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

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

ABSTRACT

Reinforcement learning provides a 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.

1 PROBLEM FORMULATION AND ALGORITHM DESIGN

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 $\pi_\theta(a|s)$. 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 $A^{\pi_\theta}(s,a)=Q^{\pi_\theta}(s,a)-V^{\pi_\theta}(s)$.

The natural policy gradient (NPG) algorithm defines $F(\theta)$ to be the Fisher information matrix (FIM) induced by π_θ, and performs natural gradient descent as follows: $\theta_{n+1}=\mathrm{Proj}_\Theta\left(\theta_n+\alpha_n F^{-1}(\theta_n)\nabla\eta(\theta_n)\right)$, where $F(\theta)=\mathbb{E}_{(s,a)\sim d^{\pi_\theta}(s,a)}[\nabla\log\pi_\theta(a|s)(\nabla\log\pi_\theta(a|s))^T]$.

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}. A vanilla unbiased FIM and gradient estimator can be obtained by the sample average of a batch of data. 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 the large variance in the estimator. An alternative FIM and gradient estimator, which reuse historical observations, are as follows:

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

where we reuse previous $K-1$ iterations' historical 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, and we summarize RNPG in Algorithm 1.

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

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 (1).
   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 when some stopping criteria are satisfied.

2 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.

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 some mild conditions, there exists a process $\theta^*(\cdot)$ to which the subsequence of $\{\theta^n(\cdot)\}_n$ converges w.p.1 in the space $\mathscr{D}^d[0,\infty)$, where $\theta^*(\cdot)$ satisfies the following ODE

$$\dot{\theta} = F^{-1}(\theta)\nabla\eta(\theta) + z,\ z \in -\mathscr{C}(\theta), \tag{2}$$

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 (2).

Note that in the update (1), 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.