 Experiment Setting and Benchmarks

In Cartpole, the goal is to balance a pole by moving a cart left or right. The state is a 4-dimensional vector (cart position/velocity, pole angle/angular velocity); the action space is binary (push left/right with fixed force). Episodes are capped at 200 steps and terminate early if the pole deviates too far from vertical or the cart moves too far from the origin. Reward is 1 per step until termination.

We compare the following algorithms: **VPG**: vanilla policy gradient; **RPG**: policy gradient with reuse; **VNPG**: vanilla natural policy gradient; **RNPG**: natural policy gradient with reuse.

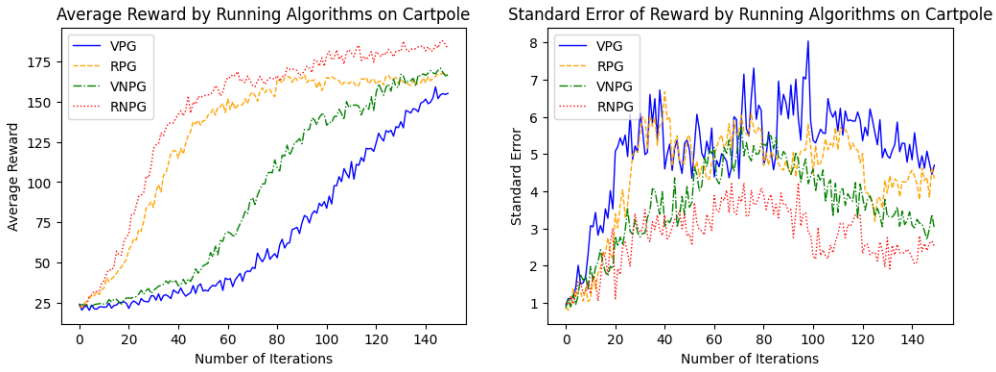

Figure 1: Mean (left) and standard error (right) of reward over $n = 150$ iterations for VPG, RPG, VNPG, and RNPG on Cartpole.

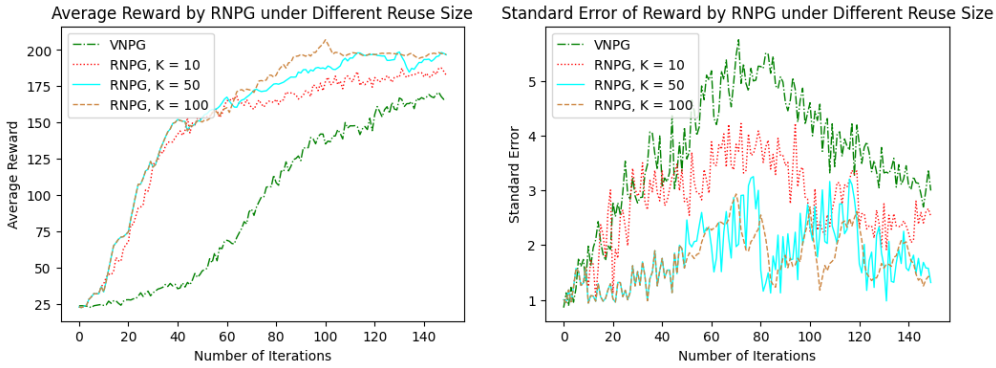

Figure 2: Mean (left) and standard error (right) over $n = 150$ iterations for RNPG under reuse sizes $K \in \{1, 10, 50, 100\}$ on Cartpole.

4.2 Experiment I: Convergence Rate on Cartpole

We run all methods on Cartpole with fixed stepsize $\alpha = 0.01$ and reuse size $K = 10$ (RNPG uses the same K for both the gradient and FIM estimators). Figure 1 shows the mean reward and standard error over 150 iterations. Reusing historical trajectories accelerates convergence for both PG and NPG (RPG vs. VPG; RNPG vs. VNPG), and substantially reduces variability across runs.

4.3 Experiment II: Empirical Study on Reuse Size

We next study the effect of the reuse size K with fixed $\alpha = 0.01$. Figure 2 reports mean reward and standard error for RNPG with $K \in \{1, 10, 50, 100\}$ on Cartpole ($K = 1$ recovers VNPG). Larger K yields faster convergence and smoother trajectories, at the cost of higher computational overhead (dominated by inverting the reuse-weighted FIM).

5 Conclusion

We introduced RNPG, a natural policy gradient method that reuses historical trajectories through principled importance weighting. Our methodology improves sample efficiency by leveraging past experience while preserving the convergence guarantees of classical NPG. Theoretical analysis established asymptotic convergence and a weak convergence rate, showing that reuse reduces estimator variance without altering limiting behavior. Numerical experiments on Cartpole demonstrated that RNPG achieves faster learning and smoother performance than VPG and VNPG, and that larger reuse sizes yield additional efficiency gains. Future directions include applying RNPG to large-scale continuous-control benchmarks, exploring adaptive strategies for choosing reuse sizes, and integrating the approach with actor–critic architectures.

References

[1] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, editors, *Advances in Neural Information Processing Systems*, pages 1057–1063, 1999.