Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

REUSING HISTORICAL OBSERVATIONS IN NATURAL POLICY GRADIENT

Yifan Lin
Enlu Zhou

School of Industrial and Systems Engineering
Georgia Institute of Technology
755 Ferst Drive NW
Atlanta, GA 30332, USA

ABSTRACT

Reinforcement learning provides a mathematical framework for learning-based control, whose success largely depends on the amount of data it can utilize. The efficient utilization of historical samples obtained from previous iterations is essential for expediting policy optimization. Empirical evidence has shown that offline variants of policy gradient methods based on importance sampling work well. However, existing literature often neglect the interdependence between observations from different iterations, and the good empirical performance lacks a rigorous theoretical justification. In this paper, we study an offline variant of the natural policy gradient method with reusing historical observations. We show that the biases of the proposed estimators of Fisher information matrix and gradient are asymptotically negligible, and reusing historical observations reduces the conditional variance of the gradient estimator. The proposed algorithm and convergence analysis could be further applied to popular policy optimization algorithms such as trust region policy optimization. Our theoretical results are verified on classical benchmarks.

1 INTRODUCTION

In challenging reinforcement learning tasks with large state and action spaces, policy optimization methods rank among the most efficacious approaches. It provides a way to directly optimize policies and handles complex and high-dimensional policy representations such as neural networks, all of which contribute to its popularity in the field. It usually works with parametric policies and employs a policy gradient approach to search for the optimal solution (e.g. Sutton et al. 1999). The gradients can be estimated using various techniques, such as the REINFORCE algorithm (Williams 1992) or actor-critic methods (e.g. Konda and Tsitsiklis 1999). These gradient estimation techniques provide a principled way to update the policy parameters based on the observed rewards and state-action trajectories.

The aforementioned on-policy gradient approach involves an iterative approach of gathering experience by interacting with the environment, typically using the currently learned policy. This experience is then utilized to improve the policy. However, in many scenarios, conducting online interactions can be impractical. This can be due to the high cost of data collection (e.g., in robotics or healthcare) or the potential dangers involved (e.g., in autonomous driving). Additionally, even in situations where online interaction is feasible, there may be a preference for utilizing previously collected data to improve the gradient estimation, especially when online data are scarce.

Reusing historical observations to accelerate the learning of the optimal policy is typically achieved by using the importance sampling (IS) technique, which could be traced back to Rubinstein and Shapiro (1990). One significant limitation of this approach in policy optimization is that it can suffer from high variance caused by the importance weights, particularly when the trajectory is long (this is often referred to as trajectory-based or episode-based approach). Liu et al. (2018) propose to apply IS directly on

979-8-3503-6966-3/23/$31.00 ©2023 IEEE

the discounted state visitation distributions to avoid the exploding variance, which is often referred to as step-based approach. Recently, Metelli et al. (2020) propose a policy optimization via importance sampling approach that mixes online and offline optimization to efficiently exploit the information contained in the collected samples. Zheng and Xie (2022) propose a variance reduction based experience replay framework that selectively reuses the most relevant samples to improve policy gradient estimation.

Apart from reusing historical observations via IS to accelerate the convergence of policy gradient algorithm, natural gradient (e.g. Amari 1998) has also been introduced to accelerate the convergence by considering the geometry of the policy parameter space (e.g. Kakade 2001). It is observed that the natural policy gradient algorithm often results in more stable updates that can prevent large policy swings and lead to smoother learning dynamics (Kakade 2001). Another benefit of natural policy gradient is its invariance to the parameterization of the policy, which allows for greater flexibility in designing the policy representation (Amari 1998).

It should be noted that most of the existing works in IS-based policy optimization assume the IS-based gradient estimator is unbiased (e.g. Zheng and Xie 2022), and the convergence analysis is based on the unbiased gradient estimator. However, it is pointed out in Eckman and Henderson (2018), Eckman and Feng (2018) and Liu and Zhou (2020) that the IS-based gradient estimator is biased in the iterative approach due to the dependence between different iterations. Regarding the biased gradient estimator, Liu and Zhou (2020) study the asymptotic convergence of the stochastic gradient descent (SGD) method with reusing historical observations.

In this paper, we propose to use IS in the natural policy gradient algorithm. The IS is used to estimate the gradient as well as the Fisher information matrix (FIM). We extend the convergence analysis of the SGD in the context of simulation optimization (Liu and Zhou 2020) to natural policy gradient in the context of reinforcement learning, and the additional bias caused by reusing historical observations in the FIM complicates the analysis. We theoretically study a mini-batch natural policy gradient with reusing historical observations (RNPG) and show the asymptotic convergence of the proposed algorithm by the ordinary differential equation (ODE) approach. We show that the bias of the natural gradient estimator with historical observations is asymptotically negligible, and RNPG shares the same limit ODE as the vanilla natural policy gradient (VNPG), which only uses samples of the current iteration for FIM and gradient estimators. Moreover, we demonstrate that the proposed RNPG can be applied to other popular policy optimization algorithms such as trust region policy optimization (TRPO, Schulman et al. 2015).

The rest of the paper is organized as follows. Section 2 gives the problem formulation and presents the RNPG algorithm. Section 3 analyzes the convergence behavior of RNPG by the ODE method. Section 4 demonstrates the performance improvement of RNPG over VNPG on a classical benchmark. Section 5 concludes the paper and outlines some future research directions.

2 PROBLEM FORMULATION AND ALGORITHM DESIGN

2.1 Preliminaries

2.1.1 Markov Decision Processes

Consider an infinite-horizon MDP defined as $(\mathscr{S}, \mathscr{A}, \mathscr{P}, \mathscr{R}, \gamma, \rho_0)$, where $\mathscr{S}$ is the state space, $\mathscr{A}$ is the action space, $\mathscr{P}$ is the transition probability with $\mathscr{P}(s_{t+1}|s_t, a_t)$ denoting the probability of transitioning to state s_{t+1} from state s_t when action a_t is taken, $\mathscr{R}$ is the reward function with $\mathscr{R}(s_t, a_t)$ denoting the cost at time stage t when action a_t is taken and state transitions from s_t, $\gamma \in (0,1)$ is the discount factor, ρ_0 is the probability for the initial state, i.e., $s_0 \sim \rho_0$.

Consider a stochastic parameterized policy $\pi_\theta : \mathscr{S} \to \Delta(\mathscr{A})$, defined as a function mapping from the state space to a probability simplex $\Delta(\cdot)$ over the action space, parameterized by $\theta \in \mathbb{R}^d$. For a particular probability (density) from this distribution we write $\pi_\theta(a|s)$. There are a large number of parameterized policy classes. For example, in the case of direct parameterization, the policies are parameterized by $\pi_\theta(a|s) = \theta_{s,a}$, where $\theta \in \Delta(\mathscr{A})^{|\mathscr{S}|}$ is within the probability simplex on the action space. In the case

of softmax parameterization, $\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{a' \in \mathscr{A}} \exp(\theta_{s,a'})}$. The policies can also be parameterized by neural networks, where significant empirical successes have been achieved in many challenging applications, such as playing Go (Silver et al. 2016).

The performance of a policy is evaluated in terms of the expected discounted return $\eta(\pi_\theta) = \mathbb{E}_{s_0,a_0,\ldots}[\sum_{t=0}^{\infty} \gamma^t \mathscr{R}(s_t,a_t)]$, where $s_0 \sim \rho_0(s_0), a_t \sim \pi_\theta(a_t \mid s_t), s_{t+1} \sim \mathscr{P}(s_{t+1} \mid s_t, a_t)$. Denote by $d^{\pi_\theta}(s)$ the discounted state visitation distribution induced by the policy π_θ, $d^{\pi_\theta}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathscr{P}(s_t = s \mid \pi_\theta)$. It is useful to define the discounted occupancy measure as $d^{\pi_\theta}(s,a) = d^{\pi_\theta}(s)\pi_\theta(a|s)$. Using the discounted occupancy measure, we can rewrite the expected discounted return as $\eta(\pi_\theta) = \mathbb{E}_{(s,a)\sim d^{\pi_\theta}(s,a)}[\mathscr{R}(s,a)]$. The goal is for the agent to find the optimal policy π_{θ^*} that maximizes the expected discounted return, or equivalently, $\theta^* = \arg\max_{\theta\in\Theta} \eta(\pi_\theta)$. We use the following standard definitions of the value function V^{π_θ}, the state-action value function Q^{π_θ}, and the advantage function A^{π_θ}: $V^{\pi_\theta}(s_t) = \mathbb{E}_{a_t,s_{t+1},\ldots}\left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})\right]$, $Q^{\pi_\theta}(s_t,a_t) = \mathbb{E}_{s_{t+1},a_{t+1},\ldots}\left[\sum_{l=0}^{\infty} \gamma^l r(s_{t+l})\right]$, and $A^{\pi_\theta}(s,a) = Q^{\pi_\theta}(s,a) - V^{\pi_\theta}(s)$.

2.1.2 Natural Policy Gradient

In the policy gradient algorithm, at each iteration n, we can iteratively update the policy parameters by

$$\theta_{n+1} = \mathrm{Proj}_\Theta\left(\theta_n + \alpha_n \nabla\eta(\theta_n)\right),$$

where α_n is the step size, $\mathrm{Proj}_\Theta(\theta)$ is a projection operator that projects the iterate of θ to the feasible parameter space Θ, and $\nabla\eta(\theta_n)$ is the policy gradient. For ease of notations, we use parameter θ to indicate a parameterized policy π_θ. The gradient is taken with respect to θ unless specified otherwise. The policy gradient (e.g. Sutton et al. 1999) is given by

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

The steepest descent direction of $\eta(\theta)$ in the policy gradient is defined as the vector $d\theta$ that minimizes $\eta(\theta + d\theta)$ under the constraint that the squared length $||d\theta||^2$ is held to a small constant. This squared length is defined with respect to some positive-definite matrix $F(\theta)$ such that $||d\theta||^2 = d\theta^T F(\theta) d\theta$. The steepest descent direction is then given by $F^{-1}(\theta)\nabla\eta(\theta)$ (Amari 1998). It can be seen that the policy gradient descent is a special case where $F(\theta)$ is the identity matrix, and the considered parameter space Θ is Euclidean. The natural policy gradient (NPG) algorithm (Kakade 2001) defines $F(\theta)$ to be the Fisher information matrix (FIM) induced by π_θ, and performs natural gradient descent as follows:

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

where $F(\theta) = \mathbb{E}_{(s,a)\sim d^{\pi_\theta}(s,a)}[\nabla\log\pi_\theta(a|s)\left(\nabla\log\pi_\theta(a|s)\right)^T]$. In practice, both the FIM and policy gradient are estimated by samples. Specifically, at each n-th iteration in stochastic natural policy gradient, we can iteratively update the policy parameters by

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

where $\widetilde{F}(\theta_n)$ and $\widetilde{\nabla\eta}(\theta_n)$ are estimators for FIM and policy gradient, respectively.

2.2 Natural Policy Gradient with Reusing Historical Observations

For ease of notations, we denote by $\xi_n^i = (s_n^i, a_n^i)$ the i-th state-action pair sampled from the discounted occupancy measure $d^{\pi_{\theta_n}}(s,a)$ at iteration n. We assume $\{\xi_n^i, i = 1,\cdots,B\}$ are independent and identically distributed (i.i.d.) samples (observations) from the stationary distribution of the Markov decision process under the policy π_{θ_n}. This i.i.d. assumption does not hold in practice (see e.g. single path generation in

Schulman et al. 2015), but it is widely used in order to show the convergence (see e.g. Zheng and Xie 2022). A vanilla baseline FIM estimator $\widetilde{F}(\theta_n)$ and gradient estimator $\widetilde{\nabla\eta}(\theta_n)$ can be obtained as:

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

where $S(\xi,\theta) = \nabla \log \pi_\theta(a|s)(\nabla \log \pi_\theta(a|s))^T$ and $G(\xi,\theta) = A^{\pi_\theta}(s,a)\nabla \log \pi_\theta(s,a)$. It is easy to see that $\widetilde{F}(\theta_n)$ and $\widetilde{\nabla\eta}(\theta_n)$ are unbiased estimators of the FIM $F(\theta_n)$ and the gradient $\nabla\eta(\theta_n)$, respectively. However, in the vanilla stochastic natural policy gradient (VNPG), a small batch size B, which is often the case when there is limited online interaction with the environment, could lead to a large variance in the estimator. An alternative FIM and gradient estimator, which reuse historical observations, are as follows (see e.g. Liu and Zhou 2020; Zheng and Xie 2022):

$$\widehat{F}(\theta_n) = \frac{1}{KB}\sum_{m=n-K+1}^{n}\sum_{i=1}^{B} \omega(\xi_m^i, \theta_n|\theta_m) S(\xi_m^i, \theta_n), \tag{2}$$

$$\widehat{\nabla\eta}(\theta_n) = \frac{1}{KB}\sum_{m=n-K+1}^{n}\sum_{i=1}^{B} \omega(\xi_m^i, \theta_n|\theta_m) G(\xi_m^i, \theta_n), \tag{3}$$

where we reuse previous $K-1$ iterations' observations, $\omega(\xi_m^i, \theta_n|\theta_m) = \frac{d^{\pi_{\theta_n}}(\xi_m^i)}{d^{\pi_{\theta_m}}(\xi_m^i)}$ is the likelihood ratio. The update of stochastic natural policy gradient with reusing historical observations (RNPG) is then as follows.

$$\theta_{n+1} = \mathrm{Proj}_\Theta\left(\theta_n + \alpha_n \widehat{F}^{-1}(\theta_n)\widehat{\nabla\eta}(\theta_n)\right). \tag{4}$$

We summarize RNPG in Algorithm 1.

Algorithm 1: Natural Gradient Descent with Reusing Historical Observations

1. At iteration $n=0$, choose an initial parameter θ_0. Draw i.i.d. samples $\{\xi_0^i, i=1,\cdots,B\}$ from discounted occupancy measure $d^{\pi_{\theta_0}}(s,a)$ by interacting with the environment.
2. At iteration $n+1$, conduct the following steps.
   2.1 Update θ_{n+1} according to (4).
   2.2 Draw i.i.d. samples $\{\xi_{n+1}^i, i=1,\cdots,B\}$ from discounted occupancy measure $d^{\pi_{\theta_{n+1}}}(s,a)$ by interacting with the environment.
   2.3 $n = n+1$. Repeat the procedure 2.
3. Output θ_n and π_{θ_n} when some stopping criteria are satisfied.

As pointed out by Liu and Zhou (2020) and prior works Eckman and Henderson (2018), and Eckman and Feng (2018), the dependence between iterations makes the FIM and the gradient estimators with historical observations $\widehat{\nabla\eta}(\theta_n)$ biased. This is in contrast to Zheng and Xie (2022), where the authors ignore the bias and their assumption of unbiased gradient estimator cannot be satisfied. It should also be noted that the likelihood ratio in (3) is usually hard to compute, since the discounted occupancy measure does not admit a closed form expression. We defer the discussion on some approximations to make Algorithm 1 more practical to Section 3.

3 CONVERGENCE ANALYSIS

In this section, we first analyze the convergence behavior of RNPG by the ordinary differential equation (ODE) method. We will show that the RNPG and VNPG share the same limit ODE, while the bias resulting from the interdependence between iterations gradually diminishes, ultimately becoming insignificant in the

asymptotic sense. Moreover, we demonstrate that RNPG effectively reduces the conditional variance of each iterate in comparison to VNPG. This observation implies that RNPG offers more stable convergence characteristics and allows for the utilization of larger step sizes. At last we apply the proposed algorithm to some popular policy optimization algorithms such as trust region policy optimization, and propose some approximations to make the proposed algorithm more practical.

3.1 Regularity Conditions for RNPG

We study the asymptotic behavior of Algorithm 1 by the ODE method (please refer to Kushner and Yin (2003) for a detailed exposition on the ODE method for stochastic approximation). The main idea is that stochastic gradient descent (SGD, and in our case is stochastic natural policy gradient, NPG) can be viewed as a noisy discretization of an ODE. Under certain conditions, the noise in NPG averages out asymptotically, such that the NPG iterates converge to the solution trajectory of the ODE. We first summarize the regularity conditions for RNPG that are used throughout the paper.

Assumption 1

- (A.1.1) The step size $\{\alpha_n\}_n$ satisfies $\sum_{n=0}^{\infty}\alpha_n^2<\infty$, $\sum_{n=0}^{\infty}\alpha_n=\infty$, $\lim_{n\to\infty}\alpha_n=0$, $\alpha_n>0, \forall n\geq 0$.
- (A.1.2) The absolute value of the reward $\mathscr{R}(s,a)$ is bounded uniformly, i.e., $\forall(s,a)\in\mathscr{S}\times\mathscr{A}$, there exists a constant $U_r>0$ such that $|\mathscr{R}(s,a)|\leq U_r$.
- (A.1.3) The policy π_θ is differentiable with respect to θ, Lipschitz continuous in θ, and has bounded norm uniformly. That is, there exist constants $L_\Theta, U_\Theta>0$ such that $||\nabla\pi_{\theta_1}(a|s)-\nabla\pi_{\theta_2}(a|s)||\leq L_\Theta||\theta_1-\theta_2||$, $\forall\theta_1,\theta_2\in\Theta$, $||\nabla\pi_\theta(a|s)||\leq U_\Theta$, $\forall(s,a)\in\mathscr{S}\times\mathscr{A}$.
- (A.1.4) $||\pi_{\theta_1}(\cdot|s)-\pi_{\theta_2}(\cdot|s)||_{TV}\leq U_\Pi||\theta_1-\theta_2||, \forall\theta_1,\theta_2\in\Theta, \forall s\in\mathscr{S}$, for some constant $U_\Pi>0$, where $||P-Q||_{TV}$ stands for total variation norm between two probability distributions P and Q with support x, i.e., $||P-Q||_{TV}=\frac{1}{2}\int_x|P(x)-Q(x)|dx$.
- (A.1.5) There exists a constant $\varepsilon_d>0$ such that the discounted occupancy distribution $d^{\pi_\theta}(s,a)\geq\varepsilon_d, \forall(s,a)\in\mathscr{S}\times\mathscr{A}, \forall\theta\in\Theta$.
- (A.1.6) Θ is a nonempty compact set in $\mathbb{R}^d$. Moreover, Θ is convex.

(A.1.1) essentially requires the step size diminishes to zero not too slow ($\sum_{n=0}^{\infty}\alpha_n^2<\infty$) nor too fast ($\sum_{n=0}^{\infty}\alpha_n=\infty$). For example, we can choose $\alpha_n=\frac{\alpha}{n}$ for some $\alpha>0$. (A.1.2) and (A.1.3) are standard assumptions on the regularity of the MDP problem and the parameterized policy. (A.1.4) is from Xu et al. (2020) and holds for any smooth policy with bounded action space. (A.1.5) ensures the discounted occupancy distribution is bounded away from zero. (A.1.6) guarantees the uniqueness of the projection in the solution iterate.

3.2 Asymptotic Convergence by the ODE Method

Before proceeding to our main convergence result, we introduce the continuous-time interpolation of the solution sequence $\{\theta_n\}$. Define $t_0=0$ and $t_n=\sum_{i=0}^{n-1}\alpha_i, n\geq 1$. For $t\geq 1$, let $N(t)$ be the unique n such that $t_n\leq t<t_{n+1}$. For $t<0$, set $N(t)=0$. Define the interpolated continuous process θ^0 as $\theta^0(0)=\theta_0$ and $\theta^0(t)=\theta_{N(t)}$ for any $t>0$, and the shifted process as $\theta^n(s)=\theta^0(s+t_n)$. To work on the projected ODE, we define a set $\mathscr{C}(\theta)$ as follows. When Θ is a hyperrectangle, for $\theta\in\Theta^0$, the interior of Θ, $\mathscr{C}(\theta)$ contains only the zero element; for $\theta\in\alpha\Theta$, the boundary of Θ, let $\mathscr{C}(\theta)$ be the infinite convex cone generated by the outer normals at Θ of the faces on which θ lies. For other more general compact spaces, we refer the readers to Chapter 4.3 in Kushner and Yin (2003) for the construction of set $\mathscr{C}(\theta)$. We then show in the following theorem the limiting behavior of the solution trajectory in Algorithm 1.

Theorem 1 Let $\mathscr{D}^d[0,\infty)$ be the space of $\mathbb{R}^d$-valued operators which are right continuous and have left-hand limits for each dimension. Under Assumption 1, there exists a process $\theta^*(\cdot)$ to which the subsequence of $\{\theta^n(\cdot)\}_n$ converges with probability one (w.p.1) in the space $\mathscr{D}^d[0,\infty)$, where $\theta^*(\cdot)$ satisfies the following

ODE

$$\dot{\theta} = F^{-1}(\theta)\nabla\eta(\theta) + z,\ z \in -\mathscr{C}(\theta), \tag{5}$$

where z is the projection term, i.e., the minimum force needed to keep the trajectory of the ODE $\theta(\cdot)$ from leaving the solution space Θ. The solution trajectory $\{\theta_n\}_n$ in Algorithm 1 also converges w.p.1 to the limit set of the ODE (5).

Before the formal proof of Theorem 1, we first give a high-level proof outline. Note that in the update (4), we can decompose the natural gradient estimation into three components: the true natural gradient, the noise caused by the simulation error, and the bias caused by reusing historical observations. We then separately analyze the noise and bias effects on the estimation of FIM and gradient, and show the noise and bias terms are asymptotically negligible.

For any $s > 0$, let $\boldsymbol{\xi}_s := (\xi_s^1, \cdots, \xi_s^B)$, $\boldsymbol{d}_s := (d_s^{\pi_\theta}(\xi_s^1), \cdots, d_s^{\pi_\theta}(\xi_s^B))$, effective memory $\boldsymbol{e}_s := (\boldsymbol{\xi}_{s-K+1}, \boldsymbol{d}_{s-K+1}, \cdots, \boldsymbol{\xi}_{s-1}, \boldsymbol{d}_{s-1})$, and non-decreasing filtration $\mathscr{F}_n := \sigma\{(\theta_s, \boldsymbol{e}_s), s \le n\}$. With an explicit projection term z_n, we can rewrite (4) as follows

$$\begin{aligned}\theta_{n+1} = \theta_n + \alpha_n \Big(F^{-1}(\theta_n)\nabla\eta(\theta_n) &+ \underbrace{F^{-1}(\theta_n)\widehat{\nabla\eta}(\theta_n) - \mathbb{E}[F^{-1}(\theta_n)\widehat{\nabla\eta}(\theta_n)|\mathscr{F}_n]}_{\delta M_n} \\ &+ \underbrace{\mathbb{E}[F^{-1}(\theta_n)\widehat{\nabla\eta}(\theta_n)|\mathscr{F}_n] - F^{-1}(\theta_n)\nabla\eta(\theta_n)}_{\zeta_n} \\ &+ \underbrace{(\widehat{F}^{-1}(\theta_n) - F^{-1}(\theta_n))\widehat{\nabla\eta}(\theta_n)}_{\iota_n} + z_n \Big),\end{aligned} \tag{6}$$

where δM_n is the noise term caused by the simulation error in the gradient estimator, ζ_n is the bias term caused by reusing historical observations in gradient estimator, and ι_n is due to the inexact estimation of FIM. We can further decompose ι_n as follows

$$\iota_n = \underbrace{(\widehat{F}^{-1}(\theta_n) - \bar{F}^{-1}(\theta_n))\widehat{\nabla\eta}(\theta_n)}_{\delta F_n} + \underbrace{(\bar{F}^{-1}(\theta_n) - F^{-1}(\theta_n))\widehat{\nabla\eta}(\theta_n)}_{D_n}, \tag{7}$$

where $\bar{F}(\theta_n) := \mathbb{E}[\widehat{F}(\theta_n)|\mathscr{F}_n]$. To prevent the FIM from becoming singular, we add a small perturbation εI_d to the FIM to ensure its positive definiteness, where $\varepsilon > 0$ is some small positive number and I_d is a d-by-d identity matrix. We will then show in the rest of the section that the continuous-time interpolations of δM_n, ζ_n and ι_n do not change asymptotically. The formal definition of zero asymptotic rate of change is given below, which is from Chapter 5.3 in Kushner and Yin (2003).

Definition 1 (Zero asymptotic rate of change) A stochastic process $X(t)$ is said to have zero asymptotic rate of change w.p.1 if for some positive number T,

$$\limsup_n \max_{j \ge n} \max_{0 \le t \le T} |X(jT + t) - X(jT)| = 0 \text{ w.p.1.}$$

We first have the following lemma to show the continuous-time interpolations of δM_n and δF_n have zero asymptotic rate of change.

Lemma 2 Let the continuous-time interpolations of δM_n and δF_n be $M(t) = \sum_{i=0}^{N(t)-1} \alpha_i \delta M_i$ and $H(t) = \sum_{i=0}^{N(t)-1} \alpha_i \delta F_i$, respectively. Then $M(t)$ and $H(t)$ have zero asymptotic rate of change w.p.1 under Assumption 1.

We then adopt the fixed-state method and apply Theorem 6.6.1 in Kushner and Yin (2003) to show the convergence of RNPG. Let $P(\boldsymbol{e}_{n+1}|\boldsymbol{e}_n,\theta_n)$ be the transition probability given the current iterate θ_n. Note that $\boldsymbol{e}_n=(\boldsymbol{\xi}_{n-K+1},\boldsymbol{d}_{n-K+1},\cdots,\boldsymbol{\xi}_{n-1},\boldsymbol{d}_{n-1})$, $\boldsymbol{e}_{n+1}=(\boldsymbol{\xi}_{n-K+2},\boldsymbol{d}_{n-K+2},\cdots,\boldsymbol{\xi}_n,\boldsymbol{d}_n)$. Given $\boldsymbol{e}_n$, the component of $\boldsymbol{e}_{n+1}$ that remains unknown are $\boldsymbol{\xi}_n$ and $\boldsymbol{d}_n$, which are random variables that only depend on θ_n. Then $\boldsymbol{e}_n$ has the Markov property: $P(\boldsymbol{e}_{n+1}|\boldsymbol{e}_m,\theta_m,m\le n)=P(\boldsymbol{e}_{n+1}|\boldsymbol{e}_n,\theta_n)$. For a fixed state θ, the transition probability $P(\boldsymbol{e}'|\boldsymbol{e},\theta)$ defines a Markov chain denoted as $\{\boldsymbol{e}_n(\theta)\}$. We expect that the probability law of the chain for a given θ is close to the probability law of the true $\{\boldsymbol{e}_n\}$ if θ_n varies slowly around θ. We are interested in $\{\boldsymbol{e}_i(\theta_n):i\ge n\}$ with initial condition $\boldsymbol{e}_n(\theta_n)=\boldsymbol{e}_n$. Thus, this process starts at value $\boldsymbol{e}_n$ at time n and evolves as if the parameter value were fixed at θ_n forever after, and the limit ODE obtained in terms of this fixed-state chain approximates that of the original iterates.

To explicitly express the estimators' dependence on the history of data $\boldsymbol{e}_b$, let

$$\widehat{\nabla\eta}(\theta,\boldsymbol{e}_m)=\frac{1}{KB}\sum_{j=m-K+1}^{m}\sum_{i=1}^{B}\frac{d^{\pi_\theta}(\xi_j^i)}{d^{\pi_{\theta_j}}(\xi_j^i)}G(\xi_j^i,\theta),\quad \widehat{F}(\theta,\boldsymbol{e}_m)=\frac{1}{KB}\sum_{j=m-K+1}^{m}\sum_{i=1}^{B}\frac{d^{\pi_\theta}(\xi_j^i)}{d^{\pi_{\theta_j}}(\xi_j^i)}S(\xi_j^i,\theta).$$

It is easy to check $\widehat{\nabla\eta}(\theta_n,\boldsymbol{e}_n)=\widehat{\nabla\eta}(\theta_n)$ and $\widehat{F}(\theta_n,\boldsymbol{e}_n)=\widehat{F}(\theta_n)$. Similarly, we can define $\bar{F}(\theta,\boldsymbol{e}_m(\theta))=\mathbb{E}[\widehat{F}(\theta,\boldsymbol{e}_m)|\boldsymbol{e}_m(\theta)=\boldsymbol{e}_m,\theta]$. Define the function $v_n(\theta,\boldsymbol{e}_n)$ and the function $\nu_n(\theta,\boldsymbol{e}_n)$ as follows

$$v_n(\theta,\boldsymbol{e}_n)=\sum_{i=n}^{\infty}\alpha_i F^{-1}(\theta)\mathbb{E}[\widehat{\nabla\eta}(\theta,\boldsymbol{e}_i(\theta))-\nabla\eta(\theta)|\boldsymbol{e}_n(\theta)=\boldsymbol{e}_n,\theta],$$

$$\nu_n(\theta,\boldsymbol{e}_n)=\sum_{i=n}^{\infty}\alpha_i(\bar{F}^{-1}(\theta,\boldsymbol{e}_i(\theta))-F^{-1}(\theta))\widehat{\nabla\eta}(\theta).$$

$v_n(\theta,\boldsymbol{e}_n)$ and $\nu_n(\theta,\boldsymbol{e}_n)$ represent the accumulated bias brought by reusing historical observations in the gradient estimator and FIM estimator, respectively, in the fixed-state chain with fixed state θ. Next we show the bias in the fixed-state chain with fixed state θ_n vanishes.

Lemma 3 Under Assumption 1, $\lim_{n\to\infty}v_n(\theta_n,\boldsymbol{e}_n)=0$ and $\lim_{n\to\infty}\nu_n(\theta_n,\boldsymbol{e}_n)=0$ w.p.1.

We then consider a perturbed iteration $\widetilde{\theta}_n=\theta_n-v_n(\theta_n,\boldsymbol{e}_n)-\nu_n(\theta_n,\boldsymbol{e}_n)$. The use of the perturbation removes $\widehat{F}^{-1}(\theta_n)\widehat{\nabla\eta}(\theta_n)$ and replaces it by $F^{-1}(\theta_n)\nabla\eta(\theta_n)$. For the gradient estimator, an error b_n (due to the replacement of θ_{n+1} by θ_n in $v_{n+1}(\theta_{n+1},\boldsymbol{e}_{n+1})$, and a new martingale difference term δB_n were introduced in the process, and similarly for the FIM estimator. We refer the readers to Chapter 6.6 in Kushner and Yin (2003) for the detailed discussion on the perturbation. Lemma 3 implies the perturbed iteration $\widetilde{\theta}_n$ asymptotically equals to θ_n. We can rewrite the perturbed iteration as follows

$$\widetilde{\theta}_{n+1}=\widetilde{\theta}_n+\alpha_n\big(F^{-1}(\theta_n)\nabla\eta(\theta_n)+\delta M_n+\delta F_n+z_n\big)+b_n+\delta B_n+u_n+\delta U_n,$$

where $b_n=v_{n+1}(\theta_{n+1},\boldsymbol{e}_{n+1})-v_{n+1}(\theta_n,\boldsymbol{e}_{n+1})$, $\delta B_n=v_{n+1}(\theta_n,\boldsymbol{e}_{n+1})-\mathbb{E}[v_{n+1}(\theta_n,\boldsymbol{e}_{n+1})|\boldsymbol{e}_n(\theta)=\boldsymbol{e}_n,\theta_n]$, $u_n=\nu_{n+1}(\theta_{n+1},\boldsymbol{e}_{n+1})-\nu_{n+1}(\theta_n,\boldsymbol{e}_{n+1})$, $\delta U_n=\nu_{n+1}(\theta_n,\boldsymbol{e}_{n+1})-\mathbb{E}[\nu_{n+1}(\theta_n,\boldsymbol{e}_{n+1})|\boldsymbol{e}_n(\theta)=\boldsymbol{e}_n,\theta_n]$. Our next step is to show the continuous-time interpolations of b_n, δB_n, u_n, δU_n have zero asymptotic rate of change.

Lemma 4 Let the continuous-time interpolations of b_n, δB_n, u_n, and δU_n be $B(t)=\sum_{i=0}^{N(t)-1}b_i$, $I(t)=\sum_{i=0}^{N(t)-1}\delta B_i$, $U(t)=\sum_{i=0}^{N(t)-1}u_i$, and $J(t)=\sum_{i=0}^{N(t)-1}\delta U_i$, respectively. Then $B(t)$, $I(t)$, $U(t)$, and $J(t)$ have zero asymptotic rate of change w.p.1 under Assumption 1.

We can then relate the bias term ζ_n in (6) and D_n in (7) to b_n, δB_n and u_n, δU_n, respectively, and show the corresponding continuous-time interpolations have zero asymptotic rate of change in the next corollary.

Corollary 5 Let the continuous-time interpolations of ζ_n and D_n be $Z(t)=\sum_{i=0}^{N(t)-1}\alpha_i\zeta_i$ and $D(t)=\sum_{i=0}^{N(t)-1}\alpha_i D_i$, respectively. Then $Z(t)$ and $D(t)$ have zero asymptotic rate of change w.p.1 under Assumption 1.

We are now ready to show the formal proof of Theorem 1.

Proof. The update (4) in Algorithm 1 can be written as:

$$\theta_{n+1} = \theta_n + \alpha_n\big(F^{-1}(\theta_n)\nabla\eta(\theta_n) + \delta M_n + \zeta_n + \delta F_n + D_n + z_n\big),$$

where δM_n is the noise term caused by the simulation error in the gradient estimator, ζ_n is the bias term caused by reusing historical observations in gradient estimator, δF_n and D_n are due to the inexact estimation of FIM. By Lemma 2, the continuous-time interpolations of δM_n and δF_n have zero asymptotic rate of change. By Corollary 5, the continuous-time interpolations of ζ_n and D_n have zero asymptotic rate of change. Therefore, the limit ODE is determined by the natural gradient $F^{-1}(\theta)\nabla\eta(\theta)$ and the projection. By Theorem 6.6.1 in Kushner and Yin (2003), the solution trajectory $\{\theta_n\}_n$ in Algorithm 1 also converges w.p.1 to the limit set of the ODE (5). □

It should be noted that Theorem 1 only shows the convergence of RNPG to the same limit set of ODE as VNPG. In the next section, we study the benefit of reusing historical observations in terms of reduced conditional variance.

3.3 Reduction of Conditional Variance

We show that reusing historical observations reduces the variance of each iterate conditioned on the history. For simplicity, we only show the reduced conditional variance by reusing historical observations in gradient estimator, while we assume the FIM can be exactly computed. Denote by θ^{RNPG} and θ^{VNPG} the iterates in RNPG and VNPG, respectively. Denote the filtration $\mathscr{F}_n^{\text{VNPG}} = \sigma\{\theta_m^{\text{VNPG}}, m \le n\}$, $\mathscr{F}_n^{\text{RNPG}} = \sigma\{\theta_m^{\text{RNPG}}, \boldsymbol{e}_m, m \le n\}$. Denote by d the dimension of θ. For any vector $V \in \mathbb{R}^d$, denote by $V^{(i)}$ the i-th dimension of V, where $i \le d$. The next theorem shows a sufficient condition on K, the number of reused iterations, for reduction in the conditional variance.

Theorem 6 If K satisfies the following condition

$$K \ge \sqrt{\frac{\max_{\theta\in\Theta}\mathbb{E}_{\xi\sim d^{\pi_\theta}}\left[((F^{-1}(\theta)(G(\xi,\theta)-\nabla\eta(\theta)))^{(i)})^2\right]}{\min_{\theta\in\Theta}\mathbb{E}_{\xi\sim d^{\pi_\theta}}\left[((F^{-1}(\theta)(G(\xi,\theta)-\nabla\eta(\theta)))^{(i)})^2\right]}},$$

then we have $\text{Var}[\theta_{n+1}^{(i),\text{RNPG}}|\mathscr{F}_n^{\text{RNPG}}] \le \text{Var}[\theta_{n+1}^{(i),\text{VNPG}}|\mathscr{F}_n^{\text{VNPG}}]$, w.p.1 $\forall n > 0, i \le d$.

3.4 Extension and Approximation

In this section, we discuss the extension of the proposed RNPG algorithm to trust region policy optimization (TRPO), which is an online natural policy gradient algorithm. With a linear approximation to the objective and quadratic approximation to the constraint, the optimization in each iteration in TRPO can be written as

$$\begin{aligned} \max_{\theta} \quad & \nabla\eta(\theta_n)(\theta-\theta_n) \\ \text{s.t.} \quad & \frac{1}{2}(\theta_n-\theta)^T F(\theta_n)(\theta_n-\theta) \le \delta. \end{aligned}$$

$F(\theta_n)_{ij} = \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}\mathbb{E}_{s\sim d^{\pi_{\theta_n}}(s)}[D_{KL}(\pi_{\theta_n}(\cdot|s)||\pi_\theta(\cdot|s))]|_{\theta=\theta_n}$, where $D_{KL}(P||Q) := \int \log\left(\frac{dP}{dQ}\right)dP$ denotes the Kullback-Leibler divergence from distribution P to distribution Q, and $F(\theta_n)$ is the same FIM as in (1).

Therefore, the update iterate in TRPO can be written as $\theta_{n+1} = \theta_n + \alpha_n F^{-1}(\theta_n)\nabla\eta(\theta_n)$. In practical implementation, TRPO performs a line search in the natural gradient direction, ensuring that the objective is improved while satisfying the nonlinear constraint. We can replace $F(\theta_n)$ and $\nabla\eta(\theta_n)$ by $\widehat{F}(\theta_n)$ and

$\widehat{\nabla\eta}(\theta_n)$ in (4) that reuse the historical observations while still ensuring the convergence of the TRPO algorithm.

Note that in Algorithm 1, we use step-based natural policy gradient algorithm. It requires a single likelihood ratio per state-action pair. However, when computing the likelihood ratio, there is usually no closed-form expression for the discounted state visitation distribution $d^{\pi_\theta}(s)$. To make the algorithm more practical, we could replace the likelihood ratio $\omega(\xi,\theta_n|\theta_m) = \frac{d^{\pi_{\theta_n}}(\xi_m)}{d^{\pi_{\theta_m}}(\xi_m)}$ by $\widehat{\omega}(\xi,\theta_n|\theta_m) = \frac{\pi(\xi_m;\theta_n)}{\pi(\xi_m;\theta_m)}$ (e.g. Degris et al. 2012). Even though it introduces additional bias into the gradient estimator, we can show in the next corollary that the solution trajectory in Algorithm 1 with the likelihood ratio $\widehat{\omega}(\xi,\theta_n|\theta_m)$ converges w.p.1 to the same limit set of the ODE (5).

Corollary 7 Under Assumption 1, the solution trajectory $\{\theta_n\}_n$ in Algorithm 1 with the likelihood ratio $\widehat{\omega}(\xi,\theta_n|\theta_m)$ converges w.p.1 to the limit set of the ODE (5).

4 NUMERICAL EXPERIMENTS

In the numerical experiment, we demonstrate the performance improvement of RNPG over VNPG on CartPole, an OpenAI benchmark problem. The goal is to balance a pole on a cart by moving the cart left or right. The state is a four-dimensional vector representing position of the cart, velocity of the cart, angle of the pole and velocity of the pole. The action space is binary: push the cart left or right with a fixed force. The environment caps episode lengths to 200 steps and ends the episode prematurely if the pole falls too far from the vertical or the cart translates too far from its origin. The agent receives a reward of one for each consecutive step before the termination. The discount factor is $\gamma = 0.99$. For the same considered problem, we compare the performance of the following algorithms. (i) vanilla policy gradient (VPG) and policy gradient with reusing historical observations (RPG); (ii) TRPO and TRPO with reusing historical observations (TRPO-R). Note that the performance difference between VNPG and TRPO has already been shown in Schulman et al. (2015), so we directly build RNPG on top of TRPO.

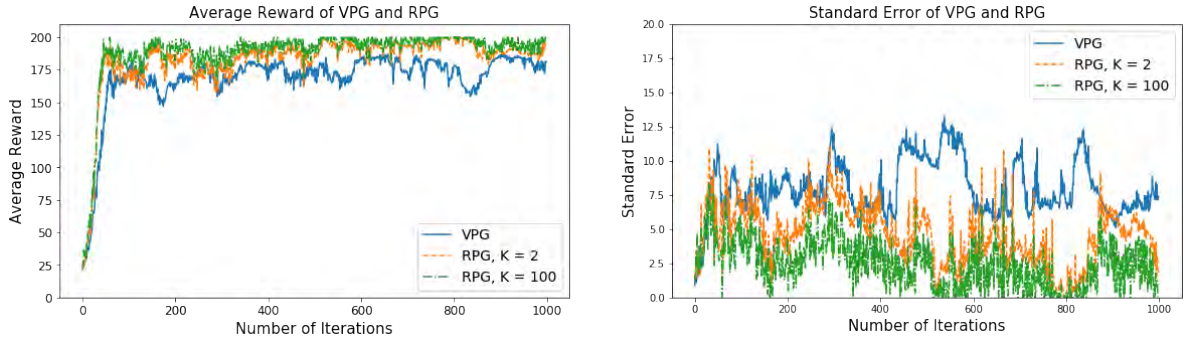

Figure 1: Mean and standard error of the reward over $n = 1000$ iterations for VPG and RPG run on CartPole.

We show the average reward over episodes (i.e., number of iterations) for different algorithms. The reward is averaged over 20 macro replications. The policy network is a fully-connected two-layer neural network with 32 neurons and Rectified Linear Unit (ReLU) activation function. We use softmax activation function on top of the neural network. The policy parameter is updated by Adam optimizer with step size (or learning rate) $\alpha = 0.005$. We should note that similar performance can be obtained by using SGD optimizer with an appropriate decay rate. For policy gradient algorithms (VPG and RPG), the number of observations generated in each iteration (i.e., batch size) is $B = 4$. For natural policy gradient algorithms (TRPO and TRPO-R), the batch size is $B = 64$. Figure 1 shows the mean and standard error of the reward over $n = 1000$ iterations for VPG and RPG algorithms, respectively. Figure 2 shows the mean and standard error of the reward over $n = 1000$ iterations for TRPO and TRPO-R algorithms, respectively. For RPG

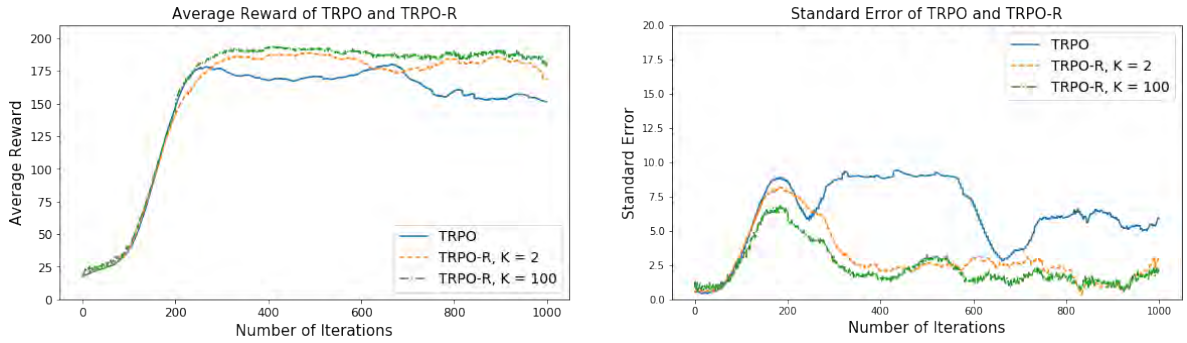

Figure 2: Mean and standard error of the reward over $n = 1000$ iterations for TRPO and TRPO-R run on CartPole.

and TRPO-R, the numbers of reused iterations are $K = 2$ and 100, respectively. We have the following observations from Figure 1 and Figure 2.

(i) Reusing historical observations accelerates the convergence of the policy gradient algorithm (the convergence of RPG is faster than that of VPG) and the natural policy gradient algorithm (the convergence of TRPO-R is faster than that of TRPO).

(ii) Both RPG and TRPO-R have a much smoother trajectory, compared with their vanilla counterpart VPG and TRPO. This can be seen from smaller standard errors of RPG and TRPO-R, compared to VPG and TRPO. It indicates that reusing historical observations reduces the variance of iterates and improves the stability of the algorithm.

(iii) For RPG and TRPO-R, as we reuse more historical observations from previous iterations (larger K), the faster the algorithm converges and the smoother the trajectory is. But this comes with the increased memory for computation.

5 CONCLUSION

In this paper, we study the convergence of an offline variant of natural policy gradient in reinforcement learning with reusing historical observations (RNPG). We show that the biases of the proposed estimators of Fisher information matrix and gradient are asymptotically negligible, and reusing historical observations reduces the conditional variance of the gradient estimator. We further demonstrate that popular policy optimization algorithms, such as trust region policy optimization, could benefit from reusing historical observations with guaranteed convergence. Two potential research directions merit further exploration in the future. First, when showing the reduction in the conditional variance, we assume the exact FIM and only consider the benefit of reusing historical observations in the gradient estimator. Extending the analysis to further consider the benefit of reusing historical observations in FIM will be left for a future work. Second, the ODE method solely examines the mean behavior of RNPG, without providing an explanation for its effectiveness resulting from the reduction in variance by reusing historical observations. The study of the improved convergence rate of RNPG will be left for a future work.

ACKNOWLEDGMENTS

The authors are grateful for the support by Air Force Office of Scientific Research (AFOSR) under Grant FA9550-19-1-0283 and Grant FA9550-22-1-0244, National Science Foundation (NSF) under Grant DMS2053489 and Artificial Intelligence Institute for Advances in Optimization (AI4OPT).

REFERENCES

Amari, S.-I. 1998. "Natural Gradient Works Efficiently in Learning". *Neural Computation* 10(2):251–276.

Degris, T., M. White, and R. S. Sutton. 2012. "Off-Policy Actor-Critic". In *Proceedings of the 29th International Conference on Machine Learning*, edited by J. Langford and J. Pineau, 179–186. Madison, Wisconsin: Omnipress.

Eckman, D. J., and M. B. Feng. 2018. "Green Simulation Optimization Using Likelihood Ratio Estimators". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 2049–2060. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Eckman, D. J., and S. G. Henderson. 2018. "Reusing Search Data in Ranking and Selection: What Could Possibly Go Wrong?". *ACM Transactions on Modeling and Computer Simulation* 28(3):1–15.

Kakade, S. M. 2001. "A Natural Policy Gradient". In *Advances in Neural Information Processing Systems*, edited by T. Dietterich, S. Becker, and Z. Ghahramani, 1531–1538. Cambridge, Massachusetts: MIT Press.

Konda, V., and J. Tsitsiklis. 1999. "Actor-Critic Algorithms". In *Advances in Neural Information Processing Systems*, edited by S. A. Solla, T. K. Leen, and K.-R. Müller, 1008–1014. Cambridge, Massachusetts: MIT Press.

Kushner, H., and G. Yin. 2003. *Stochastic Approximation and Recursive Algorithms and Applications*. New York City, New York: Springer.

Liu, Q., L. Li, Z. Tang, and D. Zhou. 2018. "Breaking the Curse of Horizon: Infinite-Horizon Off-Policy Estimation". In *Advances in Neural Information Processing Systems*, edited by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, 5361–5371. La Jolla, California: Neural Information Processing Systems Foundation, Inc.

Liu, T., and E. Zhou. 2020. "Simulation Optimization by Reusing Past Replications: Don't Be Afraid of Dependence". In *Proceedings of the 2020 Winter Simulation Conference*, edited by K.-H. G. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, 2923–2934. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Metelli, A. M., M. Papini, N. Montali, and M. Restelli. 2020. "Importance Sampling Techniques for Policy Optimization". *The Journal of Machine Learning Research* 21(1):5552–5626.

Rubinstein, R. Y., and A. Shapiro. 1990. "Optimization of Static Simulation Models by the Score Function Method". *Mathematics and Computers in Simulation* 32(4):373–392.

Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz. 2015. "Trust Region Policy Optimization". In *Proceedings of the 32nd International Conference on Machine Learning*, edited by F. Bach and D. Blei, 1889–1897. Cambridge, Massachusetts: The Journal of Machine Learning Research.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. 2016. "Mastering the Game of Go with Deep Neural Networks and Tree Search". *Nature* 529(7587):484–489.

Sutton, R. S., D. McAllester, S. Singh, and Y. Mansour. 1999. "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In *Advances in Neural Information Processing Systems*, edited by S. A. Solla, T. K. Leen, and K.-R. Müller, 1057–1063. Cambridge, Massachusetts: MIT Press.

Williams, R. J. 1992. "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". *Machine Learning* 8:229–256.

Xu, T., Z. Wang, and Y. Liang. 2020. "Improving Sample Complexity Bounds for (Natural) Actor-Critic Algorithms". In *Advances in Neural Information Processing Systems*, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, 4358–4369. La Jolla, California: Neural Information Processing Systems Foundation, Inc.

Zheng, H., and W. Xie. 2022. "Green Simulation Based Policy Optimization with Partial Historical Trajectory Reuse". In *Proceedings of the 2022 Winter Simulation Conference*, edited by B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. Corlu, L. Lee, and P. Lendermann, 168–179. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

AUTHOR BIOGRAPHIES

YIFAN LIN is a Ph.D. student in the School of Industrial and Systems Engineering at Georgia Institute of Technology. His research interests include reinforcement learning and simulation optimization. His email address is ylin429@gatech.edu.

ENLU ZHOU is a Professor in the School of Industrial and Systems Engineering at Georgia Institute of Technology. She received the B.S. degree from Zhejiang University, China, in 2004, and received the Ph.D. degree in electrical engineering from the University of Maryland, College Park, in 2009. Her research interests include control, simulation, and stochastic optimization. Her email address is enlu.zhou@isye.gatech.edu, and her web page is http://enluzhou.gatech.edu/.