ABSTRACT

For reinforcement learning on complex stochastic systems where many factors dynamically impact the output trajectories, it is desirable to effectively leverage the information from historical samples collected in previous iterations to accelerate policy optimization. Classical experience replay allows agents to remember by reusing historical observations. However, the uniform reuse strategy that treats all observations equally overlooks the relative importance of different samples. To overcome this limitation, we propose a general variance reduction based experience replay (VRER) framework that can selectively reuse the most relevant samples to improve policy gradient estimation. This selective mechanism can adaptively put more weight on past samples that are more likely to be generated by the current target distribution. Our theoretical and empirical studies show that the proposed VRER can accelerate the learning of optimal policy and enhance the performance of state-of-the-art policy optimization approaches.

Keywords Reinforcement Learning · Policy Optimization · Multiple Important Sampling · Variance Reduction · Experience Replay · Markov Decision Process

1 Introduction

In recent years, various policy optimization approaches are developed to solve challenging control problems in healthcare [Yu et al., 2021, Zheng et al., 2021], continuous control tasks [Lillicrap et al., 2016, Schulman et al., 2015a, 2017], and biomanufacturing [Zheng et al., 2022]. These approaches often consider parametric policies and search for optimal solution through policy gradient approach [Sutton and Barto, 2018], whose performance and convergence crucially depends on the accuracy of gradient estimation. Reusing historical observations is one way to improve gradient estimation, especially when historical data is scarce. In this paper, we address an important question in policy optimization methods: How to intelligently select and reuse historical samples to accelerate the learning of the optimal policy for complex stochastic systems?

According to the different basic unit of historical samples to reuse, policy gradient (PG) algorithms can be classified into episode-based and step-based approaches [Metelli et al., 2020]. Episode-based approaches can reuse historical trajectories through importance sampling (IS) strategy accounting for the distributional difference induced by the target and behavioral policies. In this case, the importance sampling weight is built on the product of likelihood ratios (LR) of state-action transitions occurring within each process trajectory. As a result, the likelihood-ratio-weighted observations can have large or even infinite variance, especially for problems with long planning horizons [Schlegel et al., 2019]. On the other hand, the step-based approaches take individual state-action transitions as the basic reuse units.
overcomes the limitation of episode-based approaches, provide a more flexible reuse strategy, and support process online control. In this paper, we will develop a general experience replay framework applicable to both episode- and step-based approaches.

For complex stochastic systems such as healthcare [Hall et al., 2012] and biopharmaceutical manufacturing [Zheng et al., 2022], each real or simulation experiment can be financially or computationally expensive. Therefore, it is important to fully utilize all available information when we solve reinforcement learning (RL) optimization problems to guide real-time decision making. On-policy methods only utilize newly generated samples to estimate the policy gradient for each policy update. Ignoring the relevant information carried with historical samples can lead to high uncertainty in policy gradient estimation. Fortunately, this information loss can be reduced through a combination of experience replay (ER) [Lin, 1992, Mnih et al., 2015, Wang et al., 2017] and off-policy optimization methods, which can store and “replay” past relevant experiences to accelerate the search for optimal policy. It often uses importance sampling (IS) mechanism to account for the likelihood mismatch induced by each target policy that we want to assess and the behavioral policy under which past samples were collected [Hesterberg, 1988].

Motivated by the multiple importance sampling (MIS) study presented in Dong et al. [2018], we create a mixture likelihood ratio metamodel that can improve the policy gradient estimation through reusing historical samples. It can compensate the mismatch between target and behavioral distributions so that each historical sample can be used to provide an unbiased prediction of the mean response across the input policy space. However, the reuse of historical samples from behavioral or proposal distributions that are far from the target distribution can lead to the inflated variance on policy gradient estimation.

To avoid this issue, we propose a general variance reduction based experience replay (VRER) framework that can selectively reuse historical samples to reduce the estimation variance of policy gradient and speed up the search for optimal policy. As policy optimization proceeds, this approach can automatically select and reuse the most relevant historical samples generated under visited behavioral policies. The computations required in the proposed selection procedure increase with the number of iterations and the size of historical samples. To circumvent this challenge, we further introduce an approximate selection strategy that dramatically reduces the computational cost. As such, our proposal is computationally efficient and it can support real-time decision making for complex stochastic systems.

Here, we summarize the key contributions and benefits of the proposed framework.

- We propose a general VRER framework for policy gradient optimization. It can accelerate the learning of optimal policy for both step-based and episode-based policy optimization algorithms. This framework is mainly inspired by mixture likelihood ratio (MLR), stochastic gradient variance reduction, and experience replay.
- Since each experiment can be very expensive for complex stochastic systems, the proposed VRER can intelligently select and reuse the most relevant historical samples to improve sample efficiency and estimation accuracy of policy gradient.
- Practically, the proposed VRER can be easily applied to most policy gradient optimization methods without structural change. It mainly requires adding the proposed selection procedure before the training step.
- Comprehensive theoretical and empirical studies demonstrate that the proposed VRER framework can efficiently utilize past samples and accelerate the learning of optimal policy for complex stochastic systems.

The organization of this paper is as follows. We review the most related literature studies in Section 2 and present the problem description of general policy gradient optimization for both finite and infinite horizon Markov decision processes (MDPs) in Section 3. We propose the (multiple) importance sampling based policy gradient estimators for both step- and episode-based algorithms in Section 4. Then, we develop the computationally efficient selection rules and propose generic VRER based policy gradient optimization in Section 5. At the end of Section 5, we also provide a finite-time convergence analysis of VRER based policy optimization and show the asymptotic properties. We conclude this paper with the comprehensive empirical study on the proposed framework in Section 6. The implementation of VRER can be found at [https://github.com/zhenghuazx/vrer_policy_gradient](https://github.com/zhenghuazx/vrer_policy_gradient).

2 Literature Review

The goal of RL is to learn the optimal policy through dynamic interactions with the systems of interest to achieve the best reward [Sutton and Barto, 2018]. Stochastic gradient descent or ascent approaches are often used to solve RL problems [Sutton et al., 1999a]. The study of policy optimization can be traced back to REINFORCE, also known as vanilla policy gradient (VPG) [Williams, 1992]. Later on, Konda and Tsitsiklis [1999], Sutton et al. [1999b] introduced a functional approximation of value function to policy gradient optimization, which is also referred as actor-critic method.
Many approaches have been proposed to improve its performance in terms of better sample efficiency, scalability, and rate of convergence in the recent years. For example, Soft Actor-Critic (SAC) incorporates the entropy measure of policy into the reward objective to encourage exploration [Haarnoja et al., 2018]. To improve training or prediction reliability, the trust region policy optimization (TRPO) enforces a KL divergence constraint on the size of policy update at each iteration to avoid too much policy parameter change in one update [Schulman et al., 2015a].

For both on-policy and off-policy RL, effective policy optimization algorithms should account for the variance of policy gradient estimation [Metelli et al., 2020]. An early attempt from Williams [1992] subtracts a baseline (some value does not depend on the action) from the gradient estimator to reduce its variance without introducing bias. Recent modern policy optimization methods typically use the value function as baseline to reduce the variance of the critic estimator; see for example [Bhatnagar et al., 2009].

Important sampling (IS) has been widely applied to improve policy optimization. Traditionally, it was often used for off-policy value evaluation and policy correction through likelihood ratio between the target policy and the behavior policy [Jiang and Li, 2016, Precup, 2000, Degris et al., 2012]. However, IS estimators can suffer from large variance, which motivates the development of variance reduction techniques [Mahmood and Sutton, 2015, Wu et al., 2018, Lyu et al., 2020]. For example, Espeholt et al. [2018] presents an IS weighted actor-critic architecture that adopts “retrace” [Munos et al., 2016], a method truncating IS weight with a upper bound to resolve the high variance issue. Metelli et al. [2020] propose the Policy Optimization via Importance Sampling (POIS) approach which optimizes a surrogate objective function accounting for the trade–off between the estimated performance improvement and variance inflation induced by IS. IS is also studied in the simulation literature; for example Feng and Staum [2017], Dong et al. [2018] propose green simulation approach that reuses outputs from past experiments to create a metamodel predicting system mean response and support experiment design.

Experience replay and its extension, such as prioritized experience replay [Schaul et al., 2016], are often used in the policy gradient and RL algorithms to reduce data correlation and improve sample efficiency. Wang et al. [2017] propose actor-critic with experience replay (ACER) that applies experience replay method to an off-policy actor-critic algorithm. This study uses the IS weight to account for the discrepancy between the proposal and target policy distributions and truncates the importance weight to avoid inflated variance.

An important perspective of policy gradient methods is to prevent dramatic updates in policy parametric space. Driven by this principle, trust region policy optimization (TRPO) [Schulman et al., 2015a] considers a surrogate objective function subject to the trust region constraint which enforces the distance between old and new candidate policies measured by KL-divergence to be small enough. Following the similar idea, proximal policy optimization (PPO) [Schulman et al., 2017] truncates the importance weights to discourage the excessively large policy update. As Metelli et al. [2020] pointed out, although TRPO and PPO represent the state–of–the–art policy optimization approaches in RL that successfully control the updates in the policy parameter space, they fail to account for the uncertainty induced by the IS procedure. Unlike these existing approaches, our proposed reuse selection rule can address both issues via comparing the difference between policy gradient estimation variance between new (target) and old (behavioral) policies. First, the variance based selection rule can avoid large variance induced by reusing out-dated samples. Second, such selection also controls the parameter update with a constraint on the distance between old and new policies, because the large mismatch between two policies leads to large likelihood ratio, which in turn causes large gradient variance.

3 Problem Description

In this section, we first review both finite- and infinite-horizon Markov decision process (MDPs) in Section 3.1. Each experiment on complex stochastic systems can be computationally burdensome and financially expensive. Thus, given a tight budget, it is important to leverage on all relevant information from current and historical experimental samples to accelerate the optimization search, especially for systems with high uncertainty. The proposed variance reduction based experience replay is applicable to general policy gradient approaches as reviewed in Section 3.2. Then, we summarize the regularity assumptions and conditions for policy gradient optimizations to support the asymptotic study of the proposed framework in Section 3.3.

3.1 Markov Decision Process

We formulate the RL problems of interest as a MDP specified by \((S, A, H, r, P, s_1)\), where \(S\), \(A\), and \(H\) are the state space, the action space, and the planning horizon, respectively. For the infinite-horizon problems, we let \(H = \infty\), and \(H\) is a constant otherwise. At each time \(t \in [H]\), the agent observes the state \(s_t \in S\), takes an action \(a_t \in A\), and receives a reward \(r_t(s_t, a_t) \in \mathbb{R}\). Here \([\cdot]\) represents the sequence of integers from 1 to \(H\). In this study, we consider stochastic policy \(\pi_\theta : S \rightarrow A\), defined as a mapping from state space to the action and parameterized by \(\theta \in \mathbb{R}^d\), i.e.,
where \( a_t \sim \pi_\theta(a|s_t) \). The probability model \( \mathbb{P} \) of the stochastic decision process (SDP) of interest is specified by the state transition probabilities \( p(s_{t+1}|s_t, a_t) \) for all \( t \), as well as the probability \( p(s_1) \) for the initial state, i.e., \( s_1 \sim p(s_1) \).

At any time \( t \), the agent observes the state \( s_t \in S \), takes an action \( a_t \in A \) by following a parametric policy distribution, \( \pi(a_t|s_t; \theta) \), and receives a reward \( r_t(s_t, a_t) \in \mathbb{R} \). The performance of the candidate policy is evaluated in terms of the expected return, i.e., the expected discounted sum of the rewards collected along the trajectory \( R(\tau) \equiv \sum_{t=1}^{H+1} \gamma^t r(s_t, a_t) \), where \( \gamma \in (0, 1) \) denotes the discount factor. Let \( R_t(H) = \sum_{t'=t}^{H+1} \gamma^{t'-t} r(s_{t'}, a_{t'}) \) denote the total discounted reward from time-step \( t \) onwards.

### (1) Finite Horizon MDPs.

In this case, we can have a finite-length process trajectory, denoted by \( \tau \equiv (s_1, a_1, s_2, a_2, \ldots, s_H, a_H, s_{H+1}) \), with probability density function

\[
D(\tau; \theta) \equiv p(s_1) \prod_{t=1}^{H} \pi_\theta(a_t|s_t)p(s_{t+1}|s_t, a_t).
\]  

For the policy \( \pi_\theta \) specified by \( \theta \), we have the state-value and action-value functions

\[
V^\pi_\theta(s_t) = \mathbb{E}[R_t^1(H)|s_t; \pi_\theta] \quad \text{and} \quad Q^\pi_\theta(s_t, a_t) = \mathbb{E}[R_t^1(H)|s_t, a_t; \pi_\theta]
\]

of the expected total discounted reward-to-go. Our goal is to find the optimal policy, denoted by \( \pi^*_\theta \), that maximizes the expected return,

\[
\theta^* = \arg\max_{\theta \in \Theta} \mu(\theta) = \mathbb{E}_\tau[R(\tau)|\theta]
\]

where \( \Theta \) represents the policy parameter space.

### (2) Infinite Horizon MDPs.

For infinite horizon case, we can further write the objective for the infinite-horizon discounted MDP as

\[
\max_{\theta \in \Theta} \mu(\theta) = \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \right| \pi_\theta]
\]

\[
= \int d^\pi_\theta(s) \int \pi_\theta(a|s)r(s, a)dsda
\]

\[
= \mathbb{E}_{s \sim d^\pi_\theta(s), a \sim \pi_\theta(a|s)}[r(s, a)],
\]

where \( d^\pi_\theta(s) \) is the discounted state distribution induced by the policy \( \pi_\theta \), defined as

\[
d^\pi_\theta(s) = (1 - \gamma) \int_{s} \sum_{t=1}^{\infty} \gamma^{t-1} p(s_1)p(s_t = s|s_1; \pi_\theta)ds_1.
\]

We denote the state–occupancy measure of state-action pair by \( d^\pi_\theta(s, a) = \pi_\theta(a|s)d^\pi_\theta(s) \). Similarly, we define the state-value and action-value functions

\[
V^\pi_\theta(s) = \lim_{H \to \infty} \mathbb{E}[R_t^1(H)|s_1 = s; \pi_\theta] = \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \right| s_1 = s; \pi_\theta]
\]

\[
Q^\pi_\theta(s, a) = \lim_{H \to \infty} \mathbb{E}[R_t^1(H)|s_1 = s, a_1 = a; \pi_\theta] = \mathbb{E} \left[ \sum_{t=1}^{\infty} \gamma^{t-1} r(s_t, a_t) \right| s_1 = s, a_1 = a; \pi_\theta]
\]

### 3.2 Generic Policy Gradient

Stochastic policy gradient ascent is a popular method to solve the RL optimization problem [2]. At each \( k \)-th iteration, we can iteratively update the policy parameters by

\[
\theta_{k+1} = \theta_k + \eta_k \hat{\nabla} \mu(\theta_k),
\]

where \( \eta_k \) is learning rate or step size and \( \hat{\nabla} \mu(\theta_k) \) is an estimator of policy gradient \( \nabla \mu(\theta_k) \). For convenience, \( \nabla \) denotes the gradient with respect to \( \theta \) unless specified otherwise.
To unify both infinite- and finite-horizon MDPs, we define the basic unit of samples for reuse as a trajectory $x = \tau$ for finite horizon MDP or a state-action pair $x = (s, a)$ for infinite horizon MDP. We let the sampling distribution of each trajectory observation $\rho(x) = D(\tau; \theta)$ for finite-horizon case and each state-action transition observation $\rho(x) = d^*(s, a)$ for infinite-horizon case. The unified notations will allow us to develop a generic VRER framework for both step- and episode-based algorithms in the following sections.

Under some regular conditions [Sutton et al., 1999b, Baxter and Bartlett, 2001, Williams, 1992], the generic policy gradient is shown as

$$\nabla \mu(\theta) \equiv \mathbb{E}_{x \sim \rho(x)}[g(x)|\theta] = \begin{cases} R(\tau) \sum_{t=1}^{H} \nabla \log \pi_{\theta}(a_t|s_t), & \text{finite horizon} \\ A^{*}(s, a) \nabla \log \pi_{\theta}(a|s), & \text{infinite horizon} \end{cases}$$  \tag{5}

where $A^{*}(s, a) \equiv Q^{*}(s, a) - V^{*}(s)$ is called advantage and it intuitively measures the extra reward that an agent can obtain by taking a particular action $a$. This leads to the scenario-based policy gradient estimate,

$$g_k(x) \equiv g(x|\theta_k) = \begin{cases} R(\tau) \sum_{t=1}^{H} \nabla \log \pi_{\theta_k}(a_t|s_t), & \text{finite horizon} \\ A^{*}(s, a) \nabla \log \pi_{\theta_k}(a|s), & \text{infinite horizon}. \end{cases}$$  \tag{6}

The baseline classical policy gradient (PG) estimator in the $k$-th iteration is given by

$$\tilde{\nabla}^{PG}_k \equiv \tilde{\nabla}^{PG}_{\theta_k} = \frac{1}{n} \sum_{j=1}^{n} g_k(x^{(k,j)}).$$  \tag{7}

This is estimated by using $n$ new generated samples in the $k$-th iteration. It can be challenging to precisely estimate the policy gradient $\nabla \mu(\theta)$ in (5), especially when process stochasticity is high and each experiment is expensive. The on-policy policy optimization algorithms only use new samples generated in the $k$-th iteration; while all historical observations are discarded. As the samples from previous iterations often have relevant information for estimating the policy gradient in the current iteration, inspired by the importance sampling design paradigm [Feng and Staum, 2017, Metelli et al., 2020], we propose reusing historical samples to improve the estimation accuracy on policy gradient with the likelihood ratio accounting for the difference between the proposal and target distributions.

### 3.3 Regularity Conditions for Generic Policy Gradient Estimation

We summarize the standard assumptions and conditions for the regularity of generic policy gradient and MDPs, which will be used throughout this paper.

**A.1** (Bounded variance) There exists a constant $\sigma > 0$ that bounds the mean square error of classical PG estimator, i.e., $\mathbb{E}||\tilde{\nabla}^{PG}_k(\theta) - \nabla \mu(\theta)||^2 \leq \sigma$ for any $\theta \in \mathbb{R}^d$.

**A.2** Suppose the reward function and the policy $\pi_{\theta}$ satisfy the following conditions:

(i) The absolute value of the reward $r(s, a)$ is bounded uniformly, i.e., there exists a constant, say $U_r > 0$ such that $|r(s, a)| \leq U_r$ for any $(s, a) \in S \times A$.

(ii) The policy $\pi_{\theta}$ is differentiable with respect to $\theta$, and the score function is Lipschitz continuous and has bounded norm, i.e., for any $(s, a) \in S \times A$, there exist positive constant bounds, denoted by $L_\theta$, $U_\theta$, and $U_r$, such that

$$\nabla \log \pi_{\theta_1}(a|s) - \nabla \log \pi_{\theta_2}(a|s) \leq L_\theta \|\theta_1 - \theta_2\|$$ \tag{8}

$$\nabla \log \pi_{\theta}(a|s) \leq U_\theta$$ \tag{9}

$$\|\pi_{\theta_1}(\cdot|s) - \pi_{\theta_2}(\cdot|s)\|_{TV} \leq U_r \|\theta_1 - \theta_2\|$$ \tag{10}

where $d_{TV}(P, Q)$ is the total variation norm between two probability measure $P$ and $Q$, which is defined as $d_{TV}(P, Q) = 1/2 \int_{X} |P(dx) - Q(dx)|$.

**A.3** (Independence) For infinite-horizon MDP, the random tuples $(s_t, a_t, s_{t+1})$ with $t = 1, 2, \ldots$, are drawn from the stationary distribution of the Markov decision process independently across time [Kumar et al., 2019].

Assumption A.1 is a standard assumption that is often used in stochastic approximation analysis. A.1 was originally introduced by Robbins and Monro [1951] and it is defined based on the bound of the $L_2$ norm inducing a metric $d(X, Y) = \|X - Y\|_2$. It states that the classical PG estimator, denoted by $\tilde{\nabla}^{PG}_k(\theta)$, is unbiased and has bounded
variance. The first two items in Assumption A.2 are standard assumptions on the regularity of the MDP problem and the parameterized policy, which are the same conditions used in many recent studies [Wu et al., 2020; Zhang et al., 2020; Kumar et al., 2019]. The third item is adopted from Xu et al. [2020] which holds for any smooth policy with bounded action space or Gaussian policy; see Lemma 1 in Xu et al. [2020]. The independence assumption A.3 does not hold in practice, but it is a standard used to deal with convergence bounds for infinite-horizon MDP problems [Kumar et al., 2019; Wang et al., 2020; Qu et al., 2021].

The uniform boundedness of the reward function \( r(s, a) \) in Assumption A.2(i) also implies that the absolute value of the Q-function is upper bounded by \( \frac{U_r}{1 - \gamma} \), since by definition

\[
|Q^*(s, a)| \leq \sum_{t=1}^{\infty} \gamma^{t-1} U_r = \frac{U_r}{1 - \gamma}, \text{ for any } (s, a) \in S \times A.
\] (11)

The same bound also applies for \( V^\pi(s) \) for any \( \theta \) and \( s \in S \)

\[
|V^\pi(s)| \leq |E[Q^*(s, a)]| \leq E[||Q^*(s, a)||] \leq \sum_{t=1}^{\infty} \gamma^{t-1} U_r = \frac{U_r}{1 - \gamma}.
\] (12)

This applies to the objective \( \mu(\theta) \) because \( |\mu(\theta)| = |E[V^\pi(s)]| \leq \frac{U_r}{1 - \gamma} \). In the similar way, we can show for a finite-horizon MDP, the objective is bounded by \( |\mu(\theta)| \leq HU_r \). Thus, we have the bound of the objective as \( U_\mu = \frac{U_r}{1 - \gamma} \) for finite horizon MDP and \( U_\mu = HU_r \) for finite horizon MDP.

Under Assumption A.2 we can establish the Lipschitz continuity of the policy gradient \( \nabla \mu(\theta) \) as shown Lemma 1. The proof is deferred to Lemma 3.2 in Zhang et al. [2020]. Lemma 1 is another essential condition to ensure the convergence of many gradient descent based algorithms; see for example Niu et al. [2011], Reddi et al. [2016], Nemirovski et al. [2009], Li and Orabona [2019]. It implies that the gradient cannot change abruptly, i.e., \( |\mu(\theta_2) - \mu(\theta_1)| = |\nabla \mu(\theta_1)^T (\theta_2 - \theta_1)| \leq L||\theta_2 - \theta_1||^2 \) (Nesterov [2003], Lemma 1.2.3).

**Lemma 1** (Zhang et al. [2020], Lemma 3.2). Under Assumption A.2, the policy gradient \( \nabla \mu(\theta) \) is Lipschitz continuous with some constant \( L > 0 \), i.e., for any \( \theta_1, \theta_2 \in \mathbb{R}^d \),

\[
||\nabla \mu(\theta_1) - \nabla \mu(\theta_2)|| \leq L||\theta_1 - \theta_2||.
\]

At the end of this section, we present Lemma 2 that establishes the boundedness of policy gradient and its stochastic estimate. There exists a constant \( M > 0 \) such that the \( L_2 \) norm of the scenario-based policy gradient estimate is bounded, i.e., \( ||g(x)|| \leq M \).

**Lemma 2** (Properties of Stochastic Policy Gradients). For any \( \theta \), the norm of the policy gradient \( \nabla \mu(\theta) \) and its scenario-based stochastic estimate \( g(x|\theta) \) are bounded, i.e.,

\[
g(x|\theta) \leq M \text{ and } \nabla \mu(\theta) \leq M
\]

where \( M = H^2U_rU_\Theta \) for finite-horizon MDP and \( M = \frac{2U_rU_\Theta}{1 - \gamma} \) for infinite-horizon MDP.

**Proof.** From Equation (9), we have the scenario-based stochastic gradient estimate

\[
||g(x|\theta)|| \leq \begin{cases} \sum_{t=1}^{H} r_t \sum_{\epsilon=1}^{H} \| \nabla \log \pi_\theta(a|s) \|, & \text{finite horizon} \\ ||Q^\pi - V^\pi|| \| \nabla \log \pi_\theta(a|s) \|, & \text{infinite horizon} \end{cases}
\]

\[
\leq \begin{cases} \frac{H^2U_rU_\Theta}{1 - \gamma}, & \text{finite horizon} \\ \frac{2U_rU_\Theta}{1 - \gamma}, & \text{infinite horizon} \end{cases}
\] (13)

where step (13) follows Assumption A.2(ii) and the boundness of value functions (Equation (11) and (12)). Let \( M = H^2U_rU_\Theta \) for finite-horizon MDP and \( M = \frac{2U_rU_\Theta}{1 - \gamma} \) for infinite-horizon MDP. The boundness of policy gradient \( \nabla \mu(\theta) \) follows the fact that

\[
||\nabla \mu(x)|| = ||E_{\pi(x)}[g(x|\theta)]|| \leq E_{\pi(x)}[||g(x|\theta)||] \leq E[M] = M,
\]

where step (⋆) follows by applying Jensen’s inequality and (**) follows by applying (13). \( \square \)
Various policy gradient algorithms are proposed during recent years. The forms of policy gradient estimators may vary from one to another. For example, the total discounted rewards of a trajectory $R(τ)$ can be replaced by the reward-to-go (future total discounted reward after taking action $a_t$), i.e., $\sum_{t'=\tau}^{T} \gamma^{t-t'} r_t'$. The advantage function can be also replaced with action value function or TD residual. Interested readers for other variants are referred to [Sutton and Barto 2018] and [Schulman et al. 2015b] for the derivations of general policy gradient algorithms. We emphasize that the proposed VRER approach is general and it can be adopted to most policy gradient optimization algorithms. In the empirical study in Section 6, we demonstrate this generality by using three well-known policy optimization algorithms, i.e., VPG, PPO, and TRPO.

4 Mixture Likelihood Ratio Assisted Policy Gradient Estimation

In this section, we describe how to utilize important sampling (IS) and multiple likelihood ratio (MLR) to improve the policy gradient estimation through reusing historical samples. In specific, at each $k$-th iteration, we estimate the performance of the candidate target policy $ρ_{θ_k}$. Let $ρ_{θ_k}(x)$ denote the target distribution at the $k$-th iteration and $ρ_{θ}(x)$ denote the proposal distribution induced by the behavioral policy from the $i$-th iteration. Let $F_k$ denote the set of all behavioral distributions that have been visited by the beginning of the $k$-th iteration. Let $U_k$ be a reuse set with $U_k \subseteq F_k$ including the MDP model candidates whose historical samples are selected and reused for estimating $∇μ(θ_k)$. Denote its cardinality as $|U_k|$. For discussions in this section we assume $U_k$ is given. We will present how to design and automatically create the reuse set $U_k$ in Section 5.

4.1 Importance Sampling based Policy Gradient Estimators

Importance sampling is a technique for evaluating the properties of a target distribution $ρ_{θ_k}(x)$ of interest by reusing samples generated from a different proposal distribution $ρ_{θ}(x)$; see [Owen 2013]. The well-known importance sampling (IS) identity $E_{x \sim ρ_{θ_k}}[g_k(x)] = E_{x \sim ρ_{θ}(x)}[g_k(x)f_{i,k}(x)]$ holds provided that $ρ_{θ_k}(x)/ρ_{θ}(x) = 0$ whenever $ρ_{θ}(x) = 0$, where $f_{i,k}(x) = ρ_{θ_k}(x)/ρ_{θ}(x)$ is likelihood ratio (LR). Intuitively, the magnitude of the likelihood ratio is associated with the discrepancy of the probability measures $ρ_{θ_k}$ and $ρ_{θ}$. Here we present IS based policy gradient estimators studied in this paper.

1) Individual Likelihood Ratio (ILR) Policy Gradient Estimator. The policy gradient $∇μ(θ_k)$ can be estimated by the individual likelihood estimator (ILR), defined in (14), by using the $n$ historical samples generated by $ρ_{θ_k}(x)$,

$$\widehat{∇μ}_{i,k}^{ILR} = \frac{1}{n} \sum_{j=1}^{n} f_{i,k}(x^{(i,j)}) g_k(x^{(i,j)})$$

with $f_{i,k}(x) = ρ_{θ_k}(x)/ρ_{θ}(x)$.

The likelihood ratio $f_{i,k}(x)$ weights the historical samples appropriately to correct the mismatch between the MDP distributions induced by the behavioral and target policies specified by parameters $θ_i$ and $θ_k$.

One way to reuse all the observations associated with the behavioral distributions included in the reuse set $U_k$ is to average the ILR estimators for all selected proposal distributions, i.e., $θ_i \in U_k$, which we call the average ILR policy gradient estimator,

$$\widehat{∇μ}^{ILR}_k = \frac{1}{|U_k|} \sum_{θ_i \in U_k} \widehat{∇μ}_{i,k}^{ILR} (θ_k) = \frac{1}{|U_k|} \sum_{θ_i \in U_k} \sum_{j=1}^{n} \frac{ρ_{θ_k}(x^{(i,j)})}{ρ_{θ}(x^{(i,j)})} g_k(x^{(i,j)})$$

For simplification, we allocate a constant number of replications (i.e., $n$) for each visit at $θ$. The estimators in (14) and (15) are unbiased due to the fact

$$E_{ρ_{θ_i}}[\widehat{∇μ}_{i,k}^{ILR}(θ_k)] = \int ρ_{θ_i}(x)/ρ_{θ}(x) g_k(x)dx = \int ρ_{θ_k}(x)/ρ_{θ}(x) g_k(x)dx = ∇μ(θ_k).$$

2) Mixture Likelihood Ratio (MLR) Policy Gradient Estimator. Though the ILR estimators [14] and [15] are unbiased, their variances could be large or even infinite, as the likelihood ratio $f_{i,k}(x)$ can be large or unbounded; see [Veitch and Guibas 1995]. Thus, the recent studies [Feng and Stamura 2017], [Dong et al. 2018] propose multiple importance sampling or MLR that uses a mixture proposal distribution to overcome this drawback, which can lead to much more stable estimation in different applications.

Instead of viewing the samples from each behavioral proposal distribution $ρ_{θ}(x)$ in isolation and forming ILR estimators, we can view all the observations associated with $θ_i \in U_k$ collectively as stratified samples from the mixture...
Although truncating the likelihood ratios (LRs) introduces a bias, this bias can be bounded as a function of the Rényi divergence and the clipping threshold; see more details in Metelli et al. [2020]. Following the similar idea, we derive an upper bound under Assumption A.1, where the bound of bias is expressed as the probability that the LRs are greater than max value $U_f$, which can control the policy gradient estimation variance inflation. In this way, the mixture likelihood ratio puts higher weight on the samples that are more likely to be generated by the target distribution $\rho_{\theta_k}(x)$ without assigning extremely large weights on the others.

**Lemma 3.** Conditional on $F_k$, the proposed MLR policy gradient estimator is unbiased:

$$\mathbb{E} \left[ \hat{\nabla}_{\mu_k}^{MLR} \bigg| F_k \right] = \mathbb{E} \left[ g_k(x) \big| \theta_k \right] = \nabla \mu(\theta_k).$$

Based on the studies, e.g., Veach and Guibas [1995] and Feng and Staum [2017], we show that the MLR estimator is unbiased in Lemma 3, see the proof in the Appendix (Section B.1). The trace of the covariance matrix $\text{Tr}(\text{Var}[\hat{\nabla}_{\mu}(\theta)])$ of the gradient estimator $\hat{\nabla}_{\mu}(\theta)$ is often considered in many gradient estimation variance reduction studies [Greensmith et al., 2004]. This motivates us to use it as the metrics to select historical behavioral policies visited in previous iterations. Since we always include the samples generated in the $k$-th iteration, this mixture likelihood ratio is bounded, i.e., $f_k(x) \leq |U_k|$, which can control the policy gradient estimation variance inflation. In this way, the mixture likelihood ratio puts higher weight on the samples that are more likely to be generated by the target distribution $\rho_{\theta_k}(x)$ without assigning extremely large weights on the others.

**Proposition 1.** Conditional on $F_k$, the total variance of the MLR policy gradient estimator (16) is smaller and equal to that of the average ILR policy gradient estimator (15).

$$\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{MLR} \bigg| F_k \right] \right) \leq \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{ILR} \bigg| F_k \right] \right).$$

(3) **Clipped Likelihood Ratio (CLR) Policy Gradient Estimator.** Another technique for overcoming the undesirable heavy-tailed behavior is weight clipping [Combes 2008]. In specific, we can truncate the likelihood ratio $f_{i,k}(x) = \rho_{\theta_i}(x) / \rho_{\theta_k}(x)$ through applying the operator $\min(f_{i,k}(x), U_f)$, which sets the max LR to be a pre-specified constant, denoted by $U_f$. Formally speaking, the CLR policy gradient estimator can be expressed as

$$\hat{\nabla}_{\mu_k}^{CLR} = \frac{1}{|U_k|} \sum_{\theta_i \in U_k} \hat{\nabla}_{\mu_i,k}^{CLR}(\theta_k) = \frac{1}{|U_k|} \sum_{\theta_i \in U_k} \sum_{j=1}^{n} \min \left( \frac{\rho_{\theta_i}(x^{(i,j)})}{\rho_{\theta_k}(x^{(i,j)})}, U_f \right) g_k(x^{(i,j)}).$$

Although truncating the likelihood ratios (LRs) introduces a bias, this bias can be bounded as a function of the Rényi divergence and the clipping threshold; see more details in Metelli et al. [2020]. Following the similar idea, we derive an upper bound under Assumption A.1 where the bound of bias is expressed as the probability that the LRs are greater than max value $U_f$. The proof of Lemma 4 can be found in the Appendix (Section B.4).

**Lemma 4.** Suppose that Assumptions A.1, A.2 and A.3 hold. For a clipped likelihood ratio based policy gradient estimator $\hat{\nabla}_{\mu_k}^{CLR}$, with a finite clipping threshold $U_f > 1$, the bias and the variance of the CLR estimator can be upper bounded as

$$\left\| \mathbb{E} \left[ \hat{\nabla}_{\mu_k}^{CLR} \right] - \nabla \mu(\theta_k) \right\| \leq Z(k),$$

$$\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{CLR} \bigg| F_k \right] \right) \leq \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{ILR} \bigg| F_k \right] \right) + \frac{1}{n|U_k|} \| \nabla \mu(\theta_k) \|^2,$$

where $Z(k) = \frac{M}{|U_k|} \sum_{\theta_i \in U_k} \mathbb{E}_k \left[ \mathbbm{1} \{ f_{i,k}(x) > U_f \} \right]$ and $M$ is the bound defined in Lemma 2.
4.2 Policy Gradient Estimation for Finite- and Infinite-Horizon MDPs

(1) Finite-Horizon MDPs. At any $k$-th iteration, importance sampling allows us to compensate the mismatch between the behavioral distribution $\rho_{\theta_k}(s,a)$ and the target distribution $\rho_{\theta_k}(s,a)$. For policy gradient estimators under finite-horizon MDP, after cancelling the term $p(s_1)\prod_{t=1}^{H-1} p(s_{t+1}|s_t,a_t)$ in the likelihood ratio $f_{i,k}(x) = \rho_{\theta_i}(x)/\rho_{\theta_i}(x)$, the ILR policy gradient estimators in (15) (similar for CLR estimator (17)) can be simplified to

$$\nabla \mu_{i,k}^{ILR} = \frac{1}{|U|} \sum_{i,k} \prod_{t=1}^{H} \frac{\pi_{\theta_i}(a_{t+1}|s_t)}{\pi_{\theta_i}(a_{t+1}|s_t)} g_k(s_{t+1},a_{t+1})$$

and

$$\nabla \mu_{k}^{MLR} = \frac{1}{|U|} \sum_{i,k} \prod_{t=1}^{H} \frac{\pi_{\theta_i}(a_t|s_t)}{\pi_{\theta_i}(a_t|s_t)} f_k(\tau),$$

with $f_k(\tau) = \frac{1}{|U|} \sum_{i,k} \prod_{t=1}^{H} \frac{\pi_{\theta_i}(a_t|s_t)}{\pi_{\theta_i}(a_t|s_t)} f_k(s_t,a_t)$. 

(2) Infinite-Horizon MDPs. For infinite-horizon MDPs, step-based policy gradient algorithms need just a single likelihood ratio per state-action transition sample, i.e.,

$$\nabla \mu(\theta_k) = E_{\rho_{\theta_k}} \left[ \frac{\rho_{\theta_k}(s,a)}{\rho_{\theta_k}(s,a)} g_k(x) \right] \approx E_{\rho_{\theta_k}} \left[ \frac{\rho_{\theta_k}(s,a)}{\rho_{\theta_k}(s,a)} \nabla \log \pi_{\theta_k}(s,a) \right] \theta_k.$$

However, it requires to estimate the stationary distribution $\pi(s,a)$ in the state–occupancy measure ratios $\rho_{\theta_k}(s,a) = \pi(\theta_k)(s,a)$. Unlike finite-horizon MDPs, those ratios cannot be computed in closed form. This problem is also known as distribution corrections (DICE) in RL. Fortunately, a list of approaches has been recently proposed to address this challenge; see Nachum et al. [2019], Yang et al. [2020]. The step-based policy gradient estimator (18) can be approximated by replacing $\rho_{\theta_k}(s,a)$ with policy $\pi_\theta(s,a)$.

$$\nabla \mu(\theta_k) \approx E_{\rho_{\theta_k}} \left[ \frac{\pi_{\theta_k}(s,a)}{\pi_{\theta_k}(s,a)} g_k(x) \right] \theta_k.$$

It can be preserved the set of local optima to which gradient ascent converges. Although biased, this estimator has been widely used in many state-of-the-art off-policy algorithms [Schulman et al. 2015a, 2017, Wang et al. 2017] due to its simplicity and computational efficiency. Following the same idea, the ILR and MLR policy gradient estimators for step-based algorithms can be approximated by replacing $\rho_{\theta_k}(s,a)$ with policy $\pi_\theta(s,a)$.

In the ILR estimator (14), the ILR

$$f_{i,k}(x) = \rho_{\theta_i}(x) = \frac{\pi_{\theta_i}(x)}{\pi_{\theta_i}(x)} \approx \frac{\pi_{\theta_i}(x)}{\pi_{\theta_i}(x)}$$

in the CLR estimator (17), while using

We conclude this section by pointing out that we proceed theoretical analysis with the unbiased policy gradient (18) in the following sections. The approximator (19) can be used in the algorithm development and the empirical study to simplify the implementation.

5 Variance Reduction Experience Replay for Policy Optimization

Before one can use the IS based policy gradient estimators developed in Section 4, there is one key remaining question: How to select the behavioral policies, among all historical candidate policies and samples visited up to the $k$-th iteration, to be included in the reuse set $U_k$? Instead of reusing all of them, reusing only the most relevant samples can reduce computational cost, stabilize likelihood ratio calculations, and reduce the gradient estimation variance. In short, a well-designed selection rule can improve sample efficiency and stability of policy gradient estimation and further accelerate the learning of optimal policy.

In Section 5.1 we propose a selection criterion to construct the reuse set $U_k$ so that it can reduce the variance in policy gradient estimation. Since the proposed selection criterion needs to assess all historical samples and associated behavioral policies to determine $U_k$, the likelihood ratio calculations can be computationally expensive as the size of the reuse set increases. Thus, we derive two computationally efficient approximated selection rules in Section 5.2. Based on the proposed selection criteria, we propose the VERR based generic policy gradient optimization algorithm (called PG-VERR) in Section 5.3 and have the convergence analysis in Section 5.4 to show that it can accelerate the policy search. In Section 5.5, we show that the size of the reuse set $|U_k|$ increases to infinity almost surely (a.s.) as the number of iterations increases $k \to \infty$. The empirical study in Section 6 shows the finite sample performance is consistent with these asymptotic analytical conclusions.
5.1 Reuse Set Selection to Reduce Policy Gradient Estimation Variance

To improve the policy gradient estimation, we want to select the reuse set $U_k$ producing stable likelihood ratios and thus reducing the estimation variance of policy gradient. Even though the target distribution can be high dimensional for complex systems with high uncertainty, the proposed VRER based policy gradient estimators tend to intelligently put more weight on historical observations that are most likely sampled from the target distribution.

Motivated by this insight, we propose a selection criterion in Theorem 1 based on the comparison of variances of the baseline PG estimator (7) and the ILR estimator (14) to dynamically determine the reuse set $U_k$ for each $k$-th iteration. The variance difference between $\hat{\nabla}_{\mu_{i,k}}^{ILR}$ and $\hat{\nabla}_{\mu_{k}}^{PG}$ is induced by the discrepancy between the proposal and target distributions, i.e., $\rho_{\theta_i}$ and $\rho_{\theta_k}$. If $\theta_i = \theta_k$, the likelihood ratio in (14) equals to 1 so $\hat{\nabla}_{\mu_{i,k}}^{ILR} = \hat{\nabla}_{\mu_{k}}^{PG}$, therefore they have the same variance. The proposed selection criterion (20) puts a constant threshold $c$ on the variance inflation of $\hat{\nabla}_{\mu_{i,k}}^{ILR}$ compared to $\hat{\nabla}_{\mu_{k}}^{PG}$. It tends to select those behavioral distributions $\rho_{\theta_i}(x)$ with $\theta_i \in F_k$ that are close to the target distribution $\rho_{\theta_k}(x)$. To ensure that new samples generated in the current $k$-th iteration are always included in the set $U_k$, we consider the constant $c > 1$. In addition, Theorem 1 demonstrates that the selection criterion (20) guides an appropriate reuse of historical samples and leads to variance reduction of the ILR and MLR policy gradient estimators compared to the baseline PG estimator by a factor of $c/|U_k|$; see (21). The proof of Theorem 1 can be found in the Appendix (Section B.3).

Theorem 1. At the $k$-th iteration with the target distribution $\rho_{\theta_k}(x)$, the reuse set $U_k$ is created to include the behavioral distributions, i.e., $\rho_{\theta_i}(x)$ with $\theta_i \in F_k$, whose the total variance of the ILR policy gradient estimator is no greater than $c$ times that of the classical PG estimator for some constant $c > 1$. Mathematically,

$$\text{Selection Rule 1: } \quad \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_{i,k}}^{ILR} \big| F_k \right] \right) \leq c \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_{k}}^{PG} \big| \theta_k \right] \right)$$

(20)

Then, based on such reuse set $U_k$, the total variance of the MLR/ILR policy gradient estimators (15) and (16) are no greater than $c/|U_k|$ times the total variance of the PG estimator,

$$\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_{i,k}}^{MLR} \big| F_k \right] \right) \leq c |U_k| \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_{k}}^{PG} \big| \theta_k \right] \right).$$

(21)

Moreover,

$$\mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_{i,k}}^{ILR} \right\|^2 \bigg| F_k \right] \leq \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_{i,k}}^{PG} \right\|^2 \bigg| \theta_k \right] + \left( 1 - \frac{c}{|U_k|} \right) \| \nabla \mu(\theta_k) \|^2.$$  

(22)

With the selection rule (20), Lemma 4 allows us to construct a similar bound on the expected squared norm of CLR policy gradient estimator as the bound (22); see Corollary 1. The proof is provided in the Appendix (Section B.5).

Corollary 1. At the $k$-th iteration with the target distribution $\rho_{\theta_k}(x)$, we apply the selection rule from Theorem 1. Then, based on such reuse set $U_k$, we have the following bound

$$\mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_{i,k}}^{CLR} \right\|^2 \bigg| F_k \right] \leq \frac{c}{|U_k|} \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_{i,k}}^{PG} \right\|^2 \bigg| \theta_k \right] + \left( 8 + \frac{1}{n |U_k|} - \frac{c}{|U_k|} \right) \| \nabla \mu(\theta_k) \| + 8 \sigma^2.$$  

(23)

At the end of this section, we provide the estimators of the conditional variances of the baseline PG estimator (7) and the ILR estimator (14) to support the implementation of the selection rule (20). Particularly, the conditional variance of the PG estimator (7) is estimated by

$$\text{Var} \left[ \hat{\nabla}_{\mu_{i,k}}^{PG} \big| \theta_k \right] = \frac{1}{n-1} \sum_{j=1}^{n} \tilde{\sigma}_{i,j} \tilde{\sigma}_{j}^\top, \quad \tilde{\sigma}_{j} = g_k \left( x^{(k,j)} \right) - \frac{1}{n} \sum_{j=1}^{n} g_k \left( x^{(k,j)} \right).$$

(24)

Its trace becomes $\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_{i,k}}^{PG} \big| \theta_k \right] \right) = \frac{1}{n-1} \sum_{j=1}^{n} \| \tilde{\sigma}_{j} \|^2$. The conditional variance estimator of the ILR gradient estimator (14) and its trace are

$$\text{Var} \left[ \hat{\nabla}_{\mu_{i,k}}^{ILR} \big| F_k \right] = \frac{1}{n-1} \sum_{j=1}^{n} \tilde{\sigma}_{i,j} \tilde{\sigma}_{j}^\top \text{ and Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_{i,k}}^{ILR} \big| F_k \right] \right) = \frac{1}{n-1} \sum_{j=1}^{n} \| \tilde{\sigma}_{i,j} \|^2,$$

(25)

where $\tilde{\sigma}_{i,j} = \frac{\rho_{\theta_i}(x^{(i,j)})}{\rho_{\theta_k}(x^{(i,j)})} g_k \left( x^{(i,j)} \right) - \hat{\nabla}_{\mu_{j,k}}^{ILR}$ for $\theta_i \in F_k$ and $j = 1, 2, \ldots, n$.

Proposition 2 shows the variance estimators of PG and ILR policy gradient estimators in (24) and (25) are conditionally unbiased; see the proof in the Appendix (Section B.6).
Proposition 2. Given $\mathcal{F}_k$, the policy gradient variance estimators $\hat{\Var} \left[ \nabla \mu_k^{PG} | \theta_k \right]$ in (24) and $\hat{\Var} \left[ \nabla \mu_k^{ILR} | \mathcal{F}_k \right]$ in (25) are unbiased.

5.2 Gradient Variance Ratio Approximation

The selection rule (20) could be computationally expensive for complex systems, especially as $|\mathcal{F}_k|$ becomes large, since it requires the repeated calculation of the variances of ILR policy gradient estimators for each $\theta_i \in \mathcal{F}_k$.

To support real-time decision making, we provide first-order approximations to the policy gradient variance, $\Tr(\Var[\nabla \mu_k^{ILR} | \mathcal{F}_k]) / \Tr(\Var[\nabla \mu_k^{PG} | \theta_k])$, for episode-based algorithms and further derive a single-observation based approximation for step-based algorithms.

Proposition 3. The total variance ratio of ILR policy gradient estimator and PG estimator has the approximation

$$\frac{\Tr(\Var[\nabla \mu_k^{ILR} | \mathcal{F}_k])}{\Tr(\Var[\nabla \mu_k^{PG} | \mathcal{F}_k])} \approx \frac{\mathbb{E}_{x \sim \rho_{\theta_k}} \left[ e^{\nu(x, \theta_k, \theta_k)} + \nu(x, \theta_k, \theta_k)^2 \right] - \mathbb{E}_{x \sim \rho_{\theta_k}} \left[ g_k(x) \right]^2}{\mathbb{E}_{x \sim \rho_{\theta_k}} \left[ g_k(x) \right]^2}$$

with

$$\nu(x, \theta_k, \theta_k) \equiv (\theta_k - \theta_i)^\top \log \rho_{\theta_k}(x).$$

Specifically, we use $\nabla \log \rho_{\theta_k}(x) = \nabla \log \pi_{\theta_k}(a|s)$ for infinite-horizon MDPs and $\nabla \log \rho_{\theta_k}(x) = \sum_{t=1}^H \nabla \log \pi_{\theta_k}(a_t|s_t)$ for finite-horizon MDPs.

Proof. Let $\Lambda(x) = \log \rho_{\theta_k}(x) / \rho_{\theta_i}(x)$. By taking Taylor expansion of $\Lambda(x)$ at $\theta_k$, we have

$$\Lambda(x) \approx (\theta_k - \theta_i)^\top \nabla \log \rho_{\theta_k}(x) - (\theta_k - \theta_i)^\top \nabla^2 \log \rho_{\theta_k}(x)(\theta_k - \theta_i).$$

By applying the well-known lemma [DeGroot and Schervish, 2012 Theorem 8.8.1], for any density function $p(x|\theta)$, the Fisher information matrix becomes,

$$I(\theta) = \mathbb{E} \left[ \nabla \log p(x|\theta) \nabla \log p(x|\theta)^\top \right] = -\mathbb{E} \left[ \nabla^2 \log p(x|\theta) \right].$$

Thus, the log likelihood ratio in (27) can be further approximated by replacing the second order derivative $\nabla^2 \log \rho_{\theta_k}(x)$ with a single observation empirical Fisher information, i.e.,

$$\Lambda(x) \approx (\theta_k - \theta_i)^\top s(x|\theta_k) + (\theta_k - \theta_i)^\top \tilde{I}(x|\theta)(\theta_k - \theta_i)$$

where $s(x|\theta_k) = \nabla \log \rho_{\theta_k}(x)$ and $\tilde{I}(x|\theta) = \nabla \log \rho_{\theta_k}(x) \nabla \log \rho_{\theta_k}(x)^\top$. Due to the fact that $(\theta_k - \theta_i)^\top s(x|\theta_k) \nabla \log \rho_{\theta_k}(x)(\theta_k - \theta_i)$, the likelihood ratio can be written as

$$\frac{\rho_{\theta_k}(x)}{\rho_{\theta_i}(x)} \approx e^{\nu(x, \theta_k, \theta_i) + \nu(x, \theta_i, \theta_i)}$$

where $\nu(x, \theta_i, \theta_k) = (\theta_k - \theta_i)^\top s(x|\theta_k)$. Let $g_{i,k}(x) = \frac{\rho_{\theta_k}(x)}{\rho_{\theta_i}(x)} g_k(x)$. Then we have

$$\mathbb{E}_{x \sim \rho_{\theta_k}} [g_{i,k}(x)^2] = \mathbb{E}_{x \sim \rho_{\theta_i}} \left[ \frac{\rho_{\theta_k}(x)}{\rho_{\theta_i}(x)} g_k(x)^2 \right] \approx \mathbb{E}_{x \sim \rho_{\theta_k}} \left[ e^{\nu(x, \theta_k, \theta_i) + \nu(x, \theta_i, \theta_i)} g_k(x)^2 \right]$$

where (29) holds by replacing the ratio $\frac{\rho_{\theta_k}(x)}{\rho_{\theta_i}(x)}$ by approximation (28). After that, we can derive the estimation variance of policy gradient, i.e.,

$$\Tr(\Var[g_{i,k}(x)]) = \Tr \left( \mathbb{E}_{x \sim \rho_{\theta_i}} [g_{i,k}(x)^2] - \mathbb{E}_{x \sim \rho_{\theta_i}} [g_{i,k}(x)] \mathbb{E}_{x \sim \rho_{\theta_i}} [g_{i,k}(x)]^\top \right)$$

$$= \Tr \left( \mathbb{E}_{x \sim \rho_{\theta_k}} [g_{i,k}(x)^2] - \left( \mathbb{E}_{x \sim \rho_{\theta_k}} [g_{i,k}(x)] \right)^2 \right)$$

$$\approx \mathbb{E}_{x \sim \rho_{\theta_k}} \left[ e^{\nu(x, \theta_k, \theta_i) + \nu(x, \theta_i, \theta_i)} \|g_k(x)\|^2 \right] - \left( \mathbb{E}_{x \sim \rho_{\theta_k}} [g_k(x)] \right)^2 .$$
Note that $\text{Var}\left[\hat{n}^{ILR}_{\mu_{i,k}} \mid \mathcal{F}_k\right] = \frac{1}{n} \text{Var}\left[\frac{\rho_{\theta_k}(x)}{\omega(x)} g_k(x)\right] = \frac{1}{n} \text{Var}[g_{i,k}(x)]$. Therefore, we have

$$\text{Tr}\left(\text{Var}\left[\hat{n}^{ILR}_{\mu_{i,k}} \mid \mathcal{F}_k\right]\right) \approx \frac{1}{n} \left(\mathbb{E}_{x \sim \rho_{\theta_k}} \left[e^{\nu(x, \theta, \theta_k) + \nu(x, \theta, \theta_k)} \|g_k(x)\|^2 - \|\mathbb{E}_{x \sim \rho_{\theta_k}} [g_k(x)]\|^2\right]\right).$$

Since $\text{Var}\left[\hat{n}^{PG}_{\mu_{k}} \mid \mathcal{F}_k\right] = \frac{1}{n} \text{Var}[g_k(x)]$, we have the approximation,

$$\frac{\text{Tr}\left(\text{Var}\left[\hat{n}^{ILR}_{\mu_{i,k}} \mid \mathcal{F}_k\right]\right)}{\text{Tr}\left(\text{Var}\left[\hat{n}^{PG}_{\mu_{k}} \mid \mathcal{F}_k\right]\right)} \approx \frac{\mathbb{E}_{x \sim \rho_{\theta_k}} \left[e^{\nu(x, \theta, \theta_k) + \nu(x, \theta, \theta_k)} \|g_k(x)\|^2 - \|\mathbb{E}_{x \sim \rho_{\theta_k}} [g_k(x)]\|^2\right]}{\text{Tr}(\text{Var}_{x \sim \rho_{\theta_k}} [g_k(x)])}. \square$$

Proposition [3] provides us a way to approximate the total variance ratio without retrieving historical samples and likelihoods. The approximation (20) only requires the calculation of $\nu(x, \theta, \theta_k) = (\theta_k - \theta) \top \nabla \log \rho_{\theta_k}(x)$ and $g_k(x)$ based on samples from the current $k$-th iteration. Then by using the approximation (20), the selection rule becomes

**Selection Rule 2:**

$$\frac{\mathbb{E}_{x \sim \rho_{\theta_k}} \left[e^{\nu(x, \theta, \theta_k) + \nu(x, \theta, \theta_k)} \|g_k(x)\|^2 - \|\mathbb{E}_{x \sim \rho_{\theta_k}} [g_k(x)]\|^2\right]}{\text{Tr}(\text{Var}_{x \sim \rho_{\theta_k}} [g_k(x)])} \leq c. \quad (31)$$

One drawback of this approximation is that the variance estimation of baseline PG still requires computing the gradient variance from independent samples. However, step-based algorithms usually do not collect samples independently. Therefore, we simplify the selection rule (31) by using a single scenario approximation on the expectations and obtain

$$\text{E}_{x \sim \rho_{\theta_k}} \left[e^{\nu(x, \theta, \theta_k) + \nu(x, \theta, \theta_k)} \|g_k(x)\|^2 - \|\mathbb{E}_{x \sim \rho_{\theta_k}} [g_k(x)]\|^2\right] \leq c \|g_k(x)\|^2 - c \|\mathbb{E}_{x \sim \rho_{\theta_k}} [g_k(x)]\|^2. \quad \text{By observing that the right hand side equals zero, this inequality can be further simplified to a new approximated selection rule, i.e.,}$$

**Selection Rule 3:**

$$\exp(\nu(x, \theta, \theta_k) + \nu(x, \theta, \theta_k)) \leq 1. \quad (32)$$

Unlike (20) and (31), this selection rule no longer requires the specification of the selection threshold constant $c$.

**Remark 1.** Unlike the episode-based approaches whose trajectory samples are generated independently from certain behavioral policy, the state transitions of step-based approach are typically correlated. This sample correlation can impact on the gradient variance estimation as the variance of corrected samples also includes covariance between samples, e.g., $\text{Var}(\sum_{i=1}^{\infty} X_i) = \sum_{i=1}^{\infty} \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$ with $X_i = g_k(x^{(k,i)})$. As a result, to use the selection rule (20) and (31) for step-based algorithms, we also need to take the sample covariance into consideration if the policy gradient estimator is based on multiple samples.

**Remark 2.** It is worth noting that the left hand side of the new criterion (32) is just the approximated likelihood ratio, which intuitively measures the closeness between policies $\theta_1$ and $\theta_2$; see Equation (28). In the empirical study, we find that the approximated selection rule works very well for step-based algorithms such as PPO and TRPO.

### 5.3 VRER Assisted Generic Policy Optimization Algorithm

In this section, we present the generic procedure of variance reduction experience replay based policy optimization, called PG-VRER, in Algorithm 1. At each $k$-th iteration, we generate $n$ new samples, denoted by $\mathcal{T}_k$, by following the candidate policy $\pi_{\theta_k}$ and update the data set $D_k$ in Step 1. We select the historical samples satisfying the selection rule (20), (31) or (32) and create the reuse set $\mathcal{U}_k$ in Step 2. Then, we compute the ILR, CLR, and/or MLR policy gradient estimate by applying (15)–(17) and update policy parameters through the offline optimization in Step 3. In specific $h$-th iteration of offline policy optimization, we use randomly selected historical samples, denoted by $D_k^{\rho_h}$, to have mini-batch stochastic gradient update, $\theta_{k+h}^{h+1} \leftarrow \theta_{k+h} + \eta_h \nabla \mu_{k,h}$. After finishing the off-line optimization, we update the set $\mathcal{F}_{k+1}$ through a circular buffer with size $B$, i.e., removing the oldest policy $\theta_{k+1-B}$ and inserting the latest one $\theta_{k+1}$ in Step 4. We repeat this whole procedure until reaching to the iteration budget, denoted by $K$.

The proposed VRER is generic and it is applicable to most policy gradient optimization algorithms. Thus, the Execution (Step 1) and Offline Optimization (Step 3) can vary from one policy gradient algorithm to anther. The proposed PG-VRER can be applicable to both episode- and step-based approaches. For episode-based approaches, although we can directly compute the selection criterion (20) based on $\text{Tr}(\text{Var}[\hat{n}^{ILR}_{\mu_{i,k}} \mid \mathcal{F}_k])$ and $\text{Tr}(\text{Var}[\hat{n}^{PG}_{\mu_{k}} \mid \mathcal{F}_k])$ by using (25) and (24), it is recommended to use the variance ratio approximation (20) together with selection rule (20) due to its computational simplicity. Basically, the implementation of the selection rule (20) can be computationally expensive.
Algorithm 1: VRER Assisted Generic Policy Gradient Algorithm (PG-VRER)

**Input:** the selection threshold constant $c$; the maximum number of iterations $K$; buffer size $B$; the number of replications per iteration (batch size) $n$; the number of iterations in offline optimization $K_{off}$; the set of historical samples $D_0$; the set of policy parameters visited so far $F_0$.

**Initialize** policy $\theta_1$ and set $\theta_0 = \theta_1$. Then store the parameter, i.e., $F_1 = F_0 \cup \{ \theta_1 \}$.

for $k = 1, 2, \ldots, K$ do

1. **Execution:**
   
   (a) Generate $n$ independent new samples, denoted by $T_k = \{ x^{(k,j)} \sim \rho_{\theta_k}(x) : j = 1, 2, \ldots, n \}$;
   
   (b) Store them $D_k \leftarrow D_{k-1} \cup T_k$;

2. **Selection:**
   
   (a) Initialize $U_k = \emptyset$ and the set $D_k^{n_0} = \emptyset$;
   
   for $\theta_i \in F_k$ do
     
     (b) Verify the selection criterion (see Table 1);
     
     if Selection Criterion is Satisfied then
       
       (c) Update the reuse set: $U_k \leftarrow U_k \cup \{ \theta_i \}$;
       
     (d) Randomly sample $n_0$ observations $T_k^{n_0} = \{ x^{(i,j)} \sim T_i : j = 1, 2, \ldots, n_0 \}$ and store them to the set $D_k^{n_0} = D_k^{n_0} \cup T_k^{n_0}$;
     
   end

3. **Offline Optimization (Algorithm-Specific):**
   
   for $h = 0, 1, \ldots, K_{off}$ do
     
     (a) Sample a mini-batch from $D_k^{n_0}$ to compute the policy gradient estimates $\hat{\nabla}_{\mu_{k,h}}^{ILR}, \hat{\nabla}_{\mu_{k,h}}^{CLR}$ or $\hat{\nabla}_{\mu_{k,h}}^{MLR}$ defined in Equation (15–17);
     
     (b) Update the policy parameters by $\theta_{k+1} \leftarrow \theta_k + \eta_k \hat{\nabla}_{\mu_{k,h}}$ using the gradient estimate $\hat{\nabla}_{\mu_{k,h}}$ from step 3(a).
     
   end

4. Set $\theta_{k+1} = \theta_{k+1}^{K_{off}}$ and $\theta_0 = \theta_{k+1}$. Store it to the set $F_{k+1} = F_k \cup \{ \theta_{k+1} \}$. If $|F_{k+1}| > B$, remove $\theta_{k+1-B}$ from $F_{k+1}$. 

end

due to the fact that it needs to repeatedly calculate the likelihood ratios between all behavioral and target distributions for sample selection and policy gradient estimation. In contrast, the approximations in (31) and (32) need to compute $\nu(x, \theta_i, \theta_k) = (\theta_k - \theta_i)^T s(x|\theta_k)$ and $g_k(x)$, where the score function $s(x|\theta_k)$ and the gradient $g_k(x)$ only need to be calculated once. That is to say, the time-varying computing complexity only comes from calculating the parameter difference $(\theta_k - \theta_i)$ for all $i < k$ with $\theta_i \in F_k$.

| Selection Rule | Approximation | Computation | Algorithm Type |
|----------------|--------------|-------------|---------------|
| (20)           | Yes          | High        | Episode-based Reuse |
| (31)           | No           | Low         | Episode-based Reuse |
| (32)           | No           | Low         | Step-based Reuse |

In summary, we list all three versions of selection rule in Table 1. Due to the computational complexity, the selection criterion (31) turns out to be more suitable to episode-based algorithms. For step-based approaches, we recommend the selection rule (32) as it does not require estimating the total variance of policy gradient with independent samples. Unlike the episode-based alternatives, the samples of step-based approaches are collected sequentially, and thus often have dependence. This impacts on policy learning and gradient variance estimation. To reduce the correlation impact, we randomly draw $n_0$ samples from historical observations following each selected policy, i.e., $D_k^{n_0} = \bigcup_{\theta_i \in U_k} T_i^{n_0} = \bigcup_{\theta_i \in U_k} \{ x^{(i,j)} \sim T_i : j = 1, 2, \ldots, n_0 \}$. In practice, in spite of the approximation error, the selection rule (32) tends to show better performance than the other two rules for step-based algorithms such as PPO and TRPO.

**Remark 3.** The replay buffer size, specified by $B$, can be fixed or infinite; for example a buffer size proportional to the number of iterations (e.g., select $B = k/3$ most recent samples). There is always a trade-off on specifying the reuse set size to balance exploration and exploitation. As Fedus et al. [2020] pointed out, there is an interplay between: (1) the improvements caused by increasing the replay capacity and covering large state-action space; and (2) the deterioration caused by having older policies in the buffer or overfitting to some "out-dated" samples. The magnitude of both effects depends on the particular settings of these quantities. Generally, as the age of the oldest policy increases, the benefits from increasing replay capacity are not as large. Thus, the buffer size $B$ of both $F_k$ and $D_k$ turns out to be critical to the
performance of most experience replay methods including the proposed VRER. We will empirically study the impact of buffer size in Section 6.

Remark 4. The importance sampling approach, used to leverage the information from historical samples, requires sampling distributions to be independent with each other. The interdependencies between historical samples can lead to an obstacle for the finite-time convergence analysis of most reusing mechanisms. To reduce this interdependence, we utilize randomly sampling strategy, i.e., draw with replacement \( n_0 \) (with \( n_0 < n \)) samples from selected historical observations with \( \theta_i \in \mathcal{U}_k \) in Step 2(d), and then use them to train the off-line policy optimization in Step 3 through mini-batch SG.

5.4 Finite-Time Convergence Analysis of Policy Optimization

We present the finite-time convergence analysis result in Theorem 2 for the proposed PG-VRER algorithm. It works for a general non-convex objective. This analytical study shows that the proposed VRER with the selection rule (20) can improve rate of convergence compared with the baseline PG approach in (7). Following the literature of non-convex optimization [Zhang et al., 2019; Zhou and Cong, 2018], we choose to state the convergence in terms of the average over all \( K \) iterations. This theorem presents the effect of learning rate \( \eta \) on the average gradient norm, as well as the balancing effect from the size of the reuse set \( |U_k| \) and the constant threshold \( c \) used for historical sample selection.

As \( |U_k| \) increases, the rate of convergence tends to increase; see (33). The proof of Theorem 2 can be found in the Appendix (Section B.7).

**Theorem 2.** Suppose Assumptions A.1, A.2 and A.3 hold. Let \( \eta_k \) denote the learning rate used in the \( k \)-th iteration and \( U_{\mu} \) is the bound of objective. If the ILR or MLR policy gradient estimators are used in Algorithm 1, we have

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \| \nabla \mu(\theta_k) \|^2 \right] \leq \frac{4U_{\mu} - 2 \eta \mu(\theta_1)}{\eta} + 4cL\sigma^2 \sum_{k=1}^{K} \frac{n_k}{|U_k|},
\]

where \( \sigma \) and \( L \) are the bounds defined in Assumption A.1 and Lemma 1, and \( c \) is the selection constant defined in Theorem 7.

Theorem 2 guarantees the local convergence of the proposed PG-VRER algorithm for general non-convex optimization problems if ILR or MLR policy gradient estimator is used in policy update. By choosing the learning rate \( \eta_k \), the convergence rate is between \( O(\frac{1}{\sqrt{K}}) \) and \( O(\frac{\ln K}{K}) \) depending on the size of the reuse set \( |U_k| \) as shown by Theorem 5.

Comparatively, under Assumptions A.1, A.2 and A.3, the baseline policy gradient methods without reusing historical samples have \( O(\frac{1}{\sqrt{K}}) \) rate of convergence with a diminishing stepsize of order \( O(\frac{1}{\sqrt{K}}) \) in the non-convex stochastic programming theory [Scaman and Malherbe, 2020; Ghadimi and Lan, 2013]. Theorem 3 sheds light on how the size of the reuse set \( |U_k| \) and the learning rate \( \eta_k \) impact on the convergence rate. If \( |U_k| \) grows in \( O(k^{1/p}) \), we can always choose a diminishing learning rate satisfying \( \max\{\frac{1}{k^{1-\frac{1}{p}}}, \frac{1}{k^{1+\frac{1}{p}}}, \frac{1}{k^{2}}\} \leq \frac{1}{\sqrt{K}} \), or equivalently set \( q \geq 2 \) and \( \frac{2}{q} = 1 - \frac{1}{p} \), such that the PG-VRER algorithm converges faster than \( O(\frac{1}{\sqrt{K}}) \). Ideally if all historical observations satisfy the selection criteria, that is \( |U_k| = k \), its optimal rate of convergence becomes \( O\left(\frac{\ln K}{K}\right) \) with a constant learning rate. The proof of Theorem 3 can be found in the Appendix (Section B.8).

**Theorem 3.** Suppose Assumptions A.1, A.2 and A.3 hold. If the size of reuse set \( |U_k| \sim O(k^{1/p}) \) and the learning rate \( \eta_k = \frac{\eta_0}{k^{1+\frac{1}{p}}} \) with \( p \geq 1, q > 1 \), the proposed PG-VRER algorithm (with ILR and MLR policy gradient estimators) has the bounded rate of convergence \( O(\max\{\frac{1}{k^{1-\frac{1}{p}}}, \frac{1}{k^{1+\frac{1}{p}}}, \frac{1}{k^{2}}\}) \).

The proposed algorithm can guarantee the convergence even with a fixed learning rate. By fixing \( \eta_k = \eta \) at the right hand side of (33), we have

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \| \nabla \mu(\theta_k) \|^2 \right] \leq \frac{4U_{\mu}/\eta - 2 \mu(\theta_1)/\eta + 4cL\sigma^2 \sum_{k=1}^{K} \frac{1}{|U_k|}}{K} \rightarrow 0 \text{ as } K \rightarrow \infty,
\]

since \( \lim_{K \rightarrow \infty} |U_k| = \infty \) almost surely which will be theoretically proved in Theorem 6.

In Theorem 4 we present an upper bound for the PG-VRER algorithm with CLR policy gradient estimator. Unsurprisingly, the convergence depends on bias introduced by clipping; see the term \( \frac{1}{K} \sum_{k=1}^{K} Z(k) \) in (34).

**Theorem 4.** Suppose Assumptions A.1, A.2 and A.3 hold. Let \( \eta_k \) denote the learning rate used in the \( k \)-th iteration and \( U_{\mu} \) is the upper bound of objective. For CLR policy gradient estimator, we have

\[
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \| \nabla \mu(\theta_k) \|^2 \right] \leq \frac{4U_{\mu} - 2 \eta \mu(\theta_1)}{\eta} + (4cL + 16)\sigma^2 \sum_{k=1}^{K} \frac{n_k}{|U_k|} + M \sum_{k=1}^{K} Z(k)
\]

(34)
where $\sigma$ and $L$ are the bounds defined in Assumption A.1 and Lemma A.2 and $c$ is the selection constant defined in Theorem 7. $M$ is the bound of policy gradient defined in Lemma 2 and $Z(k)$ is the bound of CLR policy gradient estimator defined in Lemma 2.

Lemma 5 provides a convergence guarantee for the bias of CLR policy gradient estimator with two additional Assumptions A.6 and A.7 (see Appendix B.9). We leave relaxation of those two assumptions for future work.

**Lemma 5.** Suppose Assumptions A.1, A.2 and A.3 as well as A.6 and A.7 defined in Appendix B.9 hold. Let the step size $\eta_k$ be an decreasing sequence converging to zero. Then $Z(k) \rightarrow 0$ as $k \rightarrow \infty$.

### 5.5 Asymptotic Analysis of VRER Reuse

As the size of the reuse set grows to infinity almost surely with iteration $k$, i.e., $P(\lim_{k \rightarrow \infty} |U_k| = \infty) = 1$, the variance term $\frac{4cL^2}{K} \sum_{k=1}^{K} \frac{\eta_k}{|U_k|}$ in the upper bound (33) will decrease faster than that of classic gradient descent methods. To show this asymptotic property of the reuse set $U_k$, we first introduce two additional assumptions.

**A.4 (Strong concavity)** The objective $\mu(\theta)$ is locally $\lambda$-strongly concave or equivalently $-\mu(\theta)$ is locally $\lambda$-strongly convex with a constant $\lambda > 0$, i.e., $-\mu(\theta_2) \geq -\mu(\theta_1) - \nabla \mu(\theta_1)^\top (\theta_2 - \theta_1) + \lambda \| \theta_2 - \theta_1 \|^2$ for $\forall \theta_1, \theta_2 \in B_c(\theta^*)$, where $B_c(\theta^*) \equiv \{ \theta : \| \theta - \theta^* \|^2 \leq r \}$ denotes a neighborhood around the local maximizer $\theta^*$ with radius $r > 0$.

**A.5 (Continuity)** The likelihood $\rho_\theta(x)$ and scenario-based policy gradient estimate $g(x)$ is continuous almost everywhere (a.e.) in the space of $(x, \theta)$.

Assumption A.4 is a weaker condition than the locally strong convexity (or concavity) assumption that is often used in many non-convex stochastic optimization studies; for example [Goebel and Rockafellar 2008] and [Jain and Kar 2017]. It rules out edge case, such as plateaus surrounding saddle points, and guarantees the convergence to a local optimum. Assumption A.5 is a regularity condition that is widely used in stochastic gradient descent methods enforcing the interchangeability of expectation and differentiation.

With these additional assumptions, we can establish the almost sure convergence for the proposed PG-VRER algorithm and use it for the asymptotic analysis of the reuse set $U_k$. Here, we firstly review a classical convergence result [Robbins and Siegmund 1971].

**Lemma 6 (Robbins-Siegmund).** Consider a filtration $\mathcal{C}_k$. Let $(V_k)_{k \geq 1}$, $(U_k)_{k \geq 1}$, and $(Z_k)_{k \geq 1}$ be three nonnegative $(\mathcal{C}_k)_{k \geq 1}$-adapted processes and a sequence of nonnegative numbers $(\alpha_k)_{k \geq 1}$ such that $\sum_k Z_k < \infty$ almost surely and $\prod_k (1 + \alpha_k) < \infty$ and $\forall k \in \mathbb{N}$, $\mathbb{E}[V_{k+1}|\mathcal{C}_k] + U_{k+1} \leq (1 + \alpha_k) V_k + Z_k$.

Then $(V_k)_{k \geq 1}$ converges to some random variable $V_\infty$ and $\sum_k U_k < \infty$ almost surely.

Lemma 6 has been used by [Sebbouh et al. 2021] to study the almost sure convergence of stochastic gradient descent. Following the similar idea, we can show that the policy parameters converge almost surely to some local optimum in Theorem 5. Then, given any feasible $x$, by applying continuous mapping theorem, we can have almost sure convergence of likelihood $\rho_{\theta^*}(x)$ in Proposition 4. The proofs of Theorem 5 and Proposition 4 can be found in the Appendix (Sections B.10-11).

**Theorem 5.** Assume A.1-A.5 hold. Let $\eta_k$ denote the learning rate in the $k$-th iteration and $\theta^*$ denote a local maximizer. We have $\theta_k \xrightarrow{a.s.} \theta^*$ as $k \rightarrow \infty$ if $\sum_{k=1}^{\infty} \eta_k^2 < \infty$.

**Proposition 4.** Assume A.1-A.5 hold. Let $\theta^*$ denote a local maximizer. Given any feasible $x$, we have $\rho_{\theta^*}(x) \xrightarrow{a.s.} \frac{1}{\rho_{\theta}(x)}$ as $k \rightarrow \infty$.

Finally, we present the main asymptotic property of the reuse set $U_k$ in Theorem 6, showing that its size grows and converges almost surely to infinity as $k \rightarrow \infty$ under Assumptions A.1-A.7. The proof of Theorem 6 can be found in the Appendix (Section B.12).

**Theorem 6.** Suppose Assumptions A.1-A.7 hold. The size of the reuse set $|U_k|$ increases to infinity with probability 1 as the iteration $k \rightarrow \infty$, i.e., $P(\lim_{k \rightarrow \infty} |U_k| = \infty) = 1$.

### 6 Empirical Study

In this section, we study the finite sample performance of the proposed PG-VRER framework. The empirical results demonstrate that it can enhance the state-of-art policy optimization approaches, including VPG, TRPO, and PPO, in
Section 6.1 Then, we study the pattern of experience replay (i.e., $|\mathcal{U}_b|$) in Section 6.2 and assess the effect of VRER on the policy gradient estimation variance reduction in Section 6.3. We also investigate the convergence behavior of PG-VRER with a fixed learning rate (i.e., constant $\eta_k$) in Section 6.4.

In the experiments of VPG with VRER (called VPG-VRER), we set the number of new generated samples in each iteration to be $n = 4$. For step-based algorithms, i.e., PPO and TRPO, we set batch size $n = 512$ having 4 parallel environments with each of them collecting 128 transition steps. In Section 6.4, we update policy parameters with classic SG optimizer with fixed step size. In all other cases, we use “Adam” optimizer [Kingma and Ba 2015]. We set the discount factor $\gamma = 0.99$ in all tasks. All experimental details are provided in Appendix A.

6.1 Comparison of Policy Optimization with and without VRER

Here we use three classical control benchmarks to illustrate that the proposed VRER can enhance the performance of the state-of-the-art policy optimization approaches, i.e., VPG (or REINFORCE), TRPO, and PPO algorithms. To represent the policy, VPG used a fully-connected Multilayer perceptron (MLP) with two shared hidden layers of 32 units, and ReLU (Rectified Linear Unit) activation functions. PPO used a fully-connected MLP with two shared hidden layers of 32 units, and tanh activation functions while TRPO algorithm has separate actor and critic neural network models; both of which have two layers with 32 neurons. For the problems with discrete actions, we use softmax activation function on top of the actor network, which calculates the probabilities of candidate actions. For the problem with continuous actions, we use the Gaussian policy for actor model with fixed standard deviations by following [Metelli et al., 2020].

![Figure 1: Convergence results for the VPG algorithm with and without using VRER.](image1)

![Figure 2: Convergence results for the TRPO algorithm with and without using VRER.](image2)

We set the same initial learning rate $\eta = 0.0003$ for both actor and critic in TPRO and PPO algorithms across all experiments. For VPG, the learning rate of policy model is set to be $\eta = 0.0003, 0.005$ and 0.0015 respectively for Acrobot, CartPole and Inverted Pendulum problems. The historical sample selection threshold is set to be $c = 1.5$ for all VPG experiments. The buffer size is set to $B = 400$ except PPO-Acrobot and TRPO-Inverted Pendulum case (for which the buffer size is set to be $B = 200$).

We plot the mean performance curves and 95% confidence intervals of VPG(-VRER) in Figure 1 and TRPO(-VRER) in Figure 2 and PPO(-VRER) in Figure 3. The results demonstrate that VRER improves the overall performance of the state-of-the-art policy optimization algorithms. For the VPG, we observe a significantly increased convergence speed and stability after using VRER. For the experiments related to PPO algorithm, PPO-VRER shows not only the
convergence to better optimum but also faster convergence compared to those without using VRER. In all three cases, the performance improvement for TRPO isn’t as significant as VPG and PPO. In addition, we observe a reduction in the variation of the convergences for those algorithms after implementing VRER.

Table 2 presents the performance (i.e., average total reward, standard error (SE), and percentiles) of all these algorithms over last 1000 training iterations for five different random seeds (same across all algorithms). In bold, the performances of using VRER are statistically significantly different from the baseline algorithm in each task.

| Algorithm   | Task         | Mean   | Standard Error | 5th Percentile | 95th Percentile |
|------------|--------------|--------|----------------|----------------|-----------------|
| PPO        | Acrobot      | -86.51 | 0.43           | -112.25        | -73.50          |
| PPO        | CartPole     | 168.66 | 1.06           | 86.98          | 200.00          |
| PPO        | Inverted Pendulum | 749.39 | 10.25         | 47.35          | 1000.00         |
| PPO-VRER   | Acrobot      | -86.95 | 0.40           | -108.64        | -74.00          |
| PPO-VRER   | CartPole     | 193.54 | 0.31           | 172.98         | 200.00          |
| PPO-VRER   | Inverted Pendulum | 941.55 | 3.37           | 668.54         | 1000.00         |
| TRPO       | Acrobot      | -97.41 | 0.44           | -123.00        | -81.25          |
| TRPO       | CartPole     | 197.66 | 0.26           | 179.00         | 200.00          |
| TRPO       | Inverted Pendulum | 603.77 | 11.24         | 13.00          | 1000.00         |
| TRPO-VRER  | Acrobot      | -96.93 | 0.48           | -125.00        | -79.00          |
| TRPO-VRER  | CartPole     | 199.69 | 0.05           | 198.50         | 200.00          |
| TRPO-VRER  | Inverted Pendulum | 684.87 | 9.88            | 37.50          | 1000.00         |
| VPG        | Acrobot      | -381.91 | 2.99          | -500.00        | -200.98         |
| VPG        | CartPole     | 163.69 | 1.37           | 69.61          | 200.00          |
| VPG        | Inverted Pendulum | 256.15 | 7.68           | 42.25          | 947.16          |
| VPG-VRER   | Acrobot      | -132.11 | 0.71          | -178.14        | -99.25          |
| VPG-VRER   | CartPole     | 179.21 | 0.79           | 123.11         | 200.00          |
| VPG-VRER   | Inverted Pendulum | 779.66 | 8.08            | 214.23         | 1000.00         |

Table 2 presents the performance (i.e., average total reward, standard error (SE), and percentiles) of all these algorithms over last 1000 training iterations for five different random seeds (same across all algorithms). In bold, the performances of using VRER are statistically significantly different from the baseline algorithm in each task.

In this section, we study the effects of selection constant $c$ and buffer size $B$ on the performance of VRER. Table 3 records the average rewards of TRPO-VRER with different buffer sizes. Overall, the performance of VRER is robust to the selection of buffer size $B$.

To better understand how the historical sample selection constant $c$ impacts on the performance of VRER, we conduct additional experiments by using the selection rule (20) (without using any approximation). In all experiments, we run the VPG-VRER algorithm on the CartPole for demonstration. Figure 4(b) shows the number of iterations required by VPG-VRER to solve CartPole problem (i.e., the average reward is greater than 195 over 100 consecutive iterations) as...
Table 3: Sensitivity of buffer sizes. The average reward of TRPO over last 1000 training iterations as a function of the buffer size for 10 random seeds. (mean ± 95% c.i.).

| Buffer Size | Acrobot | CartPole | Inverted Pendulum |
|-------------|---------|----------|-------------------|
| B = 200     | -106.72 ± 6.66 | 197.25 ± 1.00 | **732.92 ± 216.18** |
| B = 400     | -96.93 ± 2.80  | 199.69 ± 0.19 | 637.59 ± 207.50   |
| B = 800     | -153.06 ± 25.37| 195.97 ± 5.31 | 615.07 ± 132.89   |
| B = 1600    | -171.49 ± 62.99| 192.44 ± 6.51 | 684.87 ± 188.62   |

a function of average reward with different values of the selection constant $c$. The dashed vertical line indicates the 195 reward. In addition, we record the means and SEs of number of iterations required to solve the CartPole problem with different values of $c$ in Table 4. Overall, these empirical results indicate that the convergence behavior of the VPG-VRER algorithm is robust to the choice of $c$.

Table 4: Sensitivity of selection constants $c$.

| Sensitivity | c    | 1.2  | 1.5  | 2    | 4    | 8    |
|-------------|------|------|------|------|------|------|
| Number of Iterations | Mean  | 706.44 | 814.04 | 918.52 | 699.32 | 671.36 |
|              | SE    | 25.98 | 61.47 | 49.31 | 34.91 | 7.60  |

Figure 4: Sensitivity analyses for selection constant $c$. Result are smoothed by the moving average with a smoothing window width of 100.

Then, we study the impact of the value of selection constant $c$ on the reusing pattern of historical samples. Figure 4(a) shows the mean and 95% confidence band of the size of reuse set, i.e., $|U_k|$, when $c = 1.2, 1.5, 2, 4, 8$. It suggests that $|U_k|$ tends to linearly growing with the iteration $k$. Also, as $c$ increases, the number of reused samples tends to increase. The results obtained with $c = 2, 4, 8$ are close to each other. Basically, the value of $c$ determines the tolerance of the variance inflation of ILR policy gradient estimator $\widehat{\nabla \mu}_{ILR}^k$ relative to the baseline PG policy gradient estimator $\widehat{\nabla \mu}_{PG}^k$. When $c$ increases, more historical samples tend to be reused.

### 6.3 Gradient Variance Reduction

In this section, we present the empirical results to assess the performance of the proposed VRER in terms of reducing the policy gradient estimation variance. This study uses the selection rule (20) with the policy gradient variance estimators (24) and (25). The MLR gradient estimator (16) is used to update the policy parameters in Step 3(b) of Algorithm 1. We record the results of VPG with and without using VRER in Table 5. The original VPG without VRER leads to policy gradient estimators with high variability. By selectively reusing historical transition observations through VRER, the VPG shows a significant reduction in the estimation variance on policy gradient in all three examples. From the results in Table 5, we can see that the total variance of $\widehat{\nabla \mu}_{MLR}^k$ is consistently lower than that of $\widehat{\nabla \mu}_{PG}^k$ across different tasks over the training process (from episodes 1-500 to 1501-2000).
Table 5: The square root of total gradient variance differences in four different groups of episodes. (mean ± 95% c.i.):

| Task            | Episode       |
|-----------------|---------------|
|                 | 1-500         | 501-1000      | 1001-1500     | 1501-2000     |
| Acrobot         | 276.12 ± 10.12| 150.84 ± 6.26 | 112.91 ± 4.85| 95.54 ± 3.92 |
| CartPole        | 118.06 ± 8.10 | 168.06 ± 8.20 | 147.82 ± 6.91| 136.89 ± 12.17|
| Inverted Pendulum| 121.04 ± 23.25| 537.80 ± 52.69| 621.87 ± 93.53| 908.00 ± 89.35|

6.4 Convergence at Fixed Learning Rate

Theorem 2 shows that the PG-VRER can have asymptotic optimization convergence even with a fixed learning rate $\eta_k$ if the size of the reuse set $|U_k|$ increases to infinity as iteration $k$ grows. Here, we empirically investigate this property by comparing the convergence behaviors between VPG and VPG-VRER with the fixed learning rate $\eta = 0.005, 0.01$. To avoid the impact from approximation error of selection rule (31) and (32), in this section, we use the selection rule (20) with the policy gradient variance estimators (24) and (25). The MLR gradient estimator (16) is used for updating the policy parameters in Step 3(b) of Algorithm 1. We also fix the total number of iterations $K = 250$ and set the maximum episode length to 100 (it means that the maximum reward is 100).

The results of mean and 95% confidence interval (CI) of average reward are shown in Figure 5. They indicate that the VPG algorithm (blue line) diverges quickly after few iterations while the average reward of VPG-VRER method (orange line) continues to increase with iterations. Thus, the convergence behavior of VPG-VRER using a fixed stepsize is consistent with the theoretical conclusion in Theorem 2. As the fixed learning rate does not satisfy the Robbins–Monro conditions, the divergence of the VPG is within expectations; see Robbins and Monro [1951].

7 Conclusion

To guide real-time process control in low-data situations, we create a variance reduction experience replay approach to accelerate policy gradient optimization. The proposed selection rule guarantees the variance reduction in the policy gradient estimation through selectively reusing the most relevant historical samples and automatically allocating more weights to those samples that are more likely generated from the target distribution. The VRER approach has strong theoretical grounding since its selection rule is derived from a policy gradient estimation variance reduction criterion, capable of capturing the uncertainty induced by importance sampling. Practically speaking, it is surprisingly simple to apply VRER in most policy gradient optimization approaches as it does not require structural change of original algorithm. Practitioners can implement VRER by adding the selection procedure before training step. Both theoretical and empirical studies show that such variance reduction in gradient estimation can substantially increase the convergence...
rate of policy gradient methods and enhance the performance of state-of-the-art policy optimization algorithms, such as PPO and TRPO.

References

Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*(5):834–846, 1983. doi: 10.1109/TSMC.1983.6313077.

Jonathan Baxter and Peter L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research, 15*:319–350, 2001.

Shalabh Bhatnagar, Richard S Sutton, Mohammad Ghavamzadeh, and Mark Lee. Natural actor–critic algorithms. *Automatica, 45*(11):2471–2482, 2009.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. [http://pybullet.org](http://pybullet.org), 2021.

Thomas Degris, Martha White, and Richard S Sutton. Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*, 2012.

Morris H DeGroot and Mark J Schervish. *Probability and statistics*. Pearson, 2012.

J. Dong, M. B. Feng, and B. L. Nelson. Unbiased metamodeling via likelihood ratios. In *2018 Winter Simulation Conference (WSC)*, pages 1778–1789, Dec 2018. doi: 10.1109/WSC.2018.8632506.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, et al. Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures. In *International Conference on Machine Learning*, pages 1407–1416. PMLR, 2018.

William Fedus, Prajit Ramachandran, Rishabh Agarwal, Yoshua Bengio, Hugo Larochelle, Mark Rowland, and Will Dabney. Revisiting fundamentals of experience replay. In *International Conference on Machine Learning*, pages 3061–3071. PMLR, 2020.

Mingbin Feng and Jeremy Staum. Green simulation: Reusing the output of repeated experiments. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 27(4):23:1–23:28, October 2017. ISSN 1049-3301. doi: 10.1145/3129130.

Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Rafal Goebel and R Tyrrell Rockafellar. Local strong convexity and local lipschitz continuity of the gradient of convex functions. *Journal of Convex Analysis*, 15(2):263, 2008.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research, 5*(9), 2004.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018.

Randolph W Hall et al. *Handbook of healthcare system scheduling*. Springer, 2012.

Timothy Classen Hesterberg. *Advances in importance sampling*. PhD thesis, Citeseer, 1988.

Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics, 17*(2):295–311, 2008.

Prateek Jain and Purushottam Kar. Non-convex optimization for machine learning. *Found. Trends Mach. Learn., 10* (3-4):142–336, December 2017. ISSN 1935-8237. doi: 10.1561/2200000058.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL [http://arxiv.org/abs/1412.6980](http://arxiv.org/abs/1412.6980).
Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper/1999/file/6449f44a102fde848669bddd9eb6b76f-Paper.pdf

Harshat Kumar, Alec Koppel, and Alejandro Ribeiro. On the sample complexity of actor-critic method for reinforcement learning with function approximation. *arXiv preprint arXiv:1910.08412*, 2019.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR, 2019.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico*, 2016.

Long-Ji Lin. Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine learning*, 8(3-4):293–321, 1992.

Daoming Lyu, Qi Qi, Mohammad Ghavamzadeh, Hengshuai Yao, Tianbao Yang, and Bo Liu. Variance-reduced off-policy memory-efficient policy search. *arXiv preprint arXiv:2009.06548*, 2020.

Ashique Rupam Mahmood and Richard S Sutton. Off-policy learning based on weighted importance sampling with linear computational complexity. In *UAI*, pages 552–561. Citeseer, 2015.

Luca Martino, Victor Elvira, David Luengo, and Jukka Corander. An adaptive population importance sampler: Learning from uncertainty. *IEEE Transactions on Signal Processing*, 63(16):4422–4437, 2015.

Alberto Maria Metelli, Matteo Papini, Nico Montali, and Marcello Restelli. Importance sampling techniques for policy optimization. *J. Mach. Learn. Res.*, 21:141–1, 2020.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836.

Rémi Munos, Tom Stepleton, Anna Harutyunyan, and Marc G Bellemare. Safe and efficient off-policy reinforcement learning. *arXiv preprint arXiv:1606.02647*, 2016.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

Feng Niu, Benjamin Recht, Christopher Ré, and Stephen J Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. *arXiv preprint arXiv:1106.5730*, 2011.

Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

Shuang Qiu, Zhuorong Yang, Jieping Ye, and Zhaoran Wang. On finite-time convergence of actor-critic algorithm. *IEEE Journal on Selected Areas in Information Theory*, 2(2):652–664, 2021.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Herbert Robbins and David Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pages 233–257. Elsevier, 1971.

Kevin Scaman and Cedric Malherbe. Robustness analysis of non-convex stochastic gradient descent using biased expectations. *Advances in Neural Information Processing Systems*, 33, 2020.

Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *CoRR*, abs/1511.05952, 2016.
Matthew Schlegel, Wesley Chung, Daniel Graves, Jian Qian, and Martha White. Importance Resampling for Off-Policy Prediction. Curran Associates Inc., Red Hook, NY, USA, 2019.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International conference on machine learning, pages 1889–1897. PMLR, 2015a.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438, 2015b.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017.

Othmane Sebbouh, Robert M Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. In Conference on Learning Theory, pages 3935–3971. PMLR, 2021.

Richard S Sutton. Generalization in reinforcement learning: Successful examples using sparse coarse coding. In D. Touretzky, M.C. Mozer, and M. Hasselmo, editors, Advances in Neural Information Processing Systems, volume 8. MIT Press, 1995.

Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. A Bradford Book, Cambridge, MA, USA, 2018. ISBN 0262039249.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Proceedings of the 12th International Conference on Neural Information Processing Systems, NIPS’99, page 1057–1063, Cambridge, MA, USA, 1999a. MIT Press.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, Advances in Neural Information Processing Systems, volume 12. MIT Press, 1999b.

Eric Veach and Leonidas J Guibas. Optimally combining sampling techniques for monte carlo rendering. In Proceedings of the 22nd annual conference on Computer graphics and interactive techniques, pages 419–428, 1995.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Rémi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. In 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net, 2017.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696.

Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. In International Conference on Learning Representations, 2018.

Yue Frank Wu, Weitong ZHANG, Pan Xu, and Quanquan Gu. A finite-time analysis of two time-scale actor-critic methods. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 17617–17628. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/cc9b3c69b56df284846bf2432f1cba90-Paper.pdf.

Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for (natural) actor-critic algorithms. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 4358–4369. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/2eb24a664f5e9c18f407b2f9c73e821-Paper.pdf.

Mengjiao Yang, Ofir Nachum, Bo Dai, Lihong Li, and Dale Schuurmans. Off-policy evaluation via the regularized lagrangian. Advances in Neural Information Processing Systems, 33:6551–6561, 2020.

Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. ACM Computing Surveys (CSUR), 55(1):1–36, 2021.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Convergence and iteration complexity of policy gradient method for infinite-horizon reinforcement learning. In 2019 IEEE 58th Conference on Decision and Control (CDC), pages 7415–7422. IEEE, 2019.

Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. SIAM Journal on Control and Optimization, 58(6):3586–3612, 2020. doi: 10.1137/19M1288012. URL https://doi.org/10.1137/19M1288012.
Hua Zheng, Jiahao Zhu, Wei Xie, and Judy Zhong. Reinforcement learning assisted oxygen therapy for covid-19 patients under intensive care. *BMC medical informatics and decision making*, 21(1):1–8, 2021.

Hua Zheng, Wei Xie, Ilya O Ryzhov, and Dongming Xie. Policy optimization in bayesian network hybrid models of biomanufacturing processes. *INFORMS Journal on Computing*, 2022. In Press.

Fan Zhou and Guojing Cong. On the convergence properties of a k-step averaging stochastic gradient descent algorithm for nonconvex optimization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 3219–3227. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/447.

Shaofang Zou, Tengyu Xu, and Yingbin Liang. *Finite-Sample Analysis for SARSA with Linear Function Approximation*. Curran Associates Inc., Red Hook, NY, USA, 2019.
Appendices

Appendix A  Experiment Details

In this appendix, we report the hyperparameter values used in the experimental evaluation and some additional results and experiments. In all side-by-side comparison experiments, each pair of baseline algorithm and VRER assisted algorithm were run under same hyperparameter settings. To reproduce the results and check out the implementation details, please visit our open-sourced library: https://github.com/zhenghuazx/vrer_policy_gradient.

A.1 Hyperparameters

- **Policy architecture**: For discrete action, we adopted a softmax policy: Categorical distributions $\text{Cart}(p_i)$, where the probabilities $p_i = \pi_\theta(a_i|s)$ are modeled by softmax function (for a $m$-dimensional action)

  \[
  \pi_\theta(a_i|s) = \frac{\exp\left(\phi_\theta(s, a_i)\right)}{\sum_{i=1}^{m} \exp\left(\phi_\theta(s, a_i')\right)}. \tag{35}
  \]

  The mean function $\phi_\theta$ is a 2–layers multilayer perceptron (MLP) (32, 32) with bias (activation functions: ReLU for hidden–layers, linear for output layer).

  For continuous actions, we adopted a Gaussian policy: normal distribution $\mathcal{N}(a|\phi_\theta(s), I)$, where the mean $\phi_\theta$ is a 2–layers multilayer perceptron (MLP) (32, 32) with bias (activation functions: tanh for hidden–layers, linear for output layer the variance is state–independent and parametrized identity matrix).

- **Number of macro-replications**: 5 (95% c.i.)

- **Seeds**: 2022, 2023, 2024, 2025, 2026 (PPO, PPO-VRER, TRPO and TRPO-VRER); 2021, 2022, 2023, 2024, 2025 (VPG and VPG-VRER);

- **Policy initialization**: glorot uniform initialization for VPG(-VRER) and orthogonal initialization for PPO(-VRER) and TRPO(-VRER)

The VRER related hyperparameters include selection constant $c$, the buffer size $B$, and the number of sampled observations per iteration $n_0$.

The PPO(-VRER) hyperparameters are presented in Table 6. The hyperparameter “mini batches” represents the number of mini-batch; “batch size” represents the number of transitions per iteration; “entropy coef” represents the entropy coefficient used for entropy loss calculation; and “Lambda” represents the lambda for the general advantage estimation. The “PPO iterations” hyperparameter ($K_{off}$) represents the maximum number of actor-critic optimizations steps per train step.

|                   | CartPole | Acrobot | Inverted Pendulum |
|-------------------|----------|---------|------------------|
| step size         | 0.0003   | 0.0003  | 0.0003           |
| batch size        | 512      | 512     | 512              |
| clip norm         | 0.2      | 0.2     | 0.2              |
| entropy coef      | 0.01     | 0.01    | 0.01             |
| Lambda            | 0.95     | 0.95    | 0.95             |
| mini batches      | 128      | 128     | 128              |
| PPO iterations    | 4        | 4       | 4                |
| buffer size       | 400      | 200     | 400              |
| $n_0$             | 12       | 12      | 12               |

The TRPO(-VRER) hyperparameters are presented in Table 7. The hyperparameter “cg iterations” represents the gradient conjugation maximum number of iterations per train step; “entropy coef” represents the entropy coefficient used for entropy loss calculation; the “actor/critic iterations” hyperparameter ($K_{off}$) represents the maximum number of actor/critic optimizations steps per train step; and “clip norm” represents the the surrogate clipping coefficient.

As the VPG is an episode-based algorithm, we use the selection rule (31) with selection constant $c = 1.5$ across all experiments. Moveover, in VPG-VRER related experiment runs, we did not use the buffer and thus no buffer size and set $n_0 = n$ number of randomly sampled observations per iteration; see Table 8. In practice, we recommend users tune the selection rule constant together with the learning rate (step size).
A.2 Runtime Analysis

To better understand the computational complexity of VRER, we present the summary table of runtime in Table 9.

A.3 Control Tasks

All control systems are simulated based on the simulation model provided in the Open AI library [Brockman et al., 2016] and PyBullet library [Coumans and Bai, 2021].

CartPole is a pendulum with a center of gravity above its pivot point; see more detailed description in [Barto et al., 1983]. The goal is to keep the cart-pole balanced by applying appropriate forces to a pivot point under the center of mass and prevent the pendulum from falling over. At any time $t$, the state vector $s_t$ for this system is a four dimensional vector having components $x_t, \dot{x}_t, \alpha_t, \dot{\alpha}_t$ representing position of cart, velocity of cart, angle of pole, and rotation rate of pole respectively. The action $a_t$ consists of a force $\vec{F}_t$: left (0) and right (1). Each episode or simulation run terminates if (1) the pole angle is more than 12 degree from the vertical axis; or (2) the cart position is more than 2.4 unit from the centre; or (3) the episode length is greater than 200. The agent receives a reward of 1 for each step taken including the termination step. The problem is considered “solved” if the average reward is greater than 195 over 100 consecutive iterations.

The Acrobot environment is constructed based on Sutton’s work [Sutton, 1995]. The system consists of two links connected linearly to form a chain, with one end of the chain fixed. The joint between the two links is actuated. The goal is to apply torques on the actuated joint to swing the free end of the linear chain above a given height while starting from the initial state of hanging downwards. This problem has 3 discrete actions and 6 continuous states. The goal is to have the free end reaching a designated target height in as few steps as possible, and as such all steps that do not reach the goal incur a reward of -1. Achieving the target height results in termination with a reward of 0.

The inverted pendulum system is an example commonly found in control system textbooks and research literature. The objective of the control system is to balance the inverted pendulum by applying a force to the cart that the pendulum is attached to. The problem has one continuous action valued between -1 and 1 and five continuous states. The maximum episode length is 1000; see details in PyBullet library [Coumans and Bai, 2021].

A.4 Confidence Intervals

At each $k$-th iteration, based on the results obtained from $m_k$ macro-replications, we represent the estimation uncertainty of the output, denoted by $x_k$, by using the 95% confidence bands, $(\bar{x}_k^m - z \cdot SE_k, \bar{x}_k^m + z \cdot SE_k)$, with $z = 1.96$.

| Table 7: Hyperparamters for TRPO and TRPO-VERE |
|-----------------------------------------------|
| CartPole | Acrobot | Inverted Pendulum |
| step size | 0.0003 | 0.0003 | 0.001 |
| batch size | 512 | 512 | 512 |
| clip norm | 0.2 | 0.2 | 0.2 |
| entropy coef | 0.01 | 0.01 | 0.01 |
| mini batches | 128 | 128 | 128 |
| actor iterations | 10 | 10 | 10 |
| critic iterations | 3 | 3 | 3 |
| cg iterations | 10 | 10 | 10 |
| buffer size | 400 | 400 | 200 |
| $n_0$ | 12 | 12 | 12 |

| Table 8: Hyperparamters for VPG and VPG-VERE |
|-----------------------------------------------|
| CartPole | Acrobot | Inverted Pendulum |
| step size | 0.005 | 0.0003 | 0.0007 |
| $c$ | 1.5 | 1.5 | 1.5 |
| buffer size | - | - | - |
| $n_0$ | - | - | - |
Zheng et al.: Variance Reduction Based Experience Replay

Table 9: Runtime (in minute) of 2000 episodes with a fixed buffer size $B = 100$.

| Algorithm | Acrobat | CartPole | Inverted Pendulum |
|-----------|---------|----------|-------------------|
|           | Mean SE | Mean SE  | Mean SE           |
| PPO       | 218.29  | 12.40    | 278.62            |
| PPOVRER   | 406.82  | 64.09    | 753.34            |
| TRPO      | 292.98  | 10.64    | 353.08            |
| TRPOVRER  | 359.53  | 28.62    | 627.71            |

sample mean $\mu_k = \frac{1}{m_r} \sum_{\ell=1}^{m_r} x_{k,\ell}$, and sample standard error (SE), i.e.,

$$SE_k = \frac{1}{\sqrt{(m_r - 1)m_r}} \sum_{\ell=1}^{m_r} (x_{k,\ell} - \mu_k)^2.$$  

In our empirical study, this quantity can be the expected total reward, the size of reuse set $|\mathcal{U}_k|$, or the trace of policy gradient estimation variance matrix $\text{Tr}(\text{Var} \hat{\nabla}_\mu(\theta_k))$. For example, in the convergence plots, we record the averaged total reward $x_{k,\ell} = \frac{1}{n} \sum_{j=1}^{n} R(k,j)$; in the figure of reusing patterns (i.e., Figure 4(a)), we set $x_{k,\ell} = |\mathcal{U}_{k,\ell}|$, the size of reuse set obtained in the $\ell$-th macro-replication; and lastly we set $x_{k,\ell}$ to be the square root of the total variance or the trace of variance matrix of gradient estimator at the $k$-th iteration and the $\ell$-th macro-replication for Figure 5.

Appendix B Theoretical Analysis and Proofs

B.1 Proof of Lemma 3

**Lemma 3** Conditional on $\mathcal{F}_k$, the proposed MLR policy gradient estimators are unbiased:

$$\mathbb{E} \left[ \hat{\nabla}_\mu^{\text{MLR}} \Big| \mathcal{F}_k \right] = \mathbb{E} \left[ g_k(x) \Big| \theta_k \right] = \nabla \mu(\theta_k).$$

**Proof.** We show that the MLR policy gradient estimator listed in (16) is unbiased,

$$\mathbb{E} \left[ \hat{\nabla}_\mu^{\text{MLR}} \Big| \mathcal{F}_k \right] = \int \frac{1}{|\mathcal{U}_k|} \sum_{\theta_k \in \mathcal{U}_k} \sum_{\theta_k' \in \mathcal{U}_k} \rho_{\theta_k'}(x) g_k(x) \rho_{\theta_k'}(x) dx$$

$$= \int \rho_{\theta_k}(x) g_k(x) dx$$

$$= \mathbb{E}_{x \sim \rho_{\theta_k}} [g_k(x)].$$

\[\Box\]

B.2 Proof of Proposition 1

Before proving Lemma 4, we first introduce some additional notations. Denote the $d$-dimensional policy parameter by $\theta = (\theta^{(1)}, \ldots, \theta^{(d)})^T \in \mathbb{R}^d$. Remember that $g_k(x)$ denote the scenario-based policy gradient estimate at $\theta_k$, and it is a $d$-dimensional vector

$$g_k(x) = (g_1, \ldots, g_d)^T$$

where $g_i = \sum_{t=1}^{H} R(\tau) \frac{\partial}{\partial \theta_k} \log \pi_{\theta_k}(a_t | s_t)$ for finite horizon and $g_i = A_{\pi}(s, a) \frac{\partial}{\partial \theta_k} \log \pi_{\theta_k}(a_t | s_t)$ for infinite horizon MDP.

**Proposition 1** Conditional on $\mathcal{F}_k$, the total variance of the MLR policy gradient estimator is smaller and equal to that of the average ILR policy gradient estimator,

$$\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_\mu^{\text{MLR}} \Big| \mathcal{F}_k \right] \right) \leq \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_\mu^{\text{ILR}} \Big| \mathcal{F}_k \right] \right).$$

26
Proof. Let \( \{ \tilde{x}^{(i,j)} \sim \rho_{\theta_i}(x) : \theta_i \in \mathcal{U}_k \text{ and } j = 1, 2, \ldots, n \} \) be i.i.d. samples from the mixture distribution \( \frac{1}{|\mathcal{U}_k|} \sum_{\theta_i \in \mathcal{U}_k} \rho_{\theta_i}(x) \) and obtain the policy gradient estimator,

\[
\nabla_{\mu_k}^{Mix} = \frac{1}{|\mathcal{U}_k|} \sum_{\theta_i \in \mathcal{U}_k} \frac{1}{n} \sum_{j=1}^{n} \left[ \hat{f}_k \left( \tilde{x}^{(i,j)} \right) g_k \left( \tilde{x}^{(i,j)} \right) \right].
\]

The estimators \( \nabla_{\mu_k}^{MLR} \) and \( \nabla_{\mu_k}^{Mix} \) are similar: the former is a stratified-sampling version of the latter without considering the random sampling of parameters from the reuse set \( \mathcal{U}_k \). Clearly, \( \nabla_{\mu_k}^{Mix} \) is unbiased. Therefore,

\[
\mathbb{E} \left[ \nabla_{\mu_k}^{MLR} \right] = \mathbb{E} \left[ \nabla_{\mu_k}^{Mix} \right] = \mathbb{E}_{\mathbf{x} \sim \rho_{\theta_k}} \left[ g_k(x) \right],
\]

which equals \( \nabla_{\mu}(\theta) \) by Lemma 3. By plugging in (15), we have

\[
\text{Var} \left[ \nabla_{\mu_k}^{MLR} \mid \mathcal{F}_k \right] - \text{Var} \left[ \nabla_{\mu_k}^{Mix} \mid \mathcal{F}_k \right]
\]

\[
\overset{(*)}{=} \frac{1}{n} \text{Var} \left[ \frac{1}{|\mathcal{U}_k|} \sum_{\theta_i \in \mathcal{U}_k} \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)} g_k(x) \mid \mathcal{F}_k \right]
\]

\[
= \frac{1}{n} \int \left[ \frac{1}{|\mathcal{U}_k|} \sum_{\theta_i \in \mathcal{U}_k} \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)} g_k(x) \right]^2 d\mathbf{x}
\]

\[
- \frac{1}{n} \int \left[ \frac{1}{|\mathcal{U}_k|} \sum_{\theta_i \in \mathcal{U}_k} \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)} g_k(x) \right] g_k(x)^\top \left[ \frac{1}{|\mathcal{U}_k|} \sum_{\theta_i \in \mathcal{U}_k} \rho_{\theta_i}(x) \right] d\mathbf{x}
\]

\[
= \frac{1}{n} \int \left[ \rho_{\theta_k}(x) g_k(x) g_k(x)^\top \right] \left[ \frac{1}{|\mathcal{U}_k|} \sum_{\theta_i \in \mathcal{U}_k} \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)} \right] d\mathbf{x}
\]

\[
\overset{(36)}{=}
\]

Step (*) holds because of the conditional independence of samples \( x^{(i,j)} \sim \rho_{\theta_i}(x) \) with \( \theta_i \in \mathcal{U}_k \) and \( j = 1, 2, \ldots, n \).

Notice that \( g_k(x) g_k(x)^\top = \begin{pmatrix} g_1^2 & g_1 g_2 & \cdots & g_1 g_d \\ g_2 g_1 & g_2^2 & \cdots & g_2 g_d \\ \vdots & \vdots & \ddots & \vdots \\ g_d g_1 & g_d g_2 & \cdots & g_d^2 \end{pmatrix} \) and also \( \text{Tr} \left( g_k(x) g_k(x)^\top \right) = \sum_{i=1}^{d} g_i^2 = \| g_k(x) \|^2 \). Then the trace of covariance matrix (36) becomes

\[
\text{Tr} \left( \text{Var} \left[ \nabla_{\mu_k}^{MLR} \mid \mathcal{F}_k \right] \right) - \text{Tr} \left( \left[ \nabla_{\mu_k}^{Mix} \mid \mathcal{F}_k \right] \right)
\]

\[
= \frac{1}{n} \int \left( \rho_{\theta_k}(x) \| g_k(x) \|^2 \right)^2 \left[ \frac{1}{|\mathcal{U}_k|} \sum_{\theta_i \in \mathcal{U}_k} \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)} \right] d\mathbf{x}
\]

\[
\geq 0
\]

which holds due to the well-known inequality between arithmetic and harmonic mean, i.e.,

\[
\frac{1}{|\mathcal{U}_k|} \sum_{\theta_i \in \mathcal{U}_k} \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)} \geq \frac{1}{|\mathcal{U}_k|} \sum_{\theta_i \in \mathcal{U}_k} \rho_{\theta_i}(x).
\]

Next, we observe that \( \nabla_{\mu_k}^{MLR} \) is a stratified-sampling version of \( \nabla_{\mu_k}^{Mix} \), with \( n \) equal number of samples allocated to each behavioral policy in \( \mathcal{U}_k \). This means equally weighted strata: the sampling likelihood of the \( i \)-th stratum is \( \rho_{\theta_i}(\cdot) \). Therefore, \( \text{Tr} \left( \text{Var} \left[ \nabla_{\mu_k}^{MLR} \mid \mathcal{F}_k \right] \right) \leq \text{Tr} \left( \text{Var} \left[ \nabla_{\mu_k}^{Mix} \mid \mathcal{F}_k \right] \right) \). Combining (37) leads to the desired result

\[
\text{Tr} \left( \text{Var} \left[ \nabla_{\mu_k}^{MLR} \mid \mathcal{F}_k \right] \right) \leq \text{Tr} \left( \text{Var} \left[ \nabla_{\mu_k}^{Mix} \mid \mathcal{F}_k \right] \right).
\]

□
B.3 Proof of Theorems

**Theorem 1:** At the $k$-th iteration with the target distribution $\rho_{\theta_k}(x)$, the reuse set $U_k$ is created to include the distributions, i.e., $\rho_{\theta_k}(x)$ with $\theta_k \in F_k$, whose ILR policy gradient estimator’s total variance is no greater than $c$ times the total variance of the classical PG estimator for some constant $c > 1$. Mathematically,

$$
\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{MLR}} \left| F_k \right. \right] \right) \leq c \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{PG}} \left| \theta_k \right. \right] \right).
$$

Then, based on such reuse set $U_k$, the total variance of the MLR/ILR policy gradient estimators (15) and (16) are no greater than $\frac{c}{|U_k|}$ times the total variance of the PG estimator:

$$
\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{MLR}} \left| F_k \right. \right] \right) \leq \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{ILR}} \left| F_k \right. \right] \right) \leq \frac{c}{|U_k|} \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{PG}} \left| \theta_k \right. \right] \right).
$$

Moreover,

$$
\mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_k}^{\text{MLR}} \right\|^2 \left| F_k \right. \right] \leq \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_k}^{\text{ILR}} \right\|^2 \left| F_k \right. \right] \leq \frac{c}{|U_k|} \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_k}^{\text{PG}} \right\|^2 \left| \theta_k \right. \right] + \left( 1 - \frac{c}{|U_k|} \right) \left\| \nabla \mu(\theta_k) \right\|^2.
$$

**Proof.** Conditional on $F_k$, by Assumption A.3, the observations $x^{(i,j)}$ for any $\theta_k \in U_k$ and $j = 1, 2, \ldots, n$ are independent. Thus, we have

$$
\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{MLR}} \left| F_k \right. \right] \right) \leq \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{ILR}} \left| F_k \right. \right] \right) \leq \frac{1}{|U_k|^2} \sum_{\theta_k \in U_k} \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{ILR}} \left| F_k \right. \right] \right) \leq \frac{c}{|U_k|^2} \sum_{\theta_k \in U_k} \text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{PG}} \left| \theta_k \right. \right] \right)
$$

(38)

where inequality (38) follows Proposition 1 and (39) holds because of selection rule (20). Thus, (21) is proved. Then, we have

$$
\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{MLR}} \left| F_k \right. \right] \right) = \text{Tr} \left( \mathbb{E} \left[ \left( \hat{\nabla}_{\mu_k}^{\text{MLR}} \right)^\top \left| F_k \right. \right] - \mathbb{E} \left[ \hat{\nabla}_{\mu_k}^{\text{MLR}} \left| F_k \right. \right] \mathbb{E} \left[ \hat{\nabla}_{\mu_k}^{\text{MLR}} \left| F_k \right. \right]^\top \right) = \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_k}^{\text{MLR}} \right\|^2 \left| F_k \right. \right] - \mathbb{E} \left[ \hat{\nabla}_{\mu_k}^{\text{MLR}} \left| F_k \right. \right]^2 = \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_k}^{\text{MLR}} \right\|^2 \left| F_k \right. \right] - \left\| \nabla \mu(\theta_k) \right\|^2.
$$

(41)

where (41) follows because the MLR policy gradient estimator $\hat{\nabla}_{\mu_k}^{\text{MLR}}$ is unbiased. Similarly, we have

$$
\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{PG}} \left| \theta_k \right. \right] \right) = \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_k}^{\text{PG}} \right\|^2 \left| \theta_k \right. \right] - \left\| \nabla \mu(\theta_k) \right\|^2.
$$

(42)

$$
\text{Tr} \left( \text{Var} \left[ \hat{\nabla}_{\mu_k}^{\text{ILR}} \left| \theta_k \right. \right] \right) = \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_k}^{\text{ILR}} \right\|^2 \left| \theta_k \right. \right] - \left\| \nabla \mu(\theta_k) \right\|^2.
$$

(43)

where the equalities hold due to the unbiasedness of $\hat{\nabla}_{\mu_k}^{\text{PG}}$ and $\hat{\nabla}_{\mu_k}^{\text{ILR}}$. Then with Proposition 1 and (40)–(42), we can show

$$
\mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_k}^{\text{MLR}} \right\|^2 \left| F_k \right. \right] \leq \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_k}^{\text{ILR}} \right\|^2 \left| F_k \right. \right] \leq \frac{c}{|U_k|^2} \mathbb{E} \left[ \left\| \hat{\nabla}_{\mu_k}^{\text{PG}} \right\|^2 \left| \theta_k \right. \right] + \left( 1 - \frac{c}{|U_k|^2} \right) \left\| \nabla \mu(\theta_k) \right\|^2.
$$

(44)

It completes the proof. □
B.4 Proof of Lemma [4]

Lemma [4] Suppose that Assumptions [A.1], [A.2], and [A.3] hold. For a clipped likelihood ratio based policy gradient estimator $\widehat{\nabla \mu_k}^{\text{CLR}}$, with a finite clipping threshold $U_f > 1$, the bias and the variance of the CLR estimator can be upper bounded as

$$\text{Tr} \left( \text{Var} \left[ \widehat{\nabla \mu_k}^{\text{CLR}} \left| F_k \right. \right] \right) \leq \text{Tr} \left( \text{Var} \left[ \widehat{\nabla \mu_k} \left| F_k \right. \right] \right) + \frac{1}{\lceil n/|U_k^2| \rceil} \left\lVert \nabla \mu (\theta_k) \right\rVert^2.$$

where $Z(k)$ is the bound of stochastic gradient estimate. Remember $f_{i,k}(x) = \frac{\rho_{\theta_k}(x)}{\rho_{\theta_i}(x)}$. Let $f_{i,k}(x) = \min \left( \frac{\rho_{\theta_k}(x)}{\rho_{\theta_i}(x)}, U_f \right)$. Then we have

$$\text{Tr} \left( \text{Var} \left[ \widehat{\nabla \mu_k}^{\text{CLR}} \left| F_k \right. \right] \right) = \text{Tr} \left( \text{Var} \left[ \frac{1}{|U_k|} \sum_{i \in U_k} \sum_{j=1}^n \left( \frac{\rho_{\theta_k}(x(i,j))}{\rho_{\theta_i}(x(i,j))} - \min \left( U_f, \frac{\rho_{\theta_k}(x(i,j))}{\rho_{\theta_i}(x(i,j))} \right) \right) g_k(x(i,j)) \right] \right) \leq \frac{1}{|U_k|} \sum_{i \in U_k} \sum_{j=1}^n \text{E}_i \left[ \left( f_{i,k}(x) - f_{i,k}(x) \right) g_k(x) \right] \leq M \frac{|U_k|}{|U_k|} \sum_{i \in U_k} \text{E}_i \left[ \left\lVert f_{i,k}(x) - U_f \right\rVert \| f_{i,k}(x), U_f \| \right] \leq M \frac{|U_k|}{|U_k|} \sum_{i \in U_k} \text{E}_i \left[ \left\lVert f_{i,k}(x) - U_f \right\rVert \right] \leq M \frac{|U_k|}{|U_k|} \sum_{i \in U_k} \text{E}_i \left[ \left\lVert f_{i,k}(x) \right\rVert \right]$$(45)

where step (45) follows triangle inequality, step (46) follows from observing that the weight difference is either zero or $U_f$ based on whether $\frac{\rho_{\theta_k}(x)}{\rho_{\theta_i}(x)} > U_f$, and step (47) follows by applying Hölder's inequality and Lemma [2]. Step (48) is obtained by observing that under the indicator function we have that $f_{i,k}(x) > U_f$.

Then, we prove the bound of total variance,

$$\text{Tr} \left( \text{Var} \left[ \widehat{\nabla \mu_k}^{\text{CLR}} \left| F_k \right. \right] \right) = \text{Tr} \left( \text{Var} \left[ \frac{1}{|U_k|} \sum_{i \in U_k} \sum_{j=1}^n \min \left( \frac{\rho_{\theta_k}(x(i,j))}{\rho_{\theta_i}(x(i,j))}, U_f \right) g_k(x(i,j)) \right] \right)$$

$$= \frac{1}{n/|U_k^2|} \sum_{i \in U_k} \sum_{j=1}^n \left\lVert \text{E}_i \left[ \left( \frac{\rho_{\theta_k}(x)}{\rho_{\theta_i}(x)} \right) g_k(x) \right] \right\rVert^2$$

$$\leq \frac{1}{n/|U_k^2|} \left( \text{E}_i \left[ f_{i,k}(x) \| g_k(x) \|^2 \right] - \| \text{E}_i \left[ f_{i,k}(x) g_k(x) \right] \| + \| \text{E}_i \left[ f_{i,k}(x) g_k(x) \right] \| \right)^2$$

$$= \text{Tr} \left( \text{Var} \left[ \widehat{\nabla \mu_k}^{\text{CLR}} \left| F_k \right. \right] \right) + \frac{1}{n/|U_k^2|} \left\lVert \text{E}_k \left[ g_k(x) \right] \right\rVert^2.$$
Then the unbiasedness $\|E_k[g_k(x)]\| = \|\nabla \mu(\theta_k)\|$ gives the result stated in Lemma 4.

B.5 Proof of Corollary 1

**Corollary 1** At the $k$-th iteration with the target distribution $\rho(x)$, we apply the selection rule from Theorem 4. Then, based on such reuse set $U_k$, we have the following bound

$$
E \left[ \left\| \nabla \mu_k^{CLR} \right\|^2 | F_k \right] \leq \frac{c}{|U_k|} E \left[ \left\| \nabla \mu_k^{PG} \right\|^2 \right] + \left( 8 + \frac{1}{n|U_k|} - \frac{c}{|U_k|} \right) \|\nabla \mu(\theta_k)\| + 8\sigma^2.
$$

**Proof.** For CLR policy gradient estimator, we have

$$
\begin{align*}
\text{Tr} \left( \text{Var} \left[ \nabla \mu_k^{CLR} | F_k \right] \right) & \leq \text{Tr} \left( \text{Var} \left[ \nabla \mu_k^{ILR} | F_k \right] \right) \\
& \leq \frac{c}{|U_k|} \text{Tr} \left( \text{Var} \left[ \nabla \mu_k^{PG} | \theta_k \right] \right)
\end{align*}
$$

where (49) follows by applying Lemma 4 and (50) follows by applying (39)-(40). Then

$$
\begin{align*}
\text{Tr} \left( \text{Var} \left[ \nabla \mu_k^{CLR} | F_k \right] \right) &= \text{Tr} \left( \text{Var} \left[ \nabla \mu_k^{CLR} | \theta_k \right] \right) - \text{Tr} \left( \text{Var} \left[ \nabla \mu_k^{CLR} | F_k \right] \right) \\
& = \text{Var} \left[ \nabla \mu_k^{CLR} | \theta_k \right]
\end{align*}
$$

Rearranging the equality (51) leads to

$$
\begin{align*}
\text{Var} \left[ \nabla \mu_k^{CLR} | \theta_k \right] & \leq 8\|\nabla \mu(\theta_k)\| + 8\sigma^2 + \frac{c}{|U_k|} \text{Tr} \left( \text{Var} \left[ \nabla \mu_k^{PG} | \theta_k \right] \right) + \frac{1}{n|U_k|} \|\nabla \mu(\theta_k)\| \\
& = 8\|\nabla \mu(\theta_k)\| + 8\sigma^2 + \frac{c}{|U_k|} \left( \text{Var} \left[ \nabla \mu_k^{PG} | \theta_k \right] \right) + \frac{1}{n|U_k|} \|\nabla \mu(\theta_k)\| \\
& = \frac{c}{|U_k|} \text{Var} \left[ \nabla \mu_k^{PG} | \theta_k \right] + \left( 8 + \frac{1}{n|U_k|} - \frac{c}{|U_k|} \right) \|\nabla \mu(\theta_k)\| + 8\sigma^2
\end{align*}
$$

where (52) follows by applying Lemma 4 and (53) follows by applying (42).

B.6 Proof of Proposition 2

**Lemma 7.** [DeGroot and Schervish, 2012, Theorem 8.7.1] The covariance estimator

$$
\hat{\text{Var}}(X) = \frac{1}{n-1}(X_i - \bar{X})(X_i - \bar{X})
$$

is an unbiased estimator of the covariance $\text{Cov}(X, X) = \text{Var}(X)$ where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ and independent samples $\{X_i : i = 1, \ldots, n\}$ have same mean $E[X_i] = E[X]$ for all $i = 1, 2, \ldots, n$.

**Proposition 2.** Given $F_k$, the policy gradient variance estimators $\hat{\text{Var}} \left[ \nabla \mu_k^{PG} | \theta_k \right]$ in (24) and $\hat{\text{Var}} \left[ \nabla \mu_{i,k}^{ILR} | F_k \right]$ in (25) are unbiased.
Proof. We have derived the ILR gradient estimator and its sample covariance matrix
\[ \hat{\nabla} \mu_{k}^{ILR} = \frac{1}{n} \sum_{j=1}^{n} \hat{\nabla} \mu_{k,j}^{ILR} \quad \text{with} \quad \hat{\nabla} \mu_{k,j}^{ILR} = \frac{1}{|U_k|} \sum_{\theta \in U_k} f_k \left( x^{(i,j)} \right) g_k \left( x^{(i,j)} \right), \]
and
\[ \text{Var} \left[ \hat{\nabla} \mu_{k}^{ILR} \big| F_k \right] = \frac{1}{n-1} \sum_{j=1}^{n} \hat{\sigma}_{i,j} \hat{\sigma}_{i,j}^T \]
where \( \hat{\sigma}_{i,j} \) is the unbiasedness estimator.

B.7 Proof of Theorem 2 and 4

Lemma 8. Suppose that Assumption A.1 holds. For any feasible policy parameters \( \theta \in \mathbb{R}^d \), we have \( \mathbb{E}[\|\hat{\nabla} \mu^{PG} (\theta)\|^2] \leq 2\|\nabla \mu(\theta)\|^2 + 2\sigma^2 \).

Proof. We proceed by using Assumption A.1
\[ \mathbb{E} \left[ \|\hat{\nabla} \mu^{PG} (\theta)\|^2 \right] = \mathbb{E} \left[ \|\hat{\nabla} \mu^{PG} (\theta) - \nabla \mu(\theta) + \nabla \mu(\theta)\|^2 \right] \leq \mathbb{E} \left[ \|\hat{\nabla} \mu^{PG} (\theta) - \nabla \mu(\theta)\|^2 + \|\nabla \mu(\theta)\|^2 \right] \leq 2\mathbb{E} \left[ \|\hat{\nabla} \mu^{PG} (\theta) - \nabla \mu(\theta)\|^2 \right] + 2\sigma^2 \]
where (54) follows Cauchy–Schwarz inequality and (55) is obtained by applying Assumption A.1 This completes the proof.

Theorem 2 Suppose Assumptions A.1, A.2 and A.3 hold. Let \( \eta_k \) denote the learning rate used in the \( k \)-th iteration and \( U_\mu \) is the bound of objective. If we use the ILR and MLR policy gradient estimators in Algorithm 1, we have
\[ \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \|\nabla \mu(\theta_k)\|^2 \right] \leq \frac{\lambda U_\mu}{K} - \frac{2\mu_1}{\eta_1} + 4\epsilon L_\sigma^2 \sum_{k=1}^{K} \frac{m_k}{K}, \]
where \( \sigma \) and \( L \) are the bounds defined in Assumptions A.1 and Lemma 7 and \( c \) is the selection constant defined in Theorems 7.

Proof. Lemma 1 implies the L-Lipschitz property, \( |\mu(\theta_k) - \mu(\theta_{k+1})| \leq L_\theta |\theta_{k+1} - \theta_k| \quad \text{(Nesterov [2003], Lemma 1.2.3).} \)
It gives \( \mu(\theta_k) - \mu(\theta_{k+1}) \leq \langle \nabla \mu(\theta_k), \theta_{k+1} - \theta_k \rangle + \|\nabla \mu(\theta_k)\|^2. \)

By applying the policy parameter update implemented in the proposed algorithm \( \theta_{k+1} = \theta_k + \eta_k \hat{\nabla} \mu_{k}^{ILR} \), we have
\[ \mu(\theta_k) - \mu(\theta_{k+1}) \leq - \langle \nabla \mu(\theta_k), \eta_k \hat{\nabla} \mu_{k}^{ILR} (\theta_k) \rangle + \|\theta_{k+1} - \theta_k\|^2. \]

Taking the expectation of \( \langle \nabla \mu(\theta_k), \eta_k \hat{\nabla} \mu_{k}^{ILR} (\theta_k) \rangle \) conditioning on \( \mathcal{U}_k \), we have that
\[ \mathbb{E} \left[ \langle \nabla \mu(\theta_k), \eta_k \hat{\nabla} \mu_{k}^{ILR} (\theta_k) \rangle \big| F_k \right] = \eta_k \langle \nabla \mu(\theta_k), \mathbb{E} \left[ \hat{\nabla} \mu_{k}^{ILR} (\theta_k) \big| F_k \right] \rangle \overset{(*)}{=} \eta_k \|\nabla \mu(\theta_k)\|^2 \]

31
where step (a) follows because the ILR/MLR estimator $\hat{\nu}_k \sim \mu_k (\theta_k)$ is conditionally unbiased (Lemma 3). Taking the expectation of $\|\theta_{k+1} - \theta_k\|^2$ conditioning on $\mathcal{F}_k$ and then applying Theorem 1 we have

$$
\mathbb{E}\left[\|\theta_{k+1} - \theta_k\|^2 | \mathcal{F}_k\right] = \eta_k^2 \mathbb{E}\left[\|\hat{\nu}_{\mu_k} \|^2 | \mathcal{F}_k\right] \\
\leq \eta_k^2 \frac{c}{|\mathcal{U}_k|} \mathbb{E}\left[\|\nu_{\mu_k} \|^2 \right] + \eta_k^2 \left(1 - \frac{c}{|\mathcal{U}_k|}\right) \|\nu_{\mu_k}\|^2.
$$

(58)

Then by taking the unconditional expectation of both sides of (59) and plugging in (57) and (58), we have

$$
\mathbb{E}[\mu(\theta_k)] - \mathbb{E}[\mu(\theta_{k+1})] \\
\leq -\eta_k \mathbb{E}\left[\|\nabla \mu(\theta_k)\|^2\right] + L\eta_k^2 \frac{c}{|\mathcal{U}_k|} \mathbb{E}\left[\|\nu_{\mu_k} \|^2 \right] + L\eta_k^2 \left(1 - \frac{c}{|\mathcal{U}_k|}\right) \|\nu_{\mu_k}\|^2 \\
\leq -\eta_k \mathbb{E}\left[\|\nabla \mu(\theta_k)\|^2\right] + L\eta_k^2 \frac{c}{|\mathcal{U}_k|} \mathbb{E}\left[\|\nabla \mu(\theta_k)\|^2 \right] + L\eta_k^2 \mathbb{E}\left[\|\nabla \mu(\theta_k)\|^2 \right].
$$

(59)

Rearranging both sides of (59) gives

$$
(\eta_k - L\eta_k^2) \mathbb{E}\left[\|\nabla \mu(\theta_k)\|^2\right] \leq \left(\mathbb{E}[\mu(\theta_{k+1})] - \mathbb{E}[\mu(\theta_k)]\right) + L\eta_k^2 \frac{c}{|\mathcal{U}_k|} \mathbb{E}\left[\|\nu_{\mu_k} \|^2 \right] + \frac{2cL\eta_k^2 \sigma^2}{|\mathcal{U}_k|}.
$$

(60)

By applying Lemma 8, i.e., $\mathbb{E}[\|\nu_k \|^2] \leq 2\mathbb{E}[\|\nabla \mu(\theta_k)\|^2] + 2\sigma^2$, we can rewrite (60) as

$$
\eta_k(1 - L\eta_k) \mathbb{E}\left[\|\nabla \mu(\theta_k)\|^2\right] \leq \left(\mathbb{E}[\mu(\theta_{k+1})] - \mathbb{E}[\mu(\theta_k)]\right) + \frac{2cL\eta_k^2 \sigma^2}{|\mathcal{U}_k|}.
$$

(61)

By moving $\frac{2cL\eta_k^2}{|\mathcal{U}_k|} \mathbb{E}[\|\nabla \mu(\theta_k)\|^2]$ to the left hand side of equation (61), we have

$$
\eta_k \left(1 - L\eta_k \left( \frac{2c}{|\mathcal{U}_k|} + 1 \right) \right) \mathbb{E}[\|\nabla \mu(\theta_k)\|^2] \leq \left(\mathbb{E}[\mu(\theta_{k+1})] - \mathbb{E}[\mu(\theta_k)]\right) + \frac{2cL\eta_k^2 \sigma^2}{|\mathcal{U}_k|}.
$$

(62)

Consider $\eta_k$ small enough that $1 - L\eta_k \left( \frac{2c}{|\mathcal{U}_k|} + 1 \right) \geq \frac{1}{2}$ or equivalently $\eta_k \leq \frac{\|\mathcal{U}_k\|}{4cL + 2L|\mathcal{U}_k|}$. If learning rate is non-increasing, this condition can be simplified as initial learning rate $\eta_1 \leq \frac{1}{3cL + 2\sigma}$. Then it proceeds with

$$
\eta_k \left(1 - L\eta_k \left( \frac{2c}{|\mathcal{U}_k|} + 1 \right) \right) \mathbb{E}[\|\nabla \mu(\theta_k)\|^2] \geq \frac{\eta_k}{2} \mathbb{E}[\|\nabla \mu(\theta_k)\|^2].
$$

Thus, by dividing $\eta_k$ from both side of (62), we have

$$
\frac{1}{2} \mathbb{E}[\|\nabla \mu(\theta_k)\|^2] \leq \left(\frac{1}{\eta_k} \mathbb{E}[\mu(\theta_{k+1})] - \frac{1}{\eta_k} \mathbb{E}[\mu(\theta_k)]\right) + \frac{2cL\eta_k^2 \sigma^2}{|\mathcal{U}_k|}.
$$

(63)

Then, by applying (63) and $|\mu(\theta_k)| \leq U_\mu$ and summing over $k = 1, 2, \ldots, K$, we have

$$
\frac{1}{2} \sum_{k=1}^K \mathbb{E}[\|\nabla \mu(\theta_k)\|^2] \\
\leq -\frac{1}{\eta_1} \mathbb{E}[\mu(\theta_1)] + \frac{1}{\eta_k} \mathbb{E}[\mu(\theta_{K+1})] + \sum_{k=2}^K \left(\frac{1}{\eta_{k-1}} - \frac{1}{\eta_k}\right) \mathbb{E}[\mu(\theta_k)] + 2cL\sigma^2 \sum_{k=1}^K \frac{\eta_k}{|\mathcal{U}_k|} \\
\leq -\frac{1}{\eta_1} \mu(\theta_1) + \frac{1}{\eta_k} U_\mu + \sum_{k=2}^K \left(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}\right) U_\mu + 2cL\sigma^2 \sum_{k=1}^K \frac{\eta_k}{|\mathcal{U}_k|} \\
\leq -\frac{1}{\eta_1} \mu(\theta_1) + \frac{2}{\eta_k} U_\mu + 2cL\sigma^2 \sum_{k=1}^K \frac{\eta_k}{|\mathcal{U}_k|}.
$$

(64)

Dividing both side by $K$ gives the result stated in Theorem 2.
Theorem 4 Suppose Assumptions A.1, A.2, and A.3 hold. Let $\eta_k$ denote the learning rate used in the $k$-th iteration and $U_\mu$ is the bound of the objective. For CLR policy gradient estimator, we have
\begin{equation}
\frac{1}{K} \sum_{k=1}^{K} \mathbb{E} \left[ \| \nabla \mu (\theta_k) \|^2 \right] \leq \frac{4 \eta_k U_\mu - \frac{2}{\eta_k} \mu (\theta_1) + (4cL + 16)\sigma^2 \sum_{k=1}^{K} \frac{\eta_k}{|U_k|} + M \sum_{k=1}^{K} Z(k)}{K}
\end{equation}
where $\sigma$ and $L$ are the bounds defined in Assumption A.1 and Lemma 7 and $c$ is the selection constant defined in Theorem 7. $M$ is the bound of policy gradient defined in Lemma 2 and $Z(k)$ is the bound of CLR policy gradient estimator defined in Lemma 4.

Proof. To simplify the proof, we only focus on the different part from Theorem 2. Similar to the proof of Theorem 2, the Lipschitz continuity gives
\begin{equation}
\mu (\theta_k) - \mu (\theta_{k+1}) \leq \langle \nabla \mu (\theta_k), \theta_k - \theta_{k+1} \rangle + L \| \theta_{k+1} - \theta_k \|^2.
\end{equation}

By plugging the policy parameter update $\theta_{k+1} = \theta_k + \eta_k \nabla \mu (\theta_k)$ into (65), we have
\begin{equation}
\mu (\theta_k) - \mu (\theta_{k+1}) \leq - \langle \nabla \mu (\theta_k), \eta_k \nabla \mu_k^{CLR} (\theta_k) \rangle + L \| \theta_{k+1} - \theta_k \|^2.
\end{equation}

Taking the expectation of $\left< \nabla \mu (\theta_k), \eta_k \nabla \mu_k^{CLR} (\theta_k) \right>$ conditioning on $\mathcal{F}_k$, we get
\begin{align}
\mathbb{E} \left[ \left< \nabla \mu (\theta_k), \eta_k \nabla \mu_k^{CLR} (\theta_k) \right| \mathcal{F}_k \right] &= \eta_k \mathbb{E} \left[ \left< \nabla \mu (\theta_k), \nabla \mu (\theta_k) \right| \mathcal{F}_k \right]
\end{align}

where step (67) follows by applying Hölder’s inequality, step (68) follows by applying Lemma 4 and step (69) follows by applying Lemma 2. Taking the expectation of $\| \theta_{k+1} - \theta_k \|^2$ conditioning on $\mathcal{F}_k$ and then applying Corollary 1, we have
\begin{align}
\mathbb{E} \left[ \| \theta_{k+1} - \theta_k \|^2 \right| \mathcal{F}_k] &= \eta_k^2 \mathbb{E} \left[ \left\| \nabla \mu_k^{CLR} (\theta_k) \right\|^2 \right| \mathcal{F}_k] 
\end{align}

where step (70) follows by observing that $\frac{1}{n |U_k|} \leq 1$. Then by taking the unconditional expectation of both sides of (66) and plugging in (69) and (70), we have
\begin{align}
\mathbb{E} [\mu (\theta_k)] - \mathbb{E} [\mu (\theta_{k+1})] &\leq - \eta_k \mathbb{E} \left[ \| \nabla \mu (\theta_k) \|^2 \right] + Z(k) \eta_k M + 8 \eta_k^2 \sigma^2 \\
&\quad + L \eta_k^2 \frac{c}{|U_k|} \mathbb{E} \left[ \left\| \nabla \mu_k^{PG} \right\|^2 \right] + L \eta_k^2 \left( 9 - \frac{c}{|U_k|} \right) \mathbb{E} \left[ \| \nabla \mu (\theta_k) \|^2 \right] 
\end{align}
where step (71) follows Hölder’s inequality.

Proceeding with similar derivation as steps (60)-(62) in the proof of Theorem 2, we can have the following inequality

\[
\eta_k \left(1 - L \eta_k \left(\frac{2c}{|U_k|} + 9\right)\right) \mathbb{E} \left[||\nabla \mu(\theta_k)||^2\right] \leq Z(k) \eta_k M + 8 \eta_k^2 \sigma^2 + \left(\mathbb{E} [\mu(\theta_{k+1})] - \mathbb{E} [\mu(\theta_k)]\right) + \frac{2cL \eta_k^2 \sigma^2}{|U_k|}.
\]  

(72)

Consider \(\eta_k\) small enough such that \(1 - L \eta_k \left(\frac{2c}{|U_k|} + 9\right) \geq \frac{1}{2}\) or equivalently \(\eta_k \leq \frac{1}{4cL + 18L}|U_k|\). If the learning rate is non-increasing, this condition can be simplified as initial learning rate \(\eta_1 \leq \frac{1}{4cL + 18L}\). Then it proceeds with

\[
\frac{\eta_k}{2} \mathbb{E} \left[||\nabla \mu(\theta_k)||^2\right] \geq \left(\frac{1}{\eta_k} \mathbb{E} [\mu(\theta_{k+1})] - \frac{1}{\eta_k} \mathbb{E} [\mu(\theta_k)]\right) + \sigma^2 \frac{2cL + 8}{|U_k|} \eta_k + M Z(k).
\]

(73)

Then, by applying (73) and \(||\mu(\theta_k)|| \leq U_\mu\) and summing over \(k = 1, 2, \ldots, K\), we have

\[
\frac{1}{2} \sum_{k=1}^{K} \mathbb{E} \left[||\nabla \mu(\theta_k)||^2\right] \leq -\frac{1}{\eta_1} \mu(\theta_1) + \frac{1}{\eta_1} U_\mu + \sum_{k=2}^{K} \left(\frac{1}{\eta_k} - \frac{1}{\eta_{k-1}}\right) U_\mu + (2cL + 8)\sigma^2 \sum_{k=1}^{K} \eta_k \frac{M |U_k|}{|U_k|} + M \sum_{k=1}^{K} Z(k)
\]

\[
\leq -\frac{1}{\eta_1} \mu(\theta_1) + \frac{2}{\eta_1} U_\mu + (2cL + 8)\sigma^2 \sum_{k=1}^{K} \eta_k \frac{M |U_k|}{|U_k|} + M \sum_{k=1}^{K} Z(k).
\]

Dividing both sides by \(K\) gives the result stated in Theorem 4.

\[\square\]

B.8 Proof of Theorem 3

Lemma 9. For a non-negative integer \(n\), we have the following inequality

\[
\sum_{i=1}^{n-1} i^p < \frac{n^{p+1}}{p+1}.
\]

Proof. For any number \(r \geq 1\) and \(x \geq -1\), Bernoulli’s Inequality suggests that \(1 + rx \leq (1 + x)^r\). Let \(x = \frac{1}{i-1}\) and \(r = p + 1\). Then we have

\[
1 + \frac{p + 1}{i - 1} \leq \left(1 + \frac{1}{i - 1}\right)^{p+1} \iff (i - 1)^p (p + 1 + i - 1) \leq i^{p+1}
\]

\[
\iff (p + 1) (i - 1)^p \leq i^{p+1} - (i - 1)^{p+1}.
\]

Summing it from 1 to \(n\) and dividing by \(p + 1\) yields

\[
\sum_{i=1}^{n-1} i^p \leq \frac{n^{p+1}}{p+1}.
\]

Theorem 4. Suppose Assumptions A.1, A.2 and A.3 hold. If the size of the reuse set \(|U_k| \sim \mathcal{O}(k^{1/p})\) and the learning rate \(\eta_k = \frac{\eta_1}{k^{1/q}}\) with \(p \geq 1, q > 1\), the proposed algorithm has the bounded convergence rate \(\mathcal{O}(\max\{\frac{1}{k^{1-1/q}}, \frac{1}{k^{1/p+1/q}}\})\).
Proof. With loss of generality, we can write $|U_k| = c_u k^{1/p}$ with a constant $c_u > 0$. By applying Lemma [5] we get
\[ \sum_{i=1}^{K} \frac{1}{K^n} \sum_{k=1}^{K} \frac{1}{k^{1-p-1/q}} \leq \frac{a_{k}}{c_n(1-1/p-1/q)} (k+1)^{1-1/p-1/q} \sim O(k^{1-1/p-1/q}). \]
Then, by applying Theorem [2] we have
\[ \frac{1}{K} \sum_{k=1}^{K} \mathbb{E} [\|\nabla \mu(\theta_k)\|^2] \leq \frac{\frac{4}{n_k} U_{\mu} - \frac{2}{n_k} \mu (\theta_1) + 4cL \sigma^2 \sum_{k=1}^{K} \frac{m_k}{|U_k|}}{K} \]
\[ \leq \frac{\frac{4}{n_k} K^{1/q} U_{\mu} - \frac{2}{n_k} \mu (\theta_1) + 4cL \sigma^2 \sum_{k=1}^{K} \frac{m_k}{|U_k|} (K + 1)^{1-1/p-1/q}}{K} \]
\[ \sim O \left( \max \left\{ \frac{1}{K^{1-1/q}} \right\} \right). \] (74)

B.9 Proof of Lemma [5]

We first introduce the assumption (A.6) which is commonly used in finite-time analysis of stochastic policy optimization [Xu et al., 2020, Wu et al., 2020, Zou et al., 2019]. The second boundedness assumption (A.7) holds for any smooth function with bounded state and action space.

A.6 (Uniform ergodicity) For any $\theta \in \mathbb{R}^d$, consider the infinite-horizon MDP with policy $\pi_\theta$ and transition kernel $P(\cdot|s,a)$. Consider a Markov chain generated by the rule $a_t \sim \pi_\theta(\cdot|s_t)$ and $s_{t+1} \sim p(\cdot|s_t, a_t)$. Then there exists $\kappa_0 > 0$ and $\kappa \in (0, 1)$ such that such that
\[ d_{TV}(P(s_t \in \cdot|s_0 = s), d^{\pi_\theta}(\cdot)) \leq \kappa_0 \kappa^t, \forall t \geq 0, \forall s \in S. \]
The probability mass $P(s_{t+1} \in S|s_t = s, a_t = a) = \int_S p(dy|s,a)$ holds for any measurable set $S \subset S$.

A.7 The likelihood ratio for any $\theta_1, \theta_2 \in \mathbb{R}^d$, i.e., $\sup_{x \in \mathbb{S} \times \mathbb{A}} \left\{ \frac{\rho_{\theta_1}(x)}{\rho_{\theta_2}(x)} \right\} \leq U_{LR}.

A result comes from Assumption A.6 is the relationship between total variation distance between two stationary distribution and their L2 parametric distance; see [Xu et al., 2020] (Lemma 3), [Zou et al., 2019] and [Wu et al., 2020] (Lemma A.1) for detailed proof.

Lemma 10 [Xu et al., 2020]. For any $\theta_1, \theta_2 \in \mathbb{R}^d$ it holds that
\[ \|\rho_{\theta_1}(\cdot) - \rho_{\theta_2}(\cdot)\|_{TV} \leq C_d \|\theta_1 - \theta_2\| \]
where $C_d = U_{\pi}(\|\log_\kappa^{-1} \| + \frac{1}{1-\kappa}).$

Lemma 5 Suppose Assumptions A.1, A.2 and A.3 hold and assume A.6 and A.7 defined in Appendix B.9. Let the step size $\eta$ be an decreasing sequence converging to zero. Then $Z(k) \to 0$ as $k \to \infty.$

Proof. By applying Markov inequality, the bound of bias becomes
\[ Z(k) = \frac{M}{|U_k|} \sum_{\theta \in \Lambda_k} \Pr(f_{i,k}(\mathbf{x}) - 1 > U_f - 1) \leq \frac{M}{|U_k|} \sum_{\theta \in \Lambda_k} \frac{\mathbb{E}_{x \sim \rho_{\theta_k}}[f_{i,k}(\mathbf{x}) - 1]}{U_f - 1}. \]
Notice that the term on the right hand side is
\[ \frac{M}{|U_k|} \sum_{\theta \in \Lambda_k} \mathbb{E}_{x \sim \rho_{\theta_k}}[f_{i,k}(\mathbf{x}) - 1] = \frac{M}{|U_k|} \sum_{\theta \in \Lambda_k} \frac{1}{U_f - 1} \int_{S \times \mathbb{A}} \rho_{\theta_k}(\mathbf{x})(\rho_{\theta_k}(\mathbf{x}) - \rho_{\theta}(\mathbf{x}))d\mathbf{x}. \]
Following Assumption A.7, we can have
\[ Z(k) \leq \frac{MU_{LR}}{|U_k|(U_f - 1)} \sum_{\theta \in \Lambda_k} \int_{S \times \mathbb{A}} (\rho_{\theta_k}(\mathbf{x}) - \rho_{\theta}(\mathbf{x}))d\mathbf{x} \leq \frac{2MU_{LR}}{|U_k|(U_f - 1)} \sum_{\theta \in \Lambda_k} \|\rho_{\theta_k}(\cdot) - \rho_{\theta}(\cdot)\|_{TV} \] (75)
Case I: Infinite-Horizon. By applying Lemma 10 to step (75), we can have

$$Z(k) \leq C_d \frac{2MU_{LR}}{|U_k|(U_f - 1)} \sum_{\theta_i \in U_k} \|\theta_k - \theta_i\|_2.$$ 

Notice that $$\|\theta_k - \theta_i\| = \left\| \sum_{i'=i}^{k} \eta^c \nabla \mu^c_{i'} \right\| \leq \sum_{i'=i}^{k} \eta^c \left\| \nabla \mu^c_{i'} \right\|$$ and $$\left\| \nabla \mu^c_{i'} \right\| \leq U_M. Then we have

$$Z(k) \leq \frac{2MU_{LR}CdU_f}{|U_k|(U_f - 1)} \sum_{\theta_i \in U_k} \left( k - i \right) \eta_i \leq \frac{2M^2U_{LR}CdU_f}{|U_k|(U_f - 1)} \sum_{\theta_i \in U_k} \eta_{k-B} \leq \frac{2BM^2U_{LR}CdU_f}{|U_k|(U_f - 1)} \eta_{k-B}$$ (76)

where step (76) follows that $$\eta_k$$ is decreasing and $$i \geq k - B$$ since $$|U_k| \leq B.$$ The convergence of $$Z(k) \to 0$$ as $$k \to \infty$$ immediately follows the fact that $$\eta_{k-B} \to 0$$ for any fixed $$B$$ or any $$B$$ that is proportional to $$k$$ (e.g., $$B = k/3$$).

Case II: Finite-Horizon. In this case, we don’t need Assumption A.6.

As $$\rho_{\theta_k} = \prod_{t=1}^{H} \pi_{\theta_k}(a_t | s_t)$$ and $$\rho_{\theta} = \prod_{t=1}^{H} \pi_{\theta}(a_t | s_t),$$ the total variation distance between $$\rho_{\theta_k}$$ and $$\rho_{\theta},$$ becomes

$$\|\rho_{\theta_k}(x) - \rho_{\theta}(x)\|_{TV} \leq \int_{S} \prod_{t=1}^{H} \left( \int_{A} p(d\mathbf{s}_t) \int_{A} \pi_{\theta_k}(a_t | \mathbf{s}_t) - \pi_{\theta}(a_t | \mathbf{s}_t) \right) d\mathbf{a}_t \leq \prod_{t=1}^{H} \|\pi_{\theta_k}(\cdot | \mathbf{s}_t) - \pi_{\theta}(\cdot | \mathbf{s}_t)\|_{TV} \leq \prod_{t=1}^{H} U_{\pi} \|\theta_k - \theta_i\|_2$$ (78)

$$\leq U_{\pi} \prod_{t=1}^{H} \sum_{i'=i}^{k} \eta^c \left\| \nabla \mu^c_{i'} \right\|_2$$ (79)

where step (78) follows Assumption A.2 and step (79) holds by observing $$\|\theta_k - \theta_i\| \leq \sum_{i'=i}^{k} \eta^c \left\| \nabla \mu^c_{i'} \right\|.$$ Plugging the above result into (75) gives,

$$Z(k) \leq C_d \frac{2MU_{LR}U_{\pi}H}{|U_k|(U_f - 1)} \sum_{\theta_i \in U_k} \prod_{t=1}^{H} \sum_{i'=i}^{k} \eta^c \left\| \nabla \mu^c_{i'} \right\|_2.$$ (80)

Following the similar proof in Case I, step (80) becomes

$$Z(k) \leq \frac{2M^{H+1}U_{LR}U_{\pi}H}{|U_k|(U_f - 1)} \sum_{\theta_i \in U_k} \prod_{t=1}^{H} \sum_{i'=i}^{k} \eta^c \leq \frac{2M^{H+1}U_{LR}U_{\pi}H}{|U_k|(U_f - 1)} \sum_{\theta_i \in U_k} B^{H} \eta_{k-B}$$ (82)

$$\leq \frac{2M^{H+1}U_{LR}U_{\pi}H}{U_f - 1} B^{H} \eta_{k-B}$$ (83)
where step (81) follows by observing $\|\nabla_{\mu_{\mathcal{LR}}}^C\|_2 \leq U_f M$, step (82) follows that $\eta_k$ is decreasing and $i \geq k-B$ since $|U_k| \leq B$. The result stated in Lemma 5 immediately follows from step (83) by observing that $\eta_{k-B} \to 0$ as $k \to \infty$. \hfill \Box

B.10 Proof of Theorem 5

**Theorem 5.** Assume A.1 A.5 hold. Let $\eta_k$ denote the learning rate in the $k$-th iteration and $\theta^*$ denote a local maximizer. We have $\theta_k \xrightarrow{a.s.} \theta^*$ as $k \to \infty$ if $\sum_{k=1}^{\infty} \eta_k^2 < \infty$.

**Proof.** Step 1: Convergence in mean. The locally strong convexity of $-\mu(\theta)$ implies $\|\nabla_{\mu}(y) - \nabla_{\mu}(x)\| \geq \lambda \|x - y\|$ with a constant $\lambda > 0$ for any $x, y \in \mathcal{E}_r(\theta^*)$. By Assumption A.4, when $k$ is large enough, we have $\theta_k$ in a neighborhood of a local maximizer $\theta^*$, i.e., $\theta_k \in \mathcal{E}_r(\theta^*)$ such that $\|\nabla_{\mu}(\theta_k) - \nabla_{\mu}(\theta^*)\| \geq \lambda \|\theta_k - \theta^*\|$. By using Minkowski’s inequality, we have

$$\lambda \|\theta_k - \theta^*\| \leq \|\nabla_{\mu}(\theta_k) - \nabla_{\mu}(\theta^*)\| \leq \|\nabla_{\mu}(\theta_k)\| + \|\nabla_{\mu}(\theta^*)\| \leq \|\nabla_{\mu}(\theta_k)\|.$$ 

where (⋆) holds by applying Minkowski’s inequality and (⋆⋆) follows because $\theta^*$ is a local maximizer and $\|\nabla_{\mu}(\theta^*)\| = 0$. Then, by taking the expectation of both sides, we have $\lambda \mathbb{E}[\|\theta_k - \theta^*\|^2] \leq \mathbb{E}[\|\nabla_{\mu}(\theta_k)\|^2]$. Summing up to infinity gives

$$\frac{1}{K} \sum_{k=1}^{K} \mathbb{E}[\|\theta_k - \theta^*\|^2] \leq \frac{1}{\lambda K} \sum_{k=1}^{K} \mathbb{E}[\|\nabla_{\mu}(\theta_k)\|^2].$$

Therefore, by applying (74), we have $\mathbb{E}[\|\theta_k - \theta^*\|] \to 0$ as $k \to \infty$.

Step 2: Convergence almost surely. Following (88), we have

$$\mathbb{E}\left[\|\theta_{k+1} - \theta_k\|^2 \mid \mathcal{F}_k\right] = \eta_k^2 \mathbb{E}\left[\|\nabla_{\mu(MLR)} M\|^2 \mid \mathcal{F}_k\right]$$

$$\leq \eta_k^2 \frac{c}{|U_k|} \mathbb{E}\left[\|\nabla_{\mu(PG)} M\|^2 \mid \theta_k\right] + \eta_k^2 \left(1 - \frac{c}{|U_k|}\right) \|\nabla_{\mu}(\theta_k)\|^2$$

$$\leq \eta_k^2 \frac{2c}{|U_k|} \mathbb{E}\left[\|\nabla_{\mu}(\theta_k)\|^2 \mid \mathcal{F}_k\right] + \eta_k^2 \left(1 - \frac{c}{|U_k|}\right) \|\nabla_{\mu}(\theta_k)\|^2$$

$$= \eta_k^2 \left(1 + \frac{2c}{|U_k|} \sigma^2 + \frac{c}{|U_k|} M^2\right)$$

where (84) follows by applying Theorems 1, 85 follows by applying Lemma 8, i.e.,

$$\mathbb{E}\left[\|\nabla_{\mu(PG)} M\|^2 \mid \theta_k\right] \leq 2 \mathbb{E}\left[\|\nabla_{\mu}(\theta_k)\|^2 \mid \theta_k\right] + 2 \sigma^2;$$

and (86) follows by applying Lemma 2 and Jensen’s inequality:

$$\|\nabla_{\mu}(\theta_k)\|^2 = \|\mathbb{E}[g_k(x) \mid \theta_k]\|^2 \leq \mathbb{E}\left[\|g_k(x)\|^2 \mid \theta_k\right] \leq M^2.$$ 

By expanding the squares, we have

$$\|\theta_{k+1} - \theta^*\|^2 = \|\theta_k - \theta^*\|^2 + 2 \eta_k \left\langle \nabla_{\mu(MLR)} M, \theta_k - \theta^* \right\rangle + \|\theta_{k+1} - \theta_k\|^2.$$ 

Then taking conditional expectation $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$ gives

$$\mathbb{E}_k[\|\theta_{k+1} - \theta^*\|^2] = \|\theta_k - \theta^*\|^2 + 2 \eta_k \left\langle \nabla_{\mu}(\theta_k), \theta_k - \theta^* \right\rangle + \mathbb{E}_k[\|\theta_{k+1} - \theta_k\|^2].$$
Applying \((87)\) and A.4, i.e. \(\langle \nabla \mu(x), x - y \rangle \leq \mu(x) - \mu(y) - \lambda \| x - y \| \leq \mu(x) - \mu(y)\) gives
\[
\mathbb{E}_k \left[ \| \theta_{k+1} - \theta^* \|^2 \right] \leq \| \theta_k - \theta^* \|^2 + 2\eta_k (\mu(\theta_k) - \mu(\theta^*)) + \eta_k^2 \left( 1 + \frac{2c}{|U_k|} \sigma^2 + \frac{c}{|U_k|} M^2 \right).
\]
After rearranging, we have
\[
\mathbb{E}_k \left[ \| \theta_{k+1} - \theta^* \|^2 \right] + 2\eta_k (\mu(\theta^*) - \mu(\theta_k)) \leq \| \theta_k - \theta^* \|^2 + \eta_k^2 \left( 1 + \frac{2c}{|U_k|} \sigma^2 + \frac{c}{|U_k|} M^2 \right).
\]
Let \(V_k = \| \theta_k - \theta^* \|^2, U_k = 2\eta_k (\mu(\theta^*) - \mu(\theta_k))\) and \(Z_k = \eta_k^2 \left( 1 + \frac{2c}{|U_k|} \sigma^2 + \frac{c}{|U_k|} M^2 \right)\). Notice \(\sum_k Z_k \leq \infty\) if \(\sum_{k=1}^\infty \eta_k^2 < \infty\) (Robbins-Monro condition). Using Lemma 6 we have that \((\| \theta_k - \theta^* \|^2)_{k \geq 1}\) converges almost surely to some random variable \(V_{\infty}\).

For the converged sequence \((\| \theta_k - \theta^* \|^2)_{k \geq 1}\), \(\| \theta_k - \theta^* \|^2\) is bounded almost surely. Thus from boundary convergence theorem follows \(E[V_{\infty}] = E[\lim_{k \to \infty} \| \theta_k - \theta^* \|^2] = \lim_{k \to \infty} E[\| \theta_k - \theta^* \|^2] = 0\), which implies \(V_{\infty} = 0\) almost surely. It further implies \(\theta_k \xrightarrow{a.s.} \theta^*\).

\[\square\]

B.11 Proof of Proposition 4

Proposition 4, A.1, A.5 hold. Let \(\theta^*\) denote a local maximizer. Given any feasible observation \(x\), we have \(\rho_{\theta_k}(x) \xrightarrow{a.s.} \rho_{\theta^*}(x)\) and \(\frac{1}{\rho_{\theta_k}(x) a.s.} \to \frac{1}{\rho_{\theta^*}(x)} \) as \(k \to \infty\).

Proof. Proof. By applying Assumption A.5, Theorem 5 and Continuous Mapping Theorem, the results immediately follow, i.e., \(\rho_{\theta_k}(x) \xrightarrow{a.s.} \rho_{\theta^*}(x)\) and \(\frac{1}{\rho_{\theta_k}(x) a.s.} \to \frac{1}{\rho_{\theta^*}(x)} \).

B.12 Proof of Theorem 6

Theorem 6 Suppose Assumptions A.1-A.5 hold. The size of the reuse set \(|U_k|\) increases to infinity with probability 1 as iteration \(k \to \infty\), i.e., \(P(\lim_{k \to \infty} |U_k| = \infty) = 1\).

Proof. Let \(f(\cdot) = \| \cdot \|^2\) and define the Jensen’s gap as
\[
\epsilon_k = E[f(\gamma_k(x))] - f(E[\gamma_k(x)]) = E[\| \gamma_k(x) \|^2] - E[\| \gamma_k(x) \|^2].
\]
Define the scenario-based policy gradient estimate at true optimum \(\theta^*\) by \(g(x; \theta^*)\)

By applying \(\theta_k \xrightarrow{a.s.} \theta^*\) (Theorem 5), continuity assumption A.5, and continuous mapping theorem, at any fixed argument \(x\), we have \(g_k(x) \xrightarrow{a.s.} g(x; \theta^*)\) as \(k \to \infty\). Under the some regularity conditions, i.e., the likelihood function is dominated almost everywhere by some integrable function \(\psi(\cdot)\) in the sense that \(\rho_{\theta_k}(x) \leq \psi(x)\), applying Assumptions A.5 and Lemma 2 we have
\[
e_k = \lim_{k \to \infty} \epsilon_k = \left( \int \lim_{k \to \infty} \| \gamma_k(x) \|^2 d\gamma_k(x) \right) - E[\| \gamma(x; \theta^*) \|^2] = E[\| g(x; \theta^*) \|^2] - E[\| g(x; \theta^*) \|^2] > 0
\]
where \((\bullet)\) follows by applying dominated convergence theorem. Jensen’s inequality becomes equality only when \(f(\cdot)\) is affine and/or \(g(x; \theta^*)\) is constant. However, neither of these conditions holds, which implies that the Jensen’s gap strictly greater than zero, i.e., \(e_k > 0\). It further implies \(\lim_{k \to \infty} \epsilon_k > 0\). Otherwise there is a subsequence \(\{e_k\}_{k=1}^\infty\) that converges to 0 (which contradicts to the fact that \(e_k \leq \lim_{k \to \infty} \epsilon_k > 0\)). Let \(\epsilon = \lim_{k \to \infty} \epsilon_k\).

Define the difference \(\Delta_k \equiv \text{Tr} \left( \text{Var} \left[ \nabla \mu_{k}^{IR} \right] \right) - \epsilon \text{Tr} \left( \text{Var} \left[ \nabla \mu_{k}^{PG} \right] \theta_k \right)\). In the following proof, and write \(\mathbb{E}_k[\cdot] = E[\cdot|\theta_k]\) to represent the conditional expectation over the \(k\)-th behavioral distribution.
\[ \Delta_{i,k} = \text{tr} \left( \text{Var} \left[ \frac{1}{n} \sum_{j=1}^{n} \frac{\rho_{\theta_i}(x^{(i,j)})}{\rho_{\theta_i}(x^{(i,j)})} g_k(x^{(i,j)}) \big| \mathcal{F}_k \right] \right) - c \text{tr} \left( \text{Var} \left[ \frac{1}{n} \sum_{j=1}^{n} g_k(x^{(k,j)}) \big| \theta_k \right] \right) \]

\[ = \frac{1}{n} \text{tr} \left( \int \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)}^2 g_k(x) g_k(x)^T \rho_{\theta_i}(x) dx \right) - c \frac{1}{n} \text{tr} \left( \int g_k(x) g_k(x)^T \rho_{\theta_i}(x) dx \right) \]

\[ - \frac{1}{n} \text{tr} \left( \mathbb{E}_i \left[ \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)} g_k(x) \right] \mathbb{E}_i \left[ \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)} g_k(x) \right]^T \right) + c \frac{1}{n} \text{tr} \left( \mathbb{E}_k [g_k(x)] \mathbb{E}_k [g_k(x)^T] \right) \]

\[ = \frac{1}{n} \text{tr} \left( \int \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)}^2 g_k(x) g_k(x)^T \rho_{\theta_i}(x) dx \right) - c \frac{1}{n} \text{tr} \left( \int g_k(x) g_k(x)^T \rho_{\theta_i}(x) dx \right) \]

\[ - \frac{1}{n} \text{tr} \left( \nabla \mu(\theta_k) \nabla \mu(\theta_k)^T - c \nabla \mu(\theta_k) \nabla \mu(\theta_k)^T \right) \]  

(89)

where (89) holds by following that \( \nabla \mu(\theta_k) = \mathbb{E}[g_k(x)] \). Then, by letting \( \epsilon = \frac{c(\epsilon - 1)}{nM^2} > 0 \), (90) becomes

\[ \Delta_{i,k} = \frac{1}{n} \int \left( \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)} - 1 - \epsilon \right) \|g_k(x)\|^2 \rho_{\theta_i}(x) dx \]

\[ + \frac{c - 1}{n} \epsilon \]

\[ \leq \frac{1}{n} \int \left( \frac{\rho_{\theta_i}(x)}{\rho_{\theta_i}(x)} - 1 - \epsilon \right) \|g_k(x)\|^2 \rho_{\theta_i}(x) dx + \epsilon \frac{c - 1}{n} \epsilon \]  

(91)

where (91) holds because of Lemma 2 and (92) follows since \( \epsilon_k \) is greater than \( \epsilon \) infinitely often (i.o.) as \( \epsilon = \lim \inf \epsilon_k \).

Let \( E_{i,k} = \{ \Delta_{i,k} \leq 0 \} \) denote the event of selecting the behavioral distribution \( \theta_i \in \mathcal{F}_k \) in iteration \( k \). By Proposition 1, we have \( \rho_{\theta_i}(x) \xrightarrow{a.s.} 1 \) as \( k \to \infty \) for any \( i \in \{ \frac{k}{2}, \ldots, k \} \) as both density function a.s. converges to \( \rho_{\theta_i}(x) \). Then when \( k \) is large enough, we have \( \Delta_{i,k} \leq 0 \) for at least half of past iterations. It means \( \sum_{k} P \sum_{i=k/2}^{k} (\Delta_{i,k}) = \infty \).

By Second Borel-Cantelli Lemma II [Durrett 2019, Theorem 4.3.4], we have \( P(\limsup_{k \to \infty} E_i) = 1 \). It means that the selection criterion (20) holds infinitely often when \( k \to \infty \). It suggests \( \lim_{k \to \infty} |\mathcal{U}_k| = \infty \) almost surely.