An Improved Analysis of (Variance-Reduced) Policy Gradient and Natural Policy Gradient Methods

Yanli Liu[♮] **Kaiqing Zhang**[†] **Tamer Başar**[†] **Wotao Yin**[♮]
[♮]Department of Mathematics, University of California, Los Angeles
[†]Department of ECE and CSL, University of Illinois at Urbana-Champaign
{yanli, wotaoyin}@math.ucla.edu, {kzhang66, basar1}@illinois.edu

Abstract

In this paper, we revisit and improve the convergence of policy gradient (PG), natural PG (NPG) methods, and their variance-reduced variants, under general smooth policy parametrizations. More specifically, with the Fisher information matrix of the policy being positive definite: i) we show that a state-of-the-art variance-reduced PG method, which has only been shown to converge to stationary points, converges to the globally optimal value up to some inherent function approximation error due to policy parametrization; ii) we show that NPG enjoys a lower sample complexity; iii) we propose SRVR-NPG, which incorporates variance-reduction into the NPG update. Our improvements follow from an observation that the convergence of (variance-reduced) PG and NPG methods can improve each other: the stationary convergence analysis of PG can be applied to NPG as well, and the global convergence analysis of NPG can help to establish the global convergence of (variance-reduced) PG methods. Our analysis carefully integrates the advantages of these two lines of works. Thanks to this improvement, we have also made variance-reduction for NPG possible, with both global convergence and an efficient finite-sample complexity.

1 Introduction

Policy gradient (PG) methods, or more generally direct policy search methods, have long been recognized as one of the foundations of reinforcement learning (RL) [1]. Specifically, PG methods directly search for the optimal policy parameter that maximizes the long-term return in Markov decision processes (MDPs), following the policy gradient ascent direction [2, 3]. This search direction can be more efficient using a preconditioning matrix, e.g., using the natural PG direction [4]. These methods have achieved tremendous empirical successes recently, especially boosted by the power of (deep) neural networks for policy parametrization [5, 6, 7, 8]. These successes are primarily attributed to the fact that PG methods naturally incorporate *function approximation* for policy parametrization, in order to handle massive and even continuous state-action spaces.

In practice, the policy gradients are usually estimated via samples using Monte-Carlo rollouts and bootstrapping [2, 9]. Such stochastic PG methods notoriously suffer from very high variances, which not only destabilize but also slow down the convergence. Several conventional approaches have been advocated to reduce the variance of PG methods, e.g., by adding a baseline [3, 10], or by using function approximation for estimating the value function, namely, developing actor-critic algorithms [11, 12, 13]. More recently, motivated by the advances of variance-reduction techniques in stochastic optimization [14, 15, 16, 17], there have been surging interests in developing *variance-reduced* PG methods [18, 19, 20, 21, 22], which are shown to be faster.

In contrast to the empirical successes of PG methods, their theoretical convergence guarantees, especially *non-asymptotic global* convergence guarantees, have not been addressed satisfactorily

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

until very recently [23, 24, 25, 26, 27]. By *non-asymptotic global* convergence, here we mean the convergence behavior of PG methods from any initialization, and the quality of the point they converge to (usually enjoys global optimality up to some compatible function approximation error due to policy parametrization), after a finite number of iterations/samples. These recent prominent guarantees are normally beyond the folklore *first-order* stationary-point convergence[1], as expected from a *stochastic nonconvex optimization* perspective of solving RL with PG methods. Special landscapes of the RL objective, though nonconvex, have enabled the convergence to even globally optimal values. On the other hand, none of the aforementioned variance-reduced PG methods [18, 19, 20, 21, 22] have been shown to enjoy these desired global convergence properties. It remains unclear whether these methods can converge to beyond first-order stationary policies.

Motivated by these advances and the questions that remain to be answered, we aim in this paper to improve the convergence of PG and natural PG (NPG) methods, and their variance-reduced variants, under general smooth policy parametrizations. Our contributions are summarized as follows.

Contributions. With a focus on the conventional Monte-Carlo-based PG methods, we propose a general framework for analyzing their *global convergence*. Our contribution is three-fold: first, we establish the global convergence up to compatible function approximation errors due to policy parametrization, for a variance-reduced PG method SRVR-PG [21]; second, we improve the global convergence of NPG methods established in [27], from $\mathcal{O}\left(\varepsilon^{-4}\right)$ to $\mathcal{O}\left(\varepsilon^{-3}\right)$; third, we propose a new variance-reduced algorithm based on NPG, and establish its global convergence with an efficient sample-complexity. These improvements are based on a framework that integrates the advantages of previous analyses on (variance reduced) PG and NPG, and rely on a (mild) assumption that the Fisher information matrix induced by the policy parametrization is positive definite (see Assumption 2.1). A comparison of previous results and our improvements is laid out in Table 1.

Related Work.

Global Convergence of (Natural) PG. Recently, there has been a surging research interest in investigating the global convergence of PG and NPG methods, which is beyond the folklore convergence to first-order stationary policies. In the special case with linear dynamics and quadratic reward, [23] shows that PG methods with random search converge to the globally optimal policy with linear rates. In [24], with a simple reward-reshaping, PG methods have been shown to converge to the second-order stationary-point policies. [26] shows that for finite-MDPs and several control tasks, the nonconvex RL objective has no suboptimal local minima. [25] prove that (natural) PG methods converge to the globally optimal value when overparametrized neural networks are used for function approximation. [27] provides a fairly general characterization of global convergence for these methods, and a basic sample complexity result for sample-based NPG updates. It is also worth noting that trust-region policy optimization (TRPO) [5], as a variant of NPG, also enjoys global convergence with overparametrized neural networks [28], and for regularized MDPs [29]. Very recently, for actor-critic algorithms, a series of non-asymptotic convergence results have also been established [30, 31, 32, 33], with global convergence guarantees when natural PG/PPO are used in the actor step.

Variance-Reduction (VR) for PG. Conventional approaches to reduce the high variance in PG methods include using (natural) actor-critic algorithms [11, 12, 13], and adding baselines [3, 10]. The idea of variance reduction (VR) is first proposed to accelerate stochastic minimization. VR algorithms such as SVRG [14, 15, 16], SAGA [17], SARAH [34], and Spider [35] achieve acceleration over SGD in both convex and nonconvex settings. SVRG is also accelerated by applying a positive definite preconditioner that captures the curvature of the objective [36]. Inspired by these successes in stochastic optimization, VR is also incorporated into PG methods [18], with empirical validations for acceleration, and analyzed rigorously in [19]. Then, [20] improves the sample complexity of SVRPG, and [21] proposes a new SRVR-PG method that uses recursively updated semi-stochastic policy gradient, which leads to an improved sample complexity of $\mathcal{O}(\varepsilon^{-1.5})$ over previous works. More recently, [22] proposes a new STORM-PG method, which blends momentum in the update and matches the sample complexity of in [21], and [37] applies the idea of SARAH and considers a more general setting with regularization. Finally, heavy-ball type of momentum has also been applied to PG methods [38]. We highlight that all these sample complexity results are for first-order stationary-point convergence (which might have arbitrarily bad performance: see (2.2)), in contrast

[1] That is, finding a parameter θ such that $\|\nabla J(\theta)\|^2 \leq \varepsilon$, where J is the expected return.

to the more desired global convergence guarantees (up to some function approximation errors that can be small) that we are interested in.

NPG [27]	NPG [25]	TRPO [28]	TRPO [29]
$\mathcal{O}(\varepsilon^{-4})$	$\mathcal{O}(T_{TD}\varepsilon^{-2})$ [1]	$\mathcal{O}(\varepsilon^{-8})$	$\mathcal{O}(\varepsilon^{-4})$

NPG (2.8)	PG (2.4)	SRVR-PG [21] (Algorithm 2)	SRVR-NPG (Algorithm 1)
$\mathcal{O}(\varepsilon^{-3})$	$\mathcal{O}(\sigma^2\varepsilon^{-4})$	$\mathcal{O}\left((W+\sigma^2)\varepsilon^{-3}\right)$	$\mathcal{O}\left((W+\sigma^2)\varepsilon^{-2.5}+\varepsilon^{-3}\right)$

Table 1: Comparison of sample complexities of several methods to reach global optimality up to some compatible function approximation error (see (2.9)). Our results are listed in the second table (See App. A for their derivations). We compare the number of trajectories to reach $\varepsilon-$optimality in expectation, up to some inherent error due to the function approximation for policy parametrization (see (2.3)). σ^2 is an upper bound for the variance of gradient estimator (see Assumption 4.1), and W is an upper bound for the variance of importance weight (see Assumption 4.3).

2 Preliminaries

We first introduce some preliminaries regarding both the MDPs and policy gradient methods.

2.1 Markov Decision Processes

Consider a discounted Markov decision process defined by a tuple $(\mathcal{S}, \mathcal{A}, \mathbb{P}, R, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces of the agent, $\mathbb{P}(s' \mid s, a) : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the Markov kernel that determines the transition probability from (s, a) to state s', $\gamma \in (0, 1)$ is the discount factor, and $r : \mathcal{S} \times \mathcal{A} \to [-R, R]$ is the reward function of s and a.

At each time t, the agent executes an action $a_t \in \mathcal{A}$ given the current state $s_t \in \mathcal{S}$, following a possibly stochastic policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, i.e., $a_t \sim \pi(\cdot \mid s_t)$. Then, given the state-action pair (s_t, a_t), the agent observes a reward $r_t = r(s_t, a_t)$. Thus, under any policy π, one can define the *state-action value* function $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as

$$Q^\pi(s, a) := \mathbb{E}_{a_t \sim \pi(\cdot \mid s_t), s_{t+1} \sim \mathbb{P}(\cdot \mid s_t, a_t)} \left(\sum_{t=0}^{\infty} \gamma^t r_t \,\middle|\, s_0 = s, a_0 = a \right).$$

One can also define the *state-value* function $V^\pi : \mathcal{S} \to \mathbb{R}$, and the *advantage* function $A^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, under policy π, as $V^\pi(s) := \mathbb{E}_{a \sim \pi(\cdot \mid s)}[Q^\pi(s, a)]$ and $A^\pi(s, a) := Q^\pi(s, a) - V^\pi(s)$, respectively. Suppose that the initial state s_0 is drawn from some distribution ρ. Then, the goal of the agent is to find the optimal policy that maximizes the expected discounted return, namely,

$$\max_\pi \quad J(\pi) := \mathbb{E}_{s_0 \sim \rho}[V^\pi(s_0)]. \tag{2.1}$$

In practice, both the state and action spaces $\mathcal{S}$ and $\mathcal{A}$ can be very large. Thus, the policy π is usually parametrized as π_θ for some parameter $\theta \in \mathbb{R}^d$, using, for example, deep neural networks. As such, the goal of the agent is to maximize $J(\pi_\theta)$ in the space of the parameter θ, which naturally induces an optimization problem. Such a problem is in general nonconvex [24, 27], making it challenging to find the globally optimal policy.

For notational convenience, let us denote $J(\pi_\theta)$ by $J(\theta)$. Many of the previous works focus on establishing stationary convergence of policy gradient methods. That is, finding a θ that satisfies

$$\|\nabla J(\theta)\|^2 \leq \varepsilon. \tag{2.2}$$

Obviously, such a θ may not lead to a large $J(\theta)$. Instead, we are interested in finding a θ such that

$$J^\star - J(\theta) \leq \mathcal{O}(\sqrt{\varepsilon_{\text{bias}}}) + \varepsilon, \tag{2.3}$$

where $J^\star = \max_\pi J(\pi)$, and the $\mathcal{O}(\sqrt{\varepsilon_{\text{bias}}})$ term reflects the inherent error related to the possibly limited expressive power of the policy parametrization π_θ (see Assumption 4.4 for the definition).

[1] In [25], T_{TD} iterations of temporal difference updates are needed at each iteration, T_{TD} can be large for wide neural networks. See App. A for details.

2.2 (Natural) Policy Gradient Methods

To solve the optimization problem (2.1), one standard way is via the policy gradient (PG) method [3]. Specifically, let $\tau_i = \{s_0^i, a_0^i, s_1^i, \cdots\}$ denote the data of a sampled trajectory under policy π_θ. Then, a stochastic PG ascent update is given as

$$\theta^{k+1} = \theta^k + \eta \cdot \frac{1}{N}\sum_{i=1}^{N} g(\tau_i \mid \theta^k), \tag{2.4}$$

where $\eta > 0$ is a stepsize, N is the number of trajectories, and $g(\tau_i \mid \theta^k)$ estimates $\nabla J(\theta^k)$ using the trajectory τ_i. Common unbiased estimators of PG include REINFORCE [2], using the policy gradient theorem [39], and GPOMDP [9]. The commonly used GPOMDP estimator will be given by

$$g(\tau_i \mid \theta) = \sum_{h=0}^{\infty}\left(\sum_{t=0}^{h}\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\right)\left(\gamma^h r(s_h^i, a_h^i)\right), \tag{2.5}$$

where $\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)$ is the *score function*. If the expectation of this infinite sum exits, then (2.5) becomes an unbiased estimate of the policy gradient of the objective $J(\theta)$ defined in (2.1). This unbiasedness is established in App. B for completeness.

In practice, a *truncated* version of GPOMDP is used to approximate the infinite sum in (2.5), as

$$g(\tau_i^H \mid \theta) = \sum_{h=0}^{H-1}\left(\sum_{t=0}^{h}\nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\right)\left(\gamma^h r(s_h^i, a_h^i)\right), \tag{2.6}$$

where $\tau_i^H = \{s_0^i, a_0^i, s_1^i, \cdots, s_{H-1}^i, a_{H-1}^i, s_H^i\}$ is a truncation of the full trajectory τ_i of length H. (2.6) is thus a biased stochastic estimate of $\nabla J(\theta)$, with the bias being negligible for a large enough H. For notational simplicity, we denote the H-horizon trajectory distribution induced by the initial state distribution ρ and policy π_θ as $p_\rho^H(\cdot \mid \theta)$, that is,

$$p_\rho^H(\tau^H \mid \theta) = \rho(s_0)\prod_{h=0}^{H-1}\pi_\theta(a_h \mid s_h)\mathbb{P}(s_{h+1} \mid a_h, s_h).$$

Hereafter, unless otherwise stated, we refer to this *H-horizon trajectory* simply as *trajectory*, drawn from $p_\rho^H(\cdot \mid \theta)$.

As a significant variant of PG, NPG [4] also incorporates a preconditioning matrix $F_\rho(\theta)$, leading to the following update

$$F_\rho(\theta) = \mathbb{E}_{s\sim d_\rho^{\pi_\theta}}[F_s(\theta)], \qquad \theta^{k+1} = \theta^k + \eta \cdot F_\rho^\dagger(\theta^k)\nabla J(\theta^k), \tag{2.7}$$

where $F_s(\theta) = \mathbb{E}_{a\sim\pi_\theta(\cdot \mid s)}\left[\nabla_\theta \log \pi_\theta(a \mid s)\nabla_\theta \log \pi_\theta(a \mid s)^\top\right]$ is the Fisher information matrix of $\pi_\theta(\cdot \mid s) \in \mathcal{P}(\mathcal{A})$, $F_\rho^\dagger(\theta^k)$ is the Moore-Penrose pseudoinverse of $F_\rho(\theta^k)$, and $d_\rho^{\pi_\theta} \in \mathcal{P}(\mathcal{S})$ is the state visitation measure induced by policy π_θ and initial distribution ρ, which is defined as

$$d_\rho^{\pi_\theta}(s) := (1-\gamma)\mathbb{E}_{s_0\sim\rho}\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t = s \mid s_0, \pi_\theta).$$

The NPG update (2.7) can also be written as [4, 27]

$$\theta^{k+1} = \theta^k + \eta \cdot w^k, \qquad \text{with} \quad w^k \in \operatorname*{argmin}_{w\in\mathbb{R}^d} \ L_{\nu_\rho^{\pi_\theta}}(w;\theta), \tag{2.8}$$

where $L_{\nu_\rho^{\pi_\theta}}(w;\theta)$ is the *compatible function approximation error* defined by

$$L_{\nu_\rho^{\pi_\theta}}(w;\theta) = \mathbb{E}_{(s,a)\sim\nu_\rho^{\pi_\theta}}\left[\left(A^{\pi_\theta}(s,a) - (1-\gamma)w^\top\nabla_\theta \log \pi_\theta(a \mid s)\right)^2\right]. \tag{2.9}$$

Here, $\nu_\rho^{\pi_\theta}(s,a) = d_\rho^{\pi_\theta}(s)\pi(a \mid s)$ is the *state-action* visitation measure induced by π_θ and initial state distribution ρ, which can also be written as

$$\nu_\rho^{\pi_\theta}(s,a) := (1-\gamma)\mathbb{E}_{s_0\sim\rho}\sum_{t=0}^{\infty}\gamma^t\mathbb{P}(s_t = s, a_t = a \mid s_0, \pi_\theta). \tag{2.10}$$

For convenience, we will denote $\nu_\rho^{\pi_\theta}$ by ν^{π_θ} hereafter. In other words, the NPG update direction w^k is given by the minimizer of a stochastic optimization problem. In practice, one obtains an approximate NPG update direction w^k by SGD (see Procedure 1).

Regarding the NPG update (2.8), we make the following standing assumption on the Fisher information matrix induced by π_θ and ρ.

Assumption 2.1. For all $\theta \in \mathbb{R}^{\mathrm{d}}$, the Fisher information matrix induced by policy π_θ and initial state distribution ρ satisfies

$$F_\rho(\theta) = \mathbb{E}_{(s,a)\sim\nu_\rho^{\pi_\theta}} \left[\nabla_\theta \log \pi_\theta(a \mid s) \nabla_\theta \log \pi_\theta(a \mid s)^\top\right] \succcurlyeq \mu_F \cdot I_d$$

for some constant $\mu_F > 0$.

Assumption 2.1 essentially states that $F_\rho(\theta)$ behaves well as a preconditioner in the NPG update (2.8). This is a common (and minimal) requirement for the convergence of preconditioned algorithms in both convex and nonconvex settings in the optimization realm, for example, the quasi-Newton algorithms [40, 41, 42, 43], and their stochastic variants [44, 45, 46, 47, 36]. In the RL realm, one common example of policy parametrizations that can satisfy this assumption is the Gaussian policy [2, 48, 19, 21], where $\pi_\theta(\cdot \mid s) = \mathcal{N}(\mu_\theta(s), \Sigma)$ with mean parametrized linearly as $\mu_\theta(s) = \phi(s)^\top\theta$, where $\phi(s)$ denotes some feature matrix of proper dimensions, θ is the coefficient vector, and $\Sigma \succ 0$ is some fixed covariance matrix. In this case, the Fisher information matrix at each s becomes $\phi(s)\Sigma^{-1}\phi(s)^\top$, independent of θ, and is uniformly lower bounded (positive definite sense) if $\phi(s)$ is full-row-rank, namely, the features expanded by θ are linearly independent, which is a common requirement for linear function approximation settings [49, 50, 51]. See App. B.2 for more detailed justifications, as well as discussions on more general policy parametrizations.

In the pioneering NPG work [4], $F(\theta)$ is directly assumed to be positive definite. So is in the follow-up works on natural actor-critic algorithms [12, 13]. In fact, this way, $F(\theta)$ will define a valid Riemannian metric on the parameter space, which has been used for interpreting the desired convergence properties of natural gradient methods [52, 53]. In a recent version of [27], a relevant assumption (specifically, Assumption 6.5, item 3) is made to establish the global convergence of NPG, in which it is assumed that $\lambda_{\min}(F_\rho(\theta))$ is not too small compared with the Fisher information matrix induced by a fixed comparator policy. this can be implied by our Assumption 2.1. To sum up, the positive definiteness on the Fisher preconditioning matrix is common and not very restrictive.

In Sec. 4, we shall see that under Assumption 2.1, the stationary convergence of NPG can be analyzed, and NPG enjoys a better sample complexity of $\mathcal{O}(\varepsilon^{-3})$ in terms of its global convergence, compared with the existing sample complexity of $\mathcal{O}(\varepsilon^{-4})$ in [27]. In addition, interestingly, PG and its variance-reduced version SRVR-PG also enjoy global convergence, although the Fisher information matrix does not appear explicitly in their updates.

3 Variance-Reduced Policy Gradient Methods

Recently, [21] proposes an algorithm called Stochastic Recursive Variance Reduced Policy Gradient (SRVR-PG, see Algorithm 2), which applies variance-reduction on PG. It achieves a sample complexity of $\mathcal{O}(\varepsilon^{-1.5})$ to find an $\varepsilon-$stationary point, compared with the $\mathcal{O}(\varepsilon^{-2})$ sample complexity of stochastic PG. However, it remains unclear whether SRVR-PG converges globally. In this work, we provide an affirmative answer to this question by showing that SRVR-PG has a sample complexity of $\mathcal{O}(\varepsilon^{-3})$ to find an $\varepsilon-$optimal policy, up to some compatible function approximation error due to policy parametrization.

We also propose a new algorithm called SRVR-NPG to incorporate variance reduction into NPG, which is described in Algorithm 1. In Sec. 4, we provide a sample complexity for its global convergence, which is comparable to our improved NPG result.

In line 8 of Algorithm 1, $g_w(\tau_j^H|\theta_{t-1}^{j+1})$ is a weighted gradient estimator given by

$$g_w(\tau_j^H \mid \theta_{t-1}^{j+1}) = \sum_{h=0}^{H-1} w_{0:h}(\tau_j^H \mid \theta_{t-1}^{j+1}, \theta_t^{j+1}) \left(\sum_{t=0}^{h} \nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i)\right) \left(\gamma^h r(s_h^i, a_h^i)\right), \tag{3.1}$$

Algorithm 1 Stochastic Recursive Variance Reduced Natural Policy Gradient (SRVR-NPG)

Input: number of epochs S, epoch size m, stepsize η, batch size N, minibatch size B, truncation horizon H, initial parameter $\theta_m^0 = \theta_0 \in \mathbb{R}^{\mathrm{d}}$.

1: **for** $j \leftarrow 0, ..., S-1$ **do**
2: $\quad \theta_0^{j+1} = \theta_m^j$;
3: $\quad$ Sample $\{\tau_i^H\}_{i=1}^N$ from $p_\rho^H(\cdot \,|\, \theta_0^{j+1})$ and calculate $u_0^{j+1} = \frac{1}{N}\sum_{i=1}^N g(\tau_i^H \,|\, \theta_0^{j+1})$;
4: $\quad w_0^{j+1} = \texttt{SRVR-NPG-SGD}(\nu^{\pi_{\theta_0^{j+1}}}, \pi_{\theta_0^{j+1}}, u_0^{j+1})$; $\qquad \triangleright\ w_0^{j+1} \approx w_{0,\star}^{j+1} = F_\rho^{-1}(\theta_0^{j+1})u_0^{j+1}$;
5: $\quad \theta_1^{j+1} = \theta_0^{j+1} + \eta w_0^{j+1}$;
6: $\quad$ **for** $t \leftarrow 1, ..., m-1$ **do**
7: $\qquad$ Sample B trajectories $\{\tau_j^H\}_{j=1}^B$ from $p_\rho^H(\cdot|\theta_t^{j+1})$;
8: $\qquad u_t^{j+1} = u_{t-1}^{j+1} + \frac{1}{B}\sum_{j=1}^B \Big(g(\tau_j^H \,|\, \theta_t^{j+1}) - g_w(\tau_j^H \,|\, \theta_{t-1}^{j+1})\Big)$;
9: $\qquad w_t^{j+1} = \texttt{SRVR-NPG-SGD}(\nu^{\pi_{\theta_t^{j+1}}}, \pi_{\theta_t^{j+1}}, u_t^{j+1})$; $\qquad \triangleright\ w_t^{j+1} \approx w_{t,\star}^{j+1} = F_\rho^{-1}(\theta_t^{j+1})u_t^{j+1}$;
10: $\qquad \theta_{t+1}^{j+1} = \theta_t^{j+1} + \eta w_t^{j+1}$;
11: $\quad$ **end for**
12: **end for**
13: **return** θ_{out} chosen uniformly from $\{\theta\}_{j=1,...,S;t=0,...,m-1}$.

where the importance weight factor $w_{0:h}(\tau_j^H|\theta_{t-1}^{j+1}, \theta_t^{j+1})$ is defined by

$$w_{0:h}(\tau_j^H \,|\, \theta_{t-1}^{j+1}, \theta_t^{j+1}) = \prod_{h'=0}^{h} \frac{\pi_{\theta_{t-1}^{j+1}}(a_{h'} \,|\, s_{h'})}{\pi_{\theta_t^{j+1}}(a_{h'} \,|\, s_{h'})}. \tag{3.2}$$

This importance sampling makes u_t^{j+1} an unbiased estimator of $\nabla J^H(\theta_t^{j+1})$.

In lines 4 and 8 of Algorithm 1, w_t^{j+1} is produced by `SRVR-NPG-SGD` (see Procedure 2), which applies SGD[1] to solve the following subproblem:

$$w_t^{j+1} \approx \operatorname*{argmin}_w \left\{ \mathbb{E}_{(s,a)\sim\nu_t^{j+1}} \left[\Big(w^T \nabla_\theta \log \pi_{\theta_t^{j+1}}(a \,|\, s) \Big)^2 \right] - 2\langle w, u_t^{j+1} \rangle \right\}, \tag{3.3}$$

where ν_t^{j+1} is the state-action visitation measure induced by $\pi_{\theta_t^{j+1}}$. The exact update direction given by (3.3) is $F_\rho^{-1}(\theta_t^{j+1})u_t^{j+1}$, and as in NPG, $F_\rho(\theta_t^{j+1})$ also serves as a preconditioner.

4 Theoretical Results

Before presenting the global convergence results, we first introduce some standard assumptions.

Assumption 4.1. The truncated GPOMDP estimator $g(\tau^H \,|\, \theta)$ defined in (2.6) satisfies $\mathrm{Var}\big(g(\tau^H \,|\, \theta)\big) := \mathbb{E}[\|g(\tau^H \,|\, \theta) - \mathbb{E}[g(\tau^H \,|\, \theta)]\|^2] \le \sigma^2$ for any θ and $\tau^H \sim p_\rho^H(\cdot \,|\, \theta)$.

Assumption 4.2.

1. $\|\nabla_\theta \log \pi_\theta(a \,|\, s)\| \le G$ for any θ and $(s,a) \in \mathcal{S} \times \mathcal{A}$.
2. $\|\nabla_\theta \log \pi_{\theta_1}(a \,|\, s) - \nabla_\theta \log \pi_{\theta_2}(a \,|\, s)\| \le M\|\theta_1 - \theta_2\|$ for any θ_1, θ_2 and $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Assumption 4.3. For the importance weight $w_{0:h}(\tau^H|\theta_1, \theta_2)$ (3.2), there exists $W > 0$ such that

$$\mathrm{Var}(w_{0:h}\left(\tau^H \,|\, \theta_1, \theta_2\right)) \le W, \;\; \forall \theta_1, \theta_2 \in \mathbb{R}^{\mathrm{d}}, \tau^H \sim p_\rho^H(\cdot \,|\, \theta_2).$$

Assumptions 4.1, 4.2 and 4.3 are standard in the analysis of PG methods and their variance reduced variants [27, 19, 20, 21]. They can be verified for simple policy parametrizations such as Gaussian policies; see [19, 57, 58] for more justifications.

Following the Assumption 6.5 of [27], we assume that the policy parametrization π_θ achieves a good function approximation, as measured by the *transferred compatible function approximation error*.

[1] Following [27], we apply SGD [54] to make a fair comparison. One can also apply the SA algorithm [55] and AC-SA algorithm [56].

Assumption 4.4. For any $\theta \in \mathbb{R}^{\mathrm{d}}$, the *transferred compatible function approximation error* satisfies

$$L_{\nu^\star}(w_\star^\theta;\theta) = \mathbb{E}_{(s,a)\sim\nu^\star}\left[\left(A^{\pi_\theta}(s,a) - (1-\gamma)(w_\star^\theta)^\top \nabla_\theta \log \pi_\theta(a\,|\,s)\right)^2\right] \le \varepsilon_{\text{bias}}, \tag{4.1}$$

where $\nu^\star(s,a) = d_\rho^{\pi^\star}(s)\cdot\pi^\star(a\,|\,s)$ is the state-action distribution induced by an optimal policy $\pi^\star$ that maximizes $J(\pi)$, and $w_\star^\theta = \operatorname{argmin}_{w\in\mathbb{R}^{\mathrm{d}}} L_{\nu_\rho^{\pi_\theta}}(w;\theta)$ is the exact NPG update direction at θ.

$\varepsilon_{\text{bias}}$ reflects the error when approximating the advantage function from the score function, it measures the capacity of the parametrization π_θ. When π_θ is the softmax parametrization, we have $\varepsilon_{\text{bias}} = 0$ [27]. When π_θ is a restricted parametrization, $\varepsilon_{\text{bias}}$ is often positive as π_θ may not contain all stochastic policies. For rich neural parametrizations, $\varepsilon_{\text{bias}}$ is very small [25].

4.1 A General Framework for Global Convergence

Inspired by the global convergence analysis of NPG in [27], we present a general framework that relates the global convergence rates of these algorithms to i) their stationary convergence rate on $J(\theta)$, and ii) the difference between their update directions and exact NPG update directions.

Proposition 4.5. Let $\{\theta^k\}_{k=1}^K$ be generated by a general update of the form

$$\theta^{k+1} = \theta^k + \eta w^k, \quad k = 0, 1, ...K-1.$$

Furthermore, let $w_\star^k = F_\rho^{-1}(\theta^k)\nabla J(\theta^k)$ be the exact NPG update direction at θ^k. Then, we have

$$\begin{aligned} J(\pi^\star) - \frac{1}{K}\sum_{k=0}^{K-1} J(\theta^k) \le \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \frac{1}{\eta K}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right] \\ + \frac{M\eta}{2K}\sum_{k=0}^{K-1}\|w^k\|^2 + \frac{G}{K}\sum_{k=0}^{K-1}\|w^k - w_\star^k\|, \end{aligned} \tag{4.2}$$

where $\pi^\star$ is an optimal policy that maximizes $J(\pi)$.

The detailed proof of this global convergence framework can be found in J. To obtain a high level idea, one first starts from the $M-$smoothness of the score function to get

$$\begin{aligned} &\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^k}(\cdot\,|\,s)\right) - \mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^{k+1}}(\cdot\,|\,s)\right)\right] \\ &\ge \eta\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\cdot w_\star^k] \\ &\quad + \eta\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\cdot(w^k - w_\star^k)] - \frac{M\eta^2}{2}\|w^k\|^2. \end{aligned}$$

On the other hand, the renowned Performance Difference Lemma [59] tells us that

$$\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[A^{\pi_{\theta^k}}(s,a)] = (1-\gamma)\left(J^\star - J(\theta^k)\right).$$

To connect the advantage term $\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[A^{\pi_{\theta^k}}(s,a)]$ with the inner product term $\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\cdot w_\star^k]$, we invoke Assumption 4.4:

$$\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(A^{\pi_\theta}(s,a) - (1-\gamma)(w_\star^\theta)^\top\nabla_\theta \log \pi_\theta(a\,|\,s)\right)^2\right] \le \varepsilon_{\text{bias}}, \quad \text{for any } \theta \in \mathbb{R}^{\mathrm{d}}.$$

The final result follows from a telescoping sum on $k = 0, 1, ..., K-1$.

Several remarks are in order. The first term on the right-hand side of (4.2) reflects the function approximation error due to the parametrization π_θ, and the second term is of the form $\mathcal{O}(\frac{1}{K})$. The third term depends on the stationary convergence. With Assumption 2.1, it can be shown that[1] $\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k\|^2] \to 0$ for both NPG and SRVR-NPG. The proof follows from an optimization perspective and is inspired by the stationary convergence analysis of stochastic PG (see App. E).

With Assumption 2.1, we can also show that the last term of (4.2) is small. Take stochastic PG as an example; then, we have $w^k = \frac{1}{N}\sum_{i=1}^N g(\tau_i^H|\theta^k)$, and

$$\frac{1}{K}\sum_{k=0}^{K-1}\|w^k - w_\star^k\| \le \frac{1}{K}\sum_{k=0}^{K-1}\|w^k - \nabla J(\theta^k)\| + \frac{1}{K}\sum_{k=0}^{K-1}\left(1 + \frac{1}{\mu_F}\right)\|\nabla J(\theta^k)\|.$$

[1]The stationary convergence of SRVR-PG has been established in [21].

When H and N are large enough, w^k is a low-variance estimator of $\nabla J^H(\theta^k)$, and $\nabla J^H(\theta^k)$ is close to $\nabla J(\theta^k)$, this makes the first term above small. The second term also goes to 0 as θ^k approaches stationarity.

4.2 Global Convergence Results

By applying Proposition 4.5 on the PG, NPG, SRVR-PG, and SRVR-NPG updates and analyzing their stationary convergence, we obtain their global convergence rates. In the following, we only keep the dependences on σ^2 (the variance of the gradient estimator), W (variance of importance weight), $\frac{1}{1-\gamma}$ (the effective horizon) and ε (target accuracy). The specific choice of the parameters and sample complexities, as well as the proof, can be found in the appendix.

Theorem 4.6. In the stochastic PG (2.4) with the truncated GPOMDP estimator (2.6), take $\eta = \frac{1}{4L_J}$, $K = \mathcal{O}\left(\frac{1}{(1-\gamma)^2\varepsilon^2}\right)$, $N = \mathcal{O}\left(\frac{\sigma^2}{\varepsilon^2}\right)$, and $H = \mathcal{O}\left(\log(\frac{1}{(1-\gamma)\varepsilon})\right)$. Then, we have

$$J(\pi^\star) - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[J(\theta^k)] \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \varepsilon.$$

In total, stochastic PG samples $\mathcal{O}\left(\frac{\sigma^2}{(1-\gamma)^2\varepsilon^4}\right)$ trajectories.

Remark 4.7. $L_J = \frac{MR}{(1-\gamma)^2}$ is the Lipschitz constant of ∇J, see Lemma B.1 for details.

Remark 4.8. Theorem 4.6 improves the result of [27, Thm. 6.11] from (impractical) full gradients to sample-based stochastic gradients.

Theorem 4.9. In the NPG update (2.8), let us apply $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon^2}\right)$ iterations of SGD as in Procedure 1 to obtain an update direction. In addition, take $\eta = \frac{\mu_F^2}{4G^2L_J}$ and $K = \mathcal{O}\left(\frac{1}{(1-\gamma)^2\varepsilon}\right)$. Then,

$$J^\star - \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[J(\theta^k)] \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \varepsilon.$$

In total, NPG samples $\mathcal{O}\left(\frac{1}{(1-\gamma)^6\varepsilon^3}\right)$ trajectories.

Remark 4.10. Compared with [27, Coro. 6.10], Theorem 4.9 improves the sample complexity of NPG by $\mathcal{O}(\varepsilon^{-1})$. This is because our stationary convergence analysis on NPG allows for a constant stepsize η, while [27, Coro. 6.10] applies a stepsize of $\eta = \mathcal{O}(1/\sqrt{K})$. It is worth noting that the $\mathcal{O}(\sqrt{\varepsilon_{\text{bias}}})$ term is the same as in [27], and we also apply the average SGD [54] to solve the NPG subproblem (2.8).

Theorem 4.11. In SRVR-PG (Algorithm 2), take $\eta = \frac{1}{8L_J}$, $S = \mathcal{O}\left(\frac{1}{(1-\gamma)^{2.5}\varepsilon}\right)$, $m = \mathcal{O}\left(\frac{(1-\gamma)^{0.5}}{\varepsilon}\right)$, $B = \mathcal{O}\left(\frac{W}{(1-\gamma)^{0.5}\varepsilon}\right)$, $N = \mathcal{O}\left(\frac{\sigma^2}{\varepsilon}\right)$, and $H = \mathcal{O}\left(\log(\frac{1}{(1-\gamma)\varepsilon})\right)$. Then, we have

$$J^\star - \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[J(\theta_t^{j+1})] \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \varepsilon.$$

In total, SRVR-PG samples $\mathcal{O}\left(\frac{W+\sigma^2}{(1-\gamma)^{2.5}\varepsilon^3}\right)$ trajectories.

Remark 4.12. Theorem 4.11 establishes the global convergence of SRVR-PG proposed in [21], where only stationary convergence is shown. Also, compared with stochastic PG, SRVR-PG enjoys a better sample complexity thanks to its faster stationary convergence.

Theorem 4.13. In SRVR-NPG (Algorithm 1), let us apply $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon^2}\right)$ iterations of SGD as in Procedure 2 to obtain an update direction. In addition, take $\eta = \frac{\mu_F}{16L_J}$, $S = \mathcal{O}\left(\frac{1}{(1-\gamma)^{2.5}\varepsilon^{0.5}}\right)$, $m = \mathcal{O}\left(\frac{(1-\gamma)^{0.5}}{\varepsilon^{0.5}}\right)$, $B = \mathcal{O}\left(\frac{W}{(1-\gamma)^{0.5}\varepsilon^{1.5}}\right)$, $N = \mathcal{O}\left(\frac{\sigma^2}{\varepsilon^2}\right)$, and $H = \mathcal{O}\left(\log(\frac{1}{(1-\gamma)\varepsilon})\right)$. Then,

$$J^\star - \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[J(\theta_t^{j+1})] \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \varepsilon.$$

In total, SRVR-NPG samples $\mathcal{O}\left(\frac{W+\sigma^2}{(1-\gamma)^{2.5}\varepsilon^{2.5}} + \frac{1}{(1-\gamma)^6\varepsilon^3}\right)$ trajectories.

Remark 4.14. Compared with SRVR-PG, our SRVR-NPG has a better dependence on W and σ^2, which could be large in practice (especially W). The current sample complexity of SRVR-NPG is not better than our (improved) result of NPG since, in our analysis, the advantage of variance reduction is offset by the cost of solving the subproblems.

5 Numerical Experiments

In this section, we compare the numerical performances of stochastic PG, NPG, SRVR-PG, and SRVR-NPG. Specifically, we test on benchmark reinforcement learning environments Cartpole and Mountain Car. Our implementation is based on the implementation of SRVPG[1] and SRVR-PG[2], and can be found in the supplementary material.

For both tasks, we apply a Gaussian policy of the form $\pi_\theta(a \mid s) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(\mu_\theta(s)-a)^2}{2\sigma^2}\right)$ where the mean $\mu_\theta(s)$ is modeled by a neural network with Tanh as the activation function.

For the Cartpole problem, we apply a neural network of size 32×1 and a horizon of $H = 100$. In addition, each training algorithm uses 5000 trajectories in total. For the Mountain Car problem, we apply a neural network of size 64×1 and take $H = 1000$. 3000 trajectories are allowed for each algorithm. The numerical performance comparison, as well as the settings of algorithm-specific parameters, can be found in Figures 1 and 2. In App. O, we provide more implementation details.

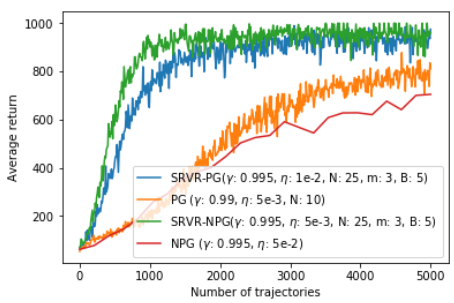

Figure 1: Numerical Performances on Cartpole. For PG, SRVR-PG and SRVR-NPG, we report the undiscounted average return averaged over 10 runs. For NPG, we report the averaged return over 40 runs. Overall, SRVR-NPG has the best performance.

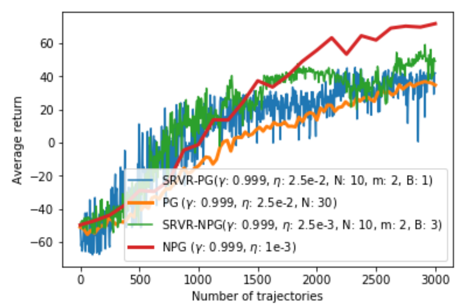

Figure 2: Numerical Performances on Mountain Car. For PG, SRVR-PG and SRVR-NPG, we report the undiscounted average return averaged over 10 runs. For NPG, we report the averaged return over 40 runs. Overall, NPG has the best performance.

6 Concluding Remarks

In this work, we have introduced a framework for analyzing the global convergence of (natural) PG methods and their variance-reduced variants, under the assumption that the Fisher information matrix is positive definite. We have established the sample complexity for the global convergence of stochastic PG and its variance-reduced variant SRVR-PG, and improved the sample complexity of NPG. In addition, we have introduced SRVR-NPG, which incorporates variance-reduction into NPG, and enjoys both global convergence guarantee and an efficient sample complexity. Our improved analysis hinges on exploiting the advantages of previous analyses on (variance reduced) PG and NPG methods, which may be of independent interest, and can be used to design faster variance-reduced NPG methods in the future.

[1] `https://github.com/Dam930/rllab`

[2] `https://github.com/xgfelicia/SRVRPG`

Broader Impact

The results of this paper improves the performance of policy-gradient methods for reinforcement learning, as well as our understanding to the existing methods. Through reinforcement learning, our study will also benefit several research communities such as machine learning and robotics. We do not believe that the results in this work will cause any ethical issue, or put anyone at a disadvantage in our society.

Acknowledgements

Yanli Liu and Wotao Yin were partially supported by the Office of Naval Research (ONR) Grant N000141712162. Yanli Liu was also supported by UCLA Dissertation Year Fellowship. Kaiqing Zhang and Tamer Başar were supported in part by the US Army Research Laboratory (ARL) Cooperative Agreement W911NF-17-2-0196, and in part by the Office of Naval Research (ONR) MURI Grant N00014-16-1-2710.

References

[1] Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT press, 2018.

[2] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256, 1992.

[3] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063, 2000.

[4] Sham M Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, pages 1531–1538, 2002.

[5] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[6] 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. *arXiv preprint arXiv:1509.02971*, 2015.

[7] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1928–1937, 2016.

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

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

[10] 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. *arXiv preprint arXiv:1803.07246*, 2018.

[11] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in Neural Information Processing Systems*, pages 1008–1014, 2000.

[12] Jan Peters and Stefan Schaal. Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190, 2008.

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

[14] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[15] Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707, 2016.

[16] Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pages 314–323, 2016.

[17] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014.

[18] Tianbing Xu, Qiang Liu, and Jian Peng. Stochastic variance reduction for policy gradient estimation. *arXiv preprint arXiv:1710.06034*, 2017.

[19] Matteo Papini, Damiano Binaghi, Giuseppe Canonaco, Matteo Pirotta, and Marcello Restelli. Stochastic variance-reduced policy gradient. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4026–4035, 2018.

[20] Pan Xu, Felicia Gao, and Quanquan Gu. An improved convergence analysis of stochastic variance-reduced policy gradient. *arXiv preprint arXiv:1905.12615*, 2019.

[21] Pan Xu, Felicia Gao, and Quanquan Gu. Sample efficient policy gradient methods with recursive variance reduction. In *International Conference on Learning Representations*, 2020.

[22] Huizhuo Yuan, Xiangru Lian, Ji Liu, and Yuren Zhou. Stochastic recursive momentum for policy gradient methods. *arXiv preprint arXiv:2003.04302*, 2020.

[23] Maryam Fazel, Rong Ge, Sham M Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, 2018.

[24] Kaiqing Zhang, Alec Koppel, Hao Zhu, and Tamer Başar. Global convergence of policy gradient methods to (almost) locally optimal policies. *arXiv preprint arXiv:1906.08383*, 2019.

[25] Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

[26] Jalaj Bhandari and Daniel Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

[27] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *arXiv preprint arXiv:1908.00261v5*, 2019.

[28] Boyi Liu, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Neural trust region/proximal policy optimization attains globally optimal policy. In *Advances in Neural Information Processing Systems*, pages 10564–10575, 2019.

[29] Lior Shani, Yonathan Efroni, and Shie Mannor. Adaptive trust region policy optimization: Global convergence and faster rates for regularized MDPs. *arXiv preprint arXiv:1909.02769*, 2019.

[30] Tengyu Xu, Zhe Wang, and Yingbin Liang. Improving sample complexity bounds for actor-critic algorithms. *arXiv preprint arXiv:2004.12956*, 2020.

[31] Tengyu Xu, Zhe Wang, and Yingbin Liang. Non-asymptotic convergence analysis of two time-scale (natural) actor-critic algorithms. *arXiv preprint arXiv:2005.03557*, 2020.

[32] Yue Wu, Weitong Zhang, Pan Xu, and Quanquan Gu. A finite time analysis of two time-scale actor critic methods. *arXiv preprint arXiv:2005.01350*, 2020.

[33] Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale framework for bilevel optimization: Complexity analysis and application to actor-critic. *arXiv preprint arXiv:2007.05170*, 2020.

[34] Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2613–2621. JMLR. org, 2017.

[35] Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

[36] Yanli Liu, Fei Feng, and Wotao Yin. Acceleration of svrg and katyusha x by inexact preconditioning. In *International Conference on Machine Learning*, pages 4003–4012, 2019.

[37] Nhan H Pham, Lam M Nguyen, Dzung T Phan, Phuong Ha Nguyen, Marten van Dijk, and Quoc Tran-Dinh. A hybrid stochastic policy gradient algorithm for reinforcement learning. *arXiv preprint arXiv:2003.00430*, 2020.

[38] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In *Thirty-seventh International Conference on Machine Learning (ICML 2020)*, 2020.

[39] Richard S Sutton, Andrew G Barto, and Ronald J Williams. Reinforcement learning is direct adaptive optimal control. *IEEE Control Systems Magazine*, 12(2):19–22, 1992.

[40] Charles George Broyden. The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90, 1970.

[41] Roger Fletcher. A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322, 1970.

[42] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26, 1970.

[43] David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.

[44] Richard H Byrd, Samantha L Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

[45] Philipp Moritz, Robert Nishihara, and Michael Jordan. A linearly-convergent stochastic l-bfgs algorithm. In *Artificial Intelligence and Statistics*, pages 249–258, 2016.

[46] Robert Gower, Donald Goldfarb, and Peter Richtárik. Stochastic block bfgs: Squeezing more curvature out of data. In *International Conference on Machine Learning*, pages 1869–1878, 2016.

[47] Xiao Wang, Shiqian Ma, Donald Goldfarb, and Wei Liu. Stochastic quasi-newton methods for nonconvex stochastic optimization. *SIAM Journal on Optimization*, 27(2):927–956, 2017.

[48] Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

[49] John N Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1075–1081, 1997.

[50] Francisco S Melo, Sean P Meyn, and M Isabel Ribeiro. An analysis of reinforcement learning with function approximation. In *International Conference on Machine Learning*, pages 664–671, 2008.

[51] Richard S Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *International Conference on Machine Learning*, pages 993–1000, 2009.

[52] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[53] James Martens. New insights and perspectives on the natural gradient method. *arXiv preprint arXiv:1412.1193*, 2014.

[54] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

[55] 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.

[56] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[57] Matteo Pirotta, Marcello Restelli, and Luca Bascetta. Adaptive step-size for policy gradient methods. In *Advances in Neural Information Processing Systems*, pages 1394–1402, 2013.

[58] Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Advances in Neural Information Processing Systems*, pages 442–450, 2010.

[59] Sham M Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *International Conference on Machine Learning*, volume 2, pages 267–274, 2002.

[60] Anirban DasGupta. The exponential family and statistical applications. In *Probability for Statistics and Machine Learning*, pages 583–612. Springer, 2011.

[61] Solomon Kullback. *Information Theory and Statistics*. Courier Corporation, 1997.

[62] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

[63] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Supplementary Materials for "An Improved Analysis of (Variance-Reduced) Policy Gradient and Natural Policy Gradient Methods"

A Derivation of Previous Complexity Bounds

In this section, we briefly explain how to derive the sample complexities bounds in the first line of Table 1.

In the most recent version of [27], a complexity bound of $\mathcal{O}(\varepsilon^{-6})$ can be obtained the taking $N = \mathcal{O}(\varepsilon^{-4})$ and $N = \mathcal{O}(\varepsilon^{-2})$ in its Corollary 6.2. Note this complexity bound can be improved to $\mathcal{O}(\varepsilon^{-4})$ if a uniform upper bound for exact NPG update directions is applied. In this case, one can apply the convergence bound of SGD instead of Projected SGD for the NPG subproblem. In this paper, we establish an upper bound for $\|\nabla J(\theta)\|$ in Lemma B.1. Therefore the exact NPG update direction is also upper bounded thanks to Assumption 2.1.

For [25], the sample complexity bound of $\mathcal{O}(T_{TD}\varepsilon^{-2})$ is achieved by its Theorem 4.13. To be specific, one takes $T = \mathcal{O}(\varepsilon^{-2})$ and $T_{\text{TD}} = \mathcal{O}(m)$ number of temporal difference updates at each iteration. Here, m is width of the neural network.

Note that in the proof of its Corollary 4.14, we can choose $m = \mathcal{O}(T^4)$ (instead of $\mathcal{O}(T^6)$) to have a convergence bound of the form $\mathcal{O}(\sqrt{\varepsilon_0}) + \varepsilon$ (instead of $\mathcal{O}(\varepsilon)$), which is similar to our $\frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \varepsilon$ convergence bound.

For [28], by the Corollary 4.10 therein, one needs to take $K = \mathcal{O}(\varepsilon^{-2})$ and $T = \mathcal{O}(K^3) = \mathcal{O}(\varepsilon^{-6})$, which results in a total sample complexity of $\mathcal{O}(\varepsilon^{-8})$.

For [29], its Theorem 5 (item 1) gives a sample complexity of $\sum_{k=1}^{N} M_k = \mathcal{O}(\varepsilon^{-4})$, where we have applied $N = \mathcal{O}(\varepsilon^{-2})$ and $M_k = \mathcal{O}(\varepsilon^{-2})$.

B Helper Lemmas

In this section, we lay out several results that will be useful in later analyses and proofs.

B.1 Properties of PG Estimator

First, for any $H > 1$, we define the H-horizon truncated versions of the return $J(\theta)$ as

$$J^H(\theta) := \mathbb{E}_{s_0 \sim \rho}\left(\sum_{t=0}^{H-1} \gamma^t r_t\right), \tag{B.1}$$

where the expectation is taken over the trajectories, starting from the state distribution ρ. Now we establish several properties of the GPOMDP policy gradient estimators and the return functions.

Lemma B.1. Recall the GPOMDP policy gradient estimate given in (2.5). The following properties hold:

- If the infinite-sum in (2.5) is well defined, $g(\tau_i \,|\, \theta)$ in (2.5) is an unbiased estimate of the PG $\nabla J(\theta)$. Similarly, the truncated GPOMDP estimate $g(\tau_i^H \,|\, \theta)$ given by (2.6) is an unbiased estimate of the PG $\nabla J^H(\theta)$.
- $J(\theta), J^H(\theta)$ are L_J-smooth, where $L_J = \frac{MR}{(1-\gamma)^2}$. Furthermore, we have $\max\left\{\|\nabla J(\theta)\|, \|\nabla J^H(\theta)\|\right\} \le \frac{GR}{(1-\gamma)^2}$.
- We also have $\|\nabla J^H(\theta) - \nabla J(\theta)\| \le GR\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)\gamma^H$.

Proof. the unbiasedness of $g(\tau_i \,|\, \theta)$ follows directly from [9]. A similar decomposition can also be done for its truncated version $g(\tau_i^H \,|\, \theta)$.

For the second argument, the smoothness proof is similar to that of Proposition 4.2 in [21]. Specifically, we have know from their Proposition 4.2 that

$$\|\nabla g(\tau_i|\theta)\|_2 \leq \frac{MR}{(1-\gamma)^2}, \qquad \|\nabla g(\tau_i^H|\theta)\|_2 \leq \frac{MR}{(1-\gamma)^2}.$$

Due to the unbiasedness of $g(\tau_i|\theta)$ and $g(\tau_i^H|\theta)$ for estimating $\nabla J(\theta)$ and $\nabla J^H(\theta)$, respectively, by applying $\mathbb{E}[\|\xi\|^2] \geq \|\mathbb{E}[\xi]\|^2$, we know that both $J(\theta)$ and $J^H(\theta)$ are L_J-smooth. For obtaining the boundedness of the gradients, similar arguments also apply.

For the third argument, one can calculate that

$$\begin{aligned}
\|g(\tau_i^H \mid \theta) - g(\tau_i \mid \theta)\| &= \left\| \sum_{h=H}^{\infty} \left(\sum_{t=0}^{h} \nabla_\theta \log \pi_\theta(a_t^i \mid s_t^i) \right) \left(\gamma^h r(s_h^i, a_h^i) \right) \right\| \\
&\leq \|GR \sum_{h=H}^{\infty} (h+1)\gamma^h\| \\
&= GR \left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2} \right) \gamma^H
\end{aligned}$$

This rest of the proof follows from the unbiasedness of $g(\tau_i|\theta)$ and $g(\tau_i^H|\theta)$ for estimating $\nabla J(\theta)$ and $\nabla J^H(\theta)$, respectively. □

B.2 On the Positive Definiteness of $F_\rho(\theta)$

Now we remark that the positive definiteness on the Fisher information matrix induced by π_θ, as stated in Assumption 2.1, is not restricted. Assumption 2.1 essentially states that $F(\theta)$ behaves well as a preconditioner in the NPG update (2.8). This is a common (and minimal) requirement for the convergence of preconditioned algorithms in both convex and nonconvex settings in the optimization realm [44, 45, 46, 47, 36].

In the RL realm, one common example of policy parametrizations that can satisfy this assumption is the Gaussian policy [2, 48, 19, 21], where $\pi_\theta(\cdot \mid s) = \mathcal{N}(\mu_\theta(s), \Sigma)$ with mean parametrized linearly as $\mu_\theta(s) = \phi(s)^\top \theta$, where $\phi(s)$ denotes some feature matrix of proper dimensions, θ is the coefficient vector, and $\Sigma \succ 0$ is some fixed covariance matrix. Suppose the action $a \in \mathcal{A} \subseteq \mathbb{R}^A$ and recall $\theta \in \mathbb{R}^d$. Thus, $\phi(s) \in \mathbb{R}^{d\times A}$. In this case, the Fisher information at each s becomes $\phi(s)\Sigma^{-1}\phi(s)^\top$, independent of θ, and is positive definite if $\phi(s)$ is full-row-rank. For the case $d < A$, which is usually the case as a lower-dimensional (than a) parameter θ is used, this can be achieved by designing the rows of $\phi(s)$ to be linearly independent, a common requirement for linear function approximation settings [49, 50, 51].

For $\mu_\theta(s)$ being nonlinear functions of θ, e.g., neural networks, the positive definiteness can still be satisfied, if the Jacobian of $\mu_\theta(s)$ at all θ uniformly satisfies the aforementioned conditions of $\phi(s)$ (the Jacobian in the linear case). In addition, beyond Gaussian policies, with the same conditions mentioned above on the feature $\phi(s)$ or the Jacobian of $\mu_\theta(s)$, Assumption 2.1 also holds more generally for any full-rank exponential family parametrization with mean parametrized by $\mu_\theta(s)$, as the Fisher information matrix, in this case, is also positive definite, in replace of the covariance matrix $\phi(s)\Sigma^{-1}\phi(s)$ in the Gaussian case [60].

Indeed, the Fisher information matrix is positive definite for any *regular* statistical model [61]. In the pioneering NPG work [4], $F(\theta)$ is directly assumed to be positive definite. So is in the follow-up works on natural actor-critic algorithms [12, 13]. In fact, this way, $F_\rho(\theta)$ will define a valid Riemannian metric on the parameter space, which has been used for interpreting the desired convergence properties of natural gradient methods [52, 53]. In sum, the positive definiteness on the Fisher preconditioning matrix is common and not restrictive.

C SGD and Sampling Procedures

C.1 SGD for Solving the Subproblems of NPG and SRVR-NPG

Similar to the Algorithm 1 of [27], we also apply the averaged SGD algorithm as in [54] to solve the subproblems of NPG and SRVR-NPG.

Procedure 1 NPG-SGD

Input: number of iterations T, stepsize $\alpha > 0$, objective function $l(w)$, initialization $w_0 = 0$.

1: **for** $t \leftarrow 0, ..., T-1$ **do**
2: $\quad w_{t+1} = w_t - \alpha \widetilde{\nabla} l(w_t);$ $\triangleright$ $l(w)$ is defined in (C.1), $\widetilde{\nabla} l(w_t)$ is defined in (C.2).
3: **end for**
4: **return** $w_{\text{out}} = \frac{1}{T}\sum_{t=1}^{T} w_t$.

Procedure 2 SRVR-NPG-SGD

Input: number of iterations T, stepsize $\alpha > 0$, objective function $l(w)$, initialization $w_0 = 0$.

1: **for** $t \leftarrow 0, ..., T-1$ **do**
2: $\quad w_{t+1} = w_t - \alpha \widetilde{\nabla} l(w_t);$ $\triangleright$ $l(w)$ is defined in (C.3), $\widetilde{\nabla} l(w_t)$ is defined in (C.4).
3: **end for**
4: **return** $w_{\text{out}} = \frac{1}{T}\sum_{t=1}^{T} w_t$.

For NPG, its subproblem (2.8) is of the form

$$w^k \in \underset{w \in \mathbb{R}^{\mathrm{d}}}{\operatorname{argmin}}\, L_{\nu^{\pi_{\theta^k}}}(w; \theta^k) = \mathbb{E}_{(s,a)\sim \nu^{\pi_{\theta^k}}}\left[\left(A^{\pi_{\theta^k}}(s,a) - (1-\gamma) w^\top \nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\right)^2\right],$$

where

$$\nu^{\pi_{\theta^k}}(s,a) = (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim\rho}\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s, a_t = a\,|\,s_0, a_0, \pi_{\theta^k}).$$

In Procedure 1, let us set

$$l(w) = \frac{1}{2(1-\gamma)^2} L_{\nu^{\pi_{\theta^k}}}(w; \theta^k). \tag{C.1}$$

Then, we can obtain a stochastic gradient at w_t by

$$\widetilde{\nabla} l(w_t) = \left((w_t)^T \nabla_\theta \log \pi_{\theta^k}(a|s) - \frac{1}{1-\gamma}\widehat{A}^{\pi_{\theta^k}}(s,a)\right)\nabla_\theta \log \pi_{\theta^k}(a|s) \tag{C.2}$$

where $(s,a) \sim \nu^{\pi_{\theta^k}}$, and $\widehat{A}^{\pi_{\theta^k}}(s,a)$ is an unbiased estimate of $A^{\pi_{\theta^k}}(s,a)$. We will describe how to obtain $(s,a) \sim \nu^{\pi_{\theta^k}}$ and $\widehat{A}^{\pi_{\theta^k}}(s,a)$ in App. C.2.

Following Corollary 6.10 of [27], we can verify that $\widetilde{\nabla} l(w_t)$ is an unbiased estimate of $\nabla l(w_t)$.

For SRVR-NPG, its subproblem (3.3) is of the form

$$w_t^{j+1} \approx \underset{w}{\operatorname{argmin}}\left\{\mathbb{E}_{(s,a)\sim \nu^{\pi_{\theta_t^{j+1}}}}\left[\left(w^T \nabla_\theta \log \pi_{\theta_t^{j+1}}(a\,|\,s)\right)^2\right] - 2\langle w, u_t^{j+1}\rangle\right\},$$

where

$$\nu^{\pi_{\theta_t^{j+1}}}(s,a) = (1-\gamma)\mathbb{E}_{(s_0,a_0)\sim\rho}\sum_{t=0}^{\infty}\gamma^t \mathbb{P}(s_t = s, a_t = a\,|\,s_0, a_0, \pi_{\theta_t^{j+1}}).$$

In Procedure 2, let us set

$$l(w) = \frac{1}{2}\left(\mathbb{E}_{(s,a)\sim \nu^{\pi_{\theta_t^{j+1}}}}\left[\left(w^T \nabla_\theta \log \pi_{\theta_t^{j+1}}(a\,|\,s)\right)^2\right] - 2\langle w, u_t^{j+1}\rangle\right). \tag{C.3}$$

Then, a stochastic gradient $\widetilde{\nabla} l(w_t)$ is given by

$$\widetilde{\nabla} l(w_t) = \left((w_t)^T \nabla_\theta \log \pi_{\theta_t^{j+1}}(a|s)\right)\nabla_\theta \log \pi_{\theta_t^{j+1}}(a|s) - u_t^{j+1}. \tag{C.4}$$

where $(s,a) \sim \nu^{\pi_{\theta_t^{j+1}}}$ is obtained in a similar way as above. It is straightforward to verify that $\widetilde{\nabla} l(w_t)$ is an unbiased estimate of $\nabla l(w_t)$.

C.2 Sampling Procedures

Sampling $(s,a) \sim \nu^{\pi_{\theta^k}}$ and Obtaining $\widehat{A}^{\pi_{\theta^k}}(s,a)$ can be done in a standard way, for example, by apply Algorithm 3 of [27]. Both of them needs to sample $\frac{1}{1-\gamma}$ state-action pairs in expectation.

D SRVR-PG Algorithm

The Stochastic Recursive Variance-Reduced PG (SRVR-PG) algorithm is introduced in [21], where a recursively updated semi-stochastic gradient u_t^{j+1} is applied as an update direction.

Algorithm 2 Stochastic Recursive Variance Reduced Policy Gradient (SRVR-PG)

Input: number of epochs S, epoch size m, stepsize η, batch size N, minibatch size B, truncation horizon H, initial parameter $\theta_m^0 = \theta_0 \in \mathbb{R}^{\mathrm{d}}$.

for $j \leftarrow 0, ..., S-1$ **do**
  $\theta_0^{j+1} = \theta_m^j$;
  Sample N trajectories $\{\tau_i^H\}_{i=1}^N$ from $p_\rho^H(\cdot|\theta_0^{j+1})$;
  $u_0^{j+1} = \frac{1}{N}\sum_{i=1}^N g(\tau_i^H|\theta_0^{j+1})$;
  $\theta_1^{j+1} = \theta_0^{j+1} - \eta\nu_0^{j+1}$;
  for $t \leftarrow 1, ..., m-1$ **do**
    Sample B trajectories $\{\tau_j^H\}_{j=1}^B$ from $p_\rho^H(\cdot|\theta_t^{j+1})$;
    $u_t^{j+1} = u_{t-1}^{j+1} + \frac{1}{B}\sum_{j=1}^B \Big(g(\tau_j^H|\theta_t^{j+1}) - g_w(\tau_j^H|\theta_{t-1}^{j+1})\Big)$;
    $\theta_{t+1}^{j+1} = \theta_t^{j+1} - \eta u_t^{j+1}$;
  end for
end for
return θ_{out} chosen uniformly from $\{\theta\}_{j=1,\ldots,S;t=0,\ldots,m-1}$.

Here, the gradient estimators g and g_w are defined in (2.6) and (3.1), respectively.

E Stationary Convergence

In this section, we proceed to establish the stationary convergence of stochastic PG, NPG, SRVR-PG, and SRVR-NPG from an optimization perspective.

The stationary convergence of stochastic PG follows from the analysis of SGD. For SRVR-PG, we adapt its analysis in [18].

For NPG and SRVR-NPG, the Fisher information matrix $F(\theta)$ is applied as a preconditioner on top of PG and SRVR-PG, respectively. Regarding $F(\theta)$, we know from Assumptions 2.1 and 4.2 that

$$\mu_F I_d \preccurlyeq F(\theta) \preccurlyeq G^2 I_d \ \text{ for any } \ \theta \in \mathbb{R}^{\mathrm{d}}.$$

Since $\mu_F > 0$, we know that $F(\theta)$ defines a nice metric around θ. Consequently, with the analysis of gradient methods in nonconvex optimization, one can show that NPG (SRVR-NPG) has a similar iteration complexity compared with PG (SRVR-PG), although at each iteration, a subproblem needs to be solved in order to obtain an approximate preconditioned update direction.

We next present the stationary convergence results, and prove them in the subsequent sections. These results are established for $J^H(\theta)$ or $J(\theta)$, and we will apply the intermediate results in their proof to establish the global convergence on $J(\theta)$ (up to function approximation errors due to policy parametrizations).

Theorem E.1. In the stochastic PG update (2.4), by choosing $\eta = \frac{1}{4L_J}$, $K = \frac{32L_J(J^{H,\star} - J^H(\theta_0))}{\varepsilon}$, and $N = \frac{6\sigma^2}{\varepsilon}$, we have

$$\frac{1}{K}\sum_{k=0}^{K-1} \mathbb{E}[\|\nabla J^H(\theta^k)\|^2] \leq \varepsilon.$$

In total, stochastic PG samples $\mathcal{O}\left(\frac{\sigma^2}{(1-\gamma)^2\varepsilon^2}\right)$ trajectories.

Theorem E.2. In the NPG update (2.8), let us apply $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon}\right)$ iterations of SGD as in Procedure 1 to obtain an update direction w^k. In addition, let us take $\eta = \frac{\mu_F^2}{4G^2L_J}$ and $K = \frac{32L_JG^4(J^\star - J(\theta_0)}{\mu_F^2\varepsilon}$. Then, we have

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J(\theta^k)\|^2] \leq \varepsilon.$$

In total, NPG samples $\mathcal{O}\left(\frac{1}{(1-\gamma)^6\varepsilon^2}\right)$ trajectories.

Corollary E.3. (Theorem 4.5 of [21]) In SRVR-PG (Algorithm 2), take $\eta = \frac{1}{4L_J}$, $N = \frac{12\sigma^2}{\varepsilon}$, $S = \frac{64MR(J^\star - J(\theta^0))}{(1-\gamma)^{2.5}\varepsilon^{0.5}}$, $m = \frac{(1-\gamma)^{0.5}}{\varepsilon^{0.5}}$, and $B = \frac{72\eta G^2(2G^2+M)(W+1)\gamma}{M(1-\gamma)^3}m$. Then, we have

$$\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\nabla J^H(\theta_t^{j+1})\|^2 \leq \varepsilon.$$

In total, SRVR-PG samples $\mathcal{O}\left(\frac{W+\sigma^2}{(1-\gamma)^{2.5}\varepsilon^{1.5}}\right)$ trajectories.

Theorem E.4. In SRVR-NPG (Algorithm 1), take $\eta = \frac{\mu_F}{8L_J}$, $S = \frac{24G^2(J^{H,\star} - J^H(\theta_0))}{\eta\varepsilon^{0.5}}$, $m = \frac{1}{\varepsilon^{0.5}}$, $B = \left(\frac{\eta}{\mu_F} + \frac{\eta}{4G^2}\right)\frac{72RG^2(2G^2+M)(W+1)\gamma}{(1-\gamma)^5}\frac{1}{L_J\varepsilon^{0.75}}$, and $N = 3\left(\frac{8G^2}{\mu_F} + 2\right)\frac{\sigma^2}{\varepsilon}$. In addition, assume that ε is small enough such that

$$\begin{aligned}\varepsilon \leq \min\Bigg\{&3\left(\frac{8G^2}{\mu_F}+2\right)\left(\frac{GR}{(1-\gamma)^2}\right)^2, 3\left(\frac{8G^2}{4}+\frac{8G^4}{4\mu_F}\right)\frac{2}{\mu_F}\left(\frac{GR}{(1-\gamma)^2}\right)^2,\\ &\left(\frac{2}{3\eta L_J}(\mu_F + \frac{\mu_F^2}{4G^2})\right)^4\Bigg\}.\end{aligned}$$

Let us also apply $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon}\right)$ iterations of SGD as in Procedure 2 to obtain an update direction w_t^{j+1}. Then, in order to have

$$\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\nabla J^H(\theta_t^{j+1})\|^2 \leq \varepsilon,$$

SRVR-NPG samples $\mathcal{O}\left(\frac{\sigma^2}{(1-\gamma)^2\varepsilon^{1.5}} + \frac{W}{(1-\gamma)^3\varepsilon^{1.75}} + \frac{1}{(1-\gamma)^6\varepsilon^2}\right)$ trajectories.

F Proof of Theorem E.1

Proof. Let $g^k = \frac{1}{N}\sum_{i=1}^N g(\tau_i^H|\theta^k)$. Then, we have

$$\begin{aligned}J^H(\theta^{k+1}) &\geq J^H(\theta^k) + \langle\nabla J^H(\theta^k), \theta^{k+1} - \theta^k\rangle - \frac{L_J}{2}\|\theta^{k+1} - \theta^k\|^2\\ &= J^H(\theta^k) + \eta\langle\nabla J^H(\theta^k), g^k\rangle - \frac{L_J\eta^2}{2}\|g^k\|^2\\ &= J^H(\theta^k) + \eta\langle\nabla J^H(\theta^k), g^k - \nabla J^H(\theta^k) + \nabla J^H(\theta^k)\rangle\\ &\quad - \frac{L_J\eta^2}{2}\|g^k - \nabla J^H(\theta^k) + \nabla J^H(\theta^k)\|^2\\ &\geq J^H(\theta^k) + \frac{\eta}{2}\|\nabla J^H(\theta^k)\|^2 - \frac{\eta}{2}\|g^k - \nabla J^H(\theta^k)\|^2\\ &\quad - L_J\eta^2\|g^k - \nabla J^H(\theta^k)\|^2 - L_J\eta^2\|\nabla J^H(\theta^k)\|^2\\ &= J^H(\theta^k) + (\frac{\eta}{2} - L_J\eta^2)\|\nabla J^H(\theta^k)\|^2 - (\frac{\eta}{2} + L_J\eta^2)\|g^k - \nabla J^H(\theta^k)\|^2,\end{aligned} \tag{F.1}$$

where we have applied Lemma B.1 in the first inequality, and Cauchy-Schwartz in the second inequality.

Taking expectation on both sides and applying Lemma B.1 and Assumption 4.1 yields

$$\mathbb{E}[J^H(\theta^{k+1})] \geq \mathbb{E}[J^H(\theta^k)] + (\frac{\eta}{2} - L_J\eta^2)\mathbb{E}[\|\nabla J^H(\theta^k)\|^2] - (\frac{\eta}{2} + L_J\eta^2)\frac{\sigma^2}{N}.$$

Let us further telescope from $k = 0$ to $K - 1$ to obtain

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J^H(\theta^k)\|^2] \leq \frac{\frac{J^{H,\star} - J^H(\theta_0)}{K} + (\frac{\eta}{2} + L_J\eta^2)\frac{\sigma^2}{N}}{\frac{\eta}{2} - L_J\eta^2}. \tag{F.2}$$

Taking $\eta = \frac{1}{4L_J}$, $K = \frac{32L_J(J^\star - J^H(\theta_0))}{\varepsilon}$, and $N = \frac{6\sigma^2}{\varepsilon}$ gives

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J^H(\theta^k)\|^2] \leq \varepsilon.$$

Finally, by applying $L_J = \frac{MR}{(1-\gamma)^2}$, we know that PG needs to sample $KN = \frac{192MR(J^{H,\star} - J^H(\theta_0))\sigma^2}{(1-\gamma)^2\varepsilon^2} = \mathcal{O}\left(\frac{\sigma^2}{(1-\gamma)^2\varepsilon^2}\right)$ trajectories. $\square$

G Proof of Theorem E.2

Before proving Theorem E.2, let us first establish the sample complexity of SGD when applied to obtain an approximate NPG update direction w^k.

Proposition G.1. In Procedure 1, take $\alpha = \frac{1}{4G^2}$ and let the objective be

$$l(w) = \frac{1}{2(1-\gamma)^2}L_{\nu^{\pi_\theta}}(w;\theta^k) = \frac{1}{2}\mathbb{E}_{(s,a)\sim\nu^{\pi_{\theta^k}}}\left[\frac{1}{1-\gamma}A^{\pi_{\theta^k}}(s,a) - w^\top\nabla_\theta\log\pi_{\theta^k}(a\,|\,s)\right]^2.$$

Let $w_\star^k$ be the minimizer of $l(w)$. Then, in order to achieve

$$\mathbb{E}[\|w_{\text{out}} - w_\star^k\|^2] \leq \varepsilon',$$

Procedure 1 requires sampling

$$\frac{4\left((\frac{G^2R}{\mu_F(1-\gamma)^2} + \frac{2}{1-\gamma})\sqrt{d} + \frac{G^2R}{\mu_F(1-\gamma)^2}\right)^2}{\mu_F\varepsilon'} = \mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon'}\right)$$

trajectories.

Proof. In this proof, we will suppress the superscript k.

Let $l^\star = \min_{w\in\mathbb{R}^d} l(w)$ and $w_\star = \operatorname{argmin}_{w\in\mathbb{R}^d} l(w)$.

By Theorem 1 of [54], we know that

$$\mathbb{E}[l(w_{\text{out}}) - l^\star] \leq \frac{2(\xi\sqrt{d} + G\mathbb{E}[\|w_\star\|])^2}{T},$$

where $l^\star$ is the minimum of $l(w)$, and ξ is defined such that

$$\mathbb{E}[g_\star(g_\star)^T] \preccurlyeq \xi^2\nabla_w^2 l(w),$$

where $g_\star$ is a stochastic gradient of $l(w)$ at $w_\star$.

Following Coro 6.10 of [27], we can take

$$\xi = \frac{G^2R}{\mu_F(1-\gamma)^2} + \frac{2R}{(1-\gamma)^2}.$$

This leads to

$$\mathbb{E}[l(w_{\text{out}}) - l(w_\star)] \leq \frac{2\left((\frac{G^2R}{\mu_F(1-\gamma)^2} + \frac{2R}{(1-\gamma)^2})\sqrt{d} + \frac{G^2R}{\mu_F(1-\gamma)^2}\right)^2}{T}.$$

Since $l(w)$ is μ_F-strongly convex, in order to achieve $\mathbb{E}[\|w_{\text{out}} - w_\star^k\|^2] \leq \varepsilon'$, let us set

$$\mathbb{E}[l(w_{\text{out}}) - l(w_\star)] \leq \frac{\mu_F}{2}\varepsilon'.$$

Then, we need

$$T = \frac{4\left((\frac{G^2R}{\mu_F(1-\gamma)^2} + \frac{2R}{(1-\gamma)^2})\sqrt{d} + \frac{G^2R}{\mu_F(1-\gamma)^2}\right)^2}{\mu_F\varepsilon'} = \mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon'}\right).$$

Since each stochastic gradient of SGD has a cost of $\frac{2}{1-\gamma}$ (see App. C), this means to sample $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon'}\right)$ trajectories. $\square$

Now, we are ready to prove Theorem E.2.

Proof of Theorem E.2. We apply SGD to obtain a w^k such that

$$\mathbb{E}\|w^k - F^{-1}(\theta^k)\nabla J(\theta^k)\|^2 \leq \frac{\mu_F^2\varepsilon}{32\eta^2G^4L_J^2\left(\frac{2G^4}{\mu_F^2}+1\right)} = \mathcal{O}\left(\varepsilon\right). \tag{G.1}$$

By Proposition G.1, we need to sample $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon}\right)$ trajectories.

From (G.1) we have

$$\mathbb{E}\|\theta^{k+1} - \theta_\star^{k+1}\|^2 = \eta^2\mathbb{E}\|w^k - F^{-1}(\theta^k)\nabla J(\theta^k)\|^2 \leq \frac{\mu_F^2\varepsilon}{32G^4L_J^2\left(\frac{2G^4}{\mu_F^2}+1\right)}, \tag{G.2}$$

where $\theta_\star^{k+1} = \theta^k + \eta F^{-1}(\theta^k)\nabla J(\theta^k)$.

By Lemma B.1 and Assumption 4.2 we have

$$\begin{aligned}
J(\theta^{k+1}) &\geq J(\theta^k) + \langle\nabla J(\theta^k), \theta_\star^{k+1} - \theta^k\rangle + \langle\nabla J(\theta^k), \theta^{k+1} - \theta_\star^{k+1}\rangle - \frac{L_J}{2}\|\theta^{k+1} - \theta^k\|^2 \\
&= J(\theta^k) + \eta\langle\nabla J(\theta^k), F^{-1}(\theta^k)\nabla J(\theta^k)\rangle \\
&\quad + \langle\nabla J(\theta^k), \theta^{k+1} - \theta_\star^{k+1}\rangle - \frac{L_J}{2}\|\theta^{k+1} - \theta^k\|^2 \\
&\geq J(\theta^k) + \frac{\eta}{G^2}\|\nabla J(\theta^k)\|^2 + \langle\nabla J(\theta^k), \theta^{k+1} - \theta_\star^{k+1}\rangle - \frac{L_J}{2}\|\theta^{k+1} - \theta^k\|^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}
J(\theta^{k+1}) &\geq J(\theta^k) + \frac{\eta}{2G^2}\|\nabla J(\theta^k)\|^2 - \frac{G^2}{2\eta}\|\theta^{k+1} - \theta_\star^{k+1}\|^2 - \frac{L_J}{2}\|\theta^{k+1} - \theta^k\|^2 \\
&\geq J(\theta^k) + \frac{\eta}{2G^2}\|\nabla J(\theta^k)\|^2 - \left(\frac{G^2}{2\eta} + L_J\right)\|\theta^{k+1} - \theta_\star^{k+1}\|^2 - L_J\|\theta_\star^{k+1} - \theta^k\|^2 \\
&\geq J(\theta^k) + \left(\frac{\eta}{2G^2} - \frac{L_J\eta^2}{\mu_F^2}\right)\|\nabla J(\theta^k)\|^2 - \left(\frac{G^2}{2\eta} + L_J\right)\|\theta^{k+1} - \theta_\star^{k+1}\|^2,
\end{aligned}$$

where we have applied Cauchy-Schwartz in the first and second inequalities, and $\theta_\star^{k+1} = \theta^k + \eta F^{-1}(\theta^k)\nabla J(\theta^k)$ in the last step.

Taking full expectation on both sides yields

$$\begin{aligned}\mathbb{E}[J(\theta^{k+1})] &\geq \mathbb{E}[J(\theta^k)] + \left(\frac{\eta}{2G^2} - \frac{L_J\eta^2}{\mu_F^2}\right)\mathbb{E}\|\nabla J(\theta^k)\|^2 - \left(\frac{G^2}{2\eta} + L_J\right)\mathbb{E}\|\theta^{k+1} - \theta_\star^{k+1}\|^2 \\ &\geq \mathbb{E}[J(\theta^k)] + \left(\frac{\eta}{2G^2} - \frac{L_J\eta^2}{\mu_F^2}\right)\mathbb{E}\|\nabla J(\theta^k)\|^2 - \left(\frac{G^2}{2\eta} + L_J\right)\frac{\mu_F^2\varepsilon}{32G^4L_J^2\left(\frac{2G^4}{\mu_F^2}+1\right)},\end{aligned}$$

where we have applied (G.2) in the second inequality.

Telescoping the above inequality from $k = 0$ to $k = K - 1$ gives

$$\frac{J^\star - J(\theta_0)}{K} \geq \left(\frac{\eta}{2G^2} - \frac{L_J\eta^2}{\mu_F^2}\right)\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|\nabla J(\theta^k)\|^2 - \left(\frac{G^2}{2\eta} + L_J\right)\frac{\mu_F^2\varepsilon}{32G^4L_J^2\left(\frac{2G^4}{\mu_F^2}+1\right)}.$$

Finally, by taking $\eta = \frac{\mu_F^2}{4G^2L_J}$ and $K = \frac{32L_JG^4(J^\star - J(\theta_0))}{\mu_F^2\varepsilon} = \mathcal{O}\left(\frac{1}{(1-\gamma)^2\varepsilon}\right)$, we arrive at

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J(\theta^k)\|^2] \leq \frac{\frac{J^\star - J(\theta_0)}{K} + \frac{\left(\frac{G^2}{2\eta}+L_J\right)\mu_F^2\varepsilon}{32G^4L_J^2\left(\frac{2G^4}{\mu_F^2}+1\right)}}{\left(\frac{\eta}{2G^2} - \frac{L_J\eta^2}{\mu_F^2}\right)} = \varepsilon. \tag{G.3}$$

Recall that at each iteration of NPG, we apply SGD as in Procedure 1 to reach (G.1). By Proposition G.1, we know that in total, NPG requires to sample

$$\frac{32L_JG^4(J^\star - J(\theta_0))}{\mu_F^2\varepsilon} \cdot \frac{4\left((\frac{G^2R}{\mu_F(1-\gamma)^2} + \frac{2R}{(1-\gamma)^2})\sqrt{d} + \frac{G^2R}{\mu_F(1-\gamma)^2}\right)^2}{\mu_F\frac{\mu_F^2\varepsilon}{32\eta^2G^4L_J^2\left(\frac{2G^4}{\mu_F^2}+1\right)}} = \mathcal{O}\left(\frac{1}{(1-\gamma)^6\varepsilon^2}\right)$$

trajectories.

□

H Proof of Theorem E.3

Proof. By Theorem 4.5 of [21], we know that if $\eta = \frac{1}{4L_J}$ and

$$B = \frac{3\eta C_\gamma m}{L_J} = \frac{72\eta G^2(2G^2 + M)(W+1)\gamma}{M(1-\gamma)^3}m,$$

then

$$\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\nabla J^H(\theta_t^{j+1})\|^2 \leq \frac{8(J^\star - J^H(\theta_0))}{\eta Sm} + \frac{6\sigma^2}{N}.$$

Therefore, taking $N = \frac{12\sigma^2}{\varepsilon}$ and $Sm = \frac{64MR(J^\star - J^H(\theta^0))}{(1-\gamma)^2\varepsilon}$ yields

$$\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\nabla J^H(\theta_t^{j+1})\|^2 \leq \varepsilon.$$

Let us take $S = \frac{64MR(J^\star - J^H(\theta^0))}{(1-\gamma)^{2.5}\varepsilon^{0.5}}$ and $m = \frac{(1-\gamma)^{0.5}}{\varepsilon^{0.5}}$. Then, the number of trajectories required by SRVR-PG is

$$\begin{aligned}S(N + mB) &= S\frac{12\sigma^2}{\varepsilon} + \frac{64MR(J^\star - J^H(\theta^0))}{(1-\gamma)^2\varepsilon}B \\ &= S\frac{12\sigma^2}{\varepsilon} + \frac{64MR(J^\star - J^H(\theta^0))}{(1-\gamma)^2\varepsilon}\frac{72\eta G^2(2G^2+M)(W+1)\gamma}{M(1-\gamma)^3}m \\ &= \mathcal{O}\left(\frac{\sigma^2}{(1-\gamma)^{2.5}\varepsilon^{1.5}} + \frac{W}{(1-\gamma)^{2.5}\varepsilon^{1.5}}\right) \\ &= \mathcal{O}\left(\frac{W + \sigma^2}{(1-\gamma)^{2.5}\varepsilon^{1.5}}\right).\end{aligned}$$

Therefore, SRVR-PG needs to sample $\mathcal{O}\left(\frac{W+\sigma^2}{(1-\gamma)^{2.5}\varepsilon^{1.5}}\right)$ trajectories. □

I Proof of Theorem E.4

In order to prove Theorem E.4, we need the following technical results.

Lemma I.1 (Equation B.10 of [20])**.** We have

$$\mathbb{E}\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 \leq \frac{C_\gamma}{B}\sum_{l=1}^{t}\mathbb{E}\|\theta_l^{j+1} - \theta_{l-1}^{j+1}\|^2 + \frac{\sigma^2}{N},$$

where

$$C_\gamma = \frac{24RG^2(2G^2+M)(W+1)\gamma}{(1-\gamma)^5}.$$

Proof. This lemma is adapted from the Equation B.10 of [20], where SRVR-PG is analyzed. It is also true for our SRVR-NPG since the update rule of u_t^{j+1} is the same for both algorithms. □

Proposition I.2. In SRVR-NPG, apply SGD as in Procedure 2 to solve the subproblems. Take $\alpha = \frac{1}{4G^2}$ and let the objective be

$$l(w) = \frac{1}{2}\left(\mathbb{E}_{(s,a)\sim\nu_t^{j+1}}[w^T\nabla_\theta \log \pi_{\theta_t^{j+1}}(a \mid s)]^2 - 2\langle \eta w, u_t^{j+1}\rangle\right).$$

Let $w_{t,\star}^{j+1} = F^{-1}(\theta_t^{j+1})\nabla J^H(\theta_t^{j+1})$ be the minimizer of $l(w)$. Assume in addition that

$$\frac{\sigma^2}{N} \leq \left(\frac{GR}{(1-\gamma)^2}\right)^2,$$
$$\varepsilon' \leq \frac{2}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2$$
$$\frac{C_\gamma m}{B}2\eta^2 \leq \frac{1}{3}\mu_F^2.$$

Then, in order to achieve

$$\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1})\|^2] \leq \varepsilon'$$

for each $s = 0, 1, ..., S-1$ and $t = 0, 1, ..., m-1$, Procedure 2 requires sampling

$$\frac{4\left(\frac{\frac{2}{\mu_F}\frac{GR}{(1-\gamma)^2}G^2 + \frac{2GR}{(1-\gamma)^2}}{\sqrt{\mu_F}}\sqrt{d} + \frac{2G^2R}{\mu_F(1-\gamma)^2}\right)^2}{\mu_F\varepsilon'} = \mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon'}\right)$$

trajectories.

Proof of Proposition I.2. Recall that we are applying SGD as in Procedure 2 to solve the SRVR-NPG subproblem (3.3).

Let us focus on $t = 0$, where $u_0^{j+1} = \frac{1}{N}\sum_{i=1}^{N} g(\tau_i^H|\theta_0^{j+1})$. As a result, $\mathbb{E}[u_0^{j+1}] = \nabla J^H(\theta_0^{j+1})$ and $\mathrm{Var}(u_0^{j+1}) \leq \frac{\sigma^2}{N}$. Therefore,

$$\begin{aligned}\mathbb{E}[\|w_{0,\star}^{j+1}\|^2] &= \mathbb{E}[\|F^{-1}(\theta_0^{j+1})u_0^{j+1}\|^2]\\ &\leq \frac{1}{\mu_F^2}\mathbb{E}[\|u_0^{j+1}\|] \leq \frac{1}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 + \frac{1}{\mu_F^2}\frac{\sigma^2}{N} \leq \frac{4}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2.\end{aligned}$$

Recall from (C.4) that a stochastic gradient $\nabla l(w_t)$ is given by

$$\nabla l(w) = \left(w^T\nabla_\theta \log \pi_{\theta_0^{j+1}}(a|s)\right)\nabla_\theta \log \pi_{\theta_0^{j+1}}(a|s) - u_0^{j+1}.$$

Here, $s, a \sim \nu^{\pi_\theta}$.

Therefore, By Theorem 1 of [54], we know that in order to reach $\mathbb{E}[\|w_0^{j+1} - (w_0^{j+1})_\star\|^2] \le \varepsilon'$, we need

$$\mathbb{E}[l(w_{\text{out}}) - l^\star] \le \frac{2(\xi\sqrt{d} + G\|w_{0,\star}^{j+1}\|)^2}{T} \le \frac{\mu_F}{2}\varepsilon',$$

where $l^\star$ is the minimum of $l(w)$, and ξ is defined such that the stochastic gradient $g_\star$ at the solution $w_{0,\star}^{j+1}$ satisfies

$$\mathbb{E}[g_\star(g_\star)^T] \preccurlyeq \xi^2\nabla_w^2 l(w).$$

Similar as Proposition G.1, we know that ξ can be chosen by

$$\xi^2 = \frac{\left(\frac{2}{\mu_F}\frac{GR}{(1-\gamma)^2}G^2 + \frac{2GR}{(1-\gamma)^2}\right)^2}{\mu_F}.$$

As a result, the number of iterations, T, should be

$$T = \frac{4\left(\frac{\frac{2}{\mu_F}\frac{GR}{(1-\gamma)^2}G^2 + \frac{2GR}{(1-\gamma)^2}}{\sqrt{\mu_F}}\sqrt{d} + \frac{2G^2R}{\mu_F(1-\gamma)^2}\right)^2}{\mu_F\varepsilon'} = \mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon'}\right).$$

Since each stochastic gradient of $l(w)$ only needs to sample a state-action pair, this is equivalent to sampling $\mathcal{O}\left(\frac{1}{(1-\gamma)^3\varepsilon'}\right)$ trajectories.

Now, let us turn to $t \ge 1$. u_t^{j+1} is an unbiased estimate of $\nabla J(\theta_t^{j+1})$, and its variance is bounded as in Lemma I.1. Therefore,

$$\begin{aligned}
\mathbb{E}[\|w_{t,\star}^{j+1})\|^2] &= \mathbb{E}[\|F^{-1}(\theta_t^{j+1})u_t^{j+1}\|^2] \le \frac{1}{\mu_F^2}\mathbb{E}[\|u_t^{j+1}\|] \\
&= \frac{1}{\mu_F^2}\mathbb{E}[\|\nabla J^H(\theta_t^{j+1})\|^2] + \frac{1}{\mu_F^2}\mathbb{E}[\|u_t^{j+1} - \nabla J^H(\theta_t^{j+1})\|^2] \\
&\le \frac{1}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 + \frac{1}{\mu_F^2}\left(\frac{C_\gamma}{B}\sum_{l=1}^t \mathbb{E}\|\theta_l^{j+1} - \theta_{l-1}^{j+1}\|^2 + \frac{\sigma^2}{N}\right) \\
&= \frac{1}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 + \frac{1}{\mu_F^2}\left(\frac{C_\gamma}{B}\sum_{l=1}^t \eta^2\mathbb{E}\|w_{l-1}^{j+1}\|^2 + \frac{\sigma^2}{N}\right) \\
&\le \frac{1}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 \\
&\quad + \frac{1}{\mu_F^2}\left(\frac{C_\gamma}{B}\sum_{l=1}^t 2\eta^2\left(\mathbb{E}\|w_{l-1,\star}^{j+1}\|^2 + \mathbb{E}\|w_{l-1,\star}^{j+1} - w_{l-1}^{j+1}\|^2\right) + \frac{\sigma^2}{N}\right).
\end{aligned} \tag{I.1}$$

Now, we are ready to prove the desired results by induction.

Assume that for all $t' < t$, we have

$$\mathbb{E}[\|w_{t',\star}^{j+1})\|] \le \frac{4}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2,$$

and we have applied

$$T = \frac{4\left(\frac{\frac{2}{\mu_F}\frac{GR}{(1-\gamma)^2}G^2 + \frac{2GR}{(1-\gamma)^2}}{\sqrt{\mu_F}}\sqrt{d} + \frac{2G^2R}{\mu_F(1-\gamma)^2}\right)^2}{\mu_F\varepsilon'} = \mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon'}\right)$$

iterations of SGD as in Procedure 2.

Similar to the case of $t = 0$, we know that this yields

$$\mathbb{E}[\|w_{t'}^{j+1} - w_{t',\star}^{j+1}\|^2] \le \varepsilon'.$$

Then, by (I.1), we have

$$
\begin{aligned}
\mathbb{E}[\|w_{t,\star}^{j+1}\|^2] &\le \frac{1}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 \\
&\quad + \frac{1}{\mu_F^2}\left(\frac{C_\gamma}{B}\sum_{l=1}^{t} 2\eta^2\left(\mathbb{E}\|w_{l-1,\star}^{j+1}\|^2 + \mathbb{E}\|w_{l-1,\star}^{j+1} - w_{l-1}^{j+1}\|^2\right) + \frac{\sigma^2}{N}\right) \\
&\le \frac{1}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 + \frac{1}{\mu_F^2}\left(\frac{C_\gamma m}{B}2\eta^2\left(\frac{4}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 + \varepsilon'\right) + \frac{\sigma^2}{N}\right) \\
&= \frac{1}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 + \frac{1}{\mu_F^2}\frac{C_\gamma m}{B}2\eta^2\frac{4}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 \\
&\quad + \frac{1}{\mu_F^2}\frac{C_\gamma m}{B}2\eta^2\varepsilon' + \frac{1}{\mu_F^2}\frac{\sigma^2}{N} \\
&\le \frac{4}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2.
\end{aligned}
$$

As a result, we can apply

$$
T = \frac{4\left(\frac{\frac{2}{\mu_F}\frac{GR}{(1-\gamma)^2}G^2 + \frac{2GR}{(1-\gamma)^2}}{\sqrt{\mu_F}}\sqrt{d} + \frac{2G^2R}{\mu_F(1-\gamma)^2}\right)^2}{\mu_F\varepsilon'} = \mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon'}\right)
$$

iterations of SGD as in Procedure 2 so that

$$
\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2] \le \varepsilon'.
$$

Since each stochastic gradient of $l(w)$ has a cost of $\frac{1}{1-\gamma}$ (see App. C), this is equivalent to sample

$$
\frac{4\left(\frac{\frac{2}{\mu_F}\frac{GR}{(1-\gamma)^2}G^2 + \frac{2GR}{(1-\gamma)^2}}{\sqrt{\mu_F}}\sqrt{d} + \frac{2G^2R}{\mu_F(1-\gamma)^2}\right)^2}{\mu_F\varepsilon'} = \mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon'}\right)
$$

trajectories. □

We are now ready to prove Theorem E.4.

Proof of Theorem E.4. Line 9 of Algorithm 1 reads

$$
w_t^{j+1} \approx \underset{w}{\operatorname{argmin}}\{\mathbb{E}_{(s,a)\sim\nu_t^{j+1}}[w^T\nabla_\theta \log \pi_{\theta_t^{j+1}}(s,a)]^2 - 2\langle \eta w, u_t^{j+1}\rangle\}.
$$

And we want to apply SGD as in Procedure 2 to obtain a w_t^{j+1} that satisfies

$$
\mathbb{E}[\|w_t^{j+1} - F^{-1}(\theta_t^{j+1})u_t^{j+1}\|^2] \le \frac{\varepsilon}{3\left(\frac{8G^2\mu_F}{4} + \frac{8G^4}{4}\right)}. \tag{I.2}
$$

Recall that the parameters S, m, B and N are chosen as

$$
\begin{aligned}
S &= \frac{24G^2(J^\star - J^H(\theta_0))}{\eta\varepsilon^{0.5}}, \\
m &= \frac{1}{\varepsilon^{0.5}}, \\
B &= \left(\frac{\eta}{\mu_F} + \frac{\eta}{4G^2}\right)\frac{4C_\gamma m}{L_J\varepsilon^{0.25}}, \\
N &= 3\left(\frac{8G^2}{\mu_F} + 2\right)\frac{\sigma^2}{\varepsilon}.
\end{aligned}
$$

Since

$$
\begin{aligned}
\varepsilon \leq \min \Big\{ & 3\left(\frac{8G^2}{\mu_F}+2\right)\left(\frac{GR}{(1-\gamma)^2}\right)^2, 3\left(\frac{8G^2}{4}+\frac{8G^4}{4\mu_F}\right)\frac{2}{\mu_F}\left(\frac{GR}{(1-\gamma)^2}\right)^2, \\
& \left(\frac{2}{3\eta L_J}(\mu_F+\frac{\mu_F^2}{4G^2})\right)^4 \Big\},
\end{aligned}
$$

the requirements of Proposition I.2 are satisfied:

$$
\begin{aligned}
\frac{\sigma^2}{N} &= \frac{\varepsilon}{3\left(\frac{8G^2}{\mu_F}+2\right)} \leq \left(\frac{GR}{(1-\gamma)^2}\right)^2, \\
\varepsilon' &= \frac{\varepsilon}{3\left(\frac{8G^2\mu_F}{4}+\frac{8G^4}{4}\right)} \leq \frac{2}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 \\
\frac{C_\gamma m}{B} 2\eta^2 &= \frac{2\eta^2 \varepsilon^{0.25}}{\left(\frac{\eta}{\mu_F}+\frac{\eta}{4G^2}\right)\frac{4}{L_J}} \leq \frac{1}{3}\mu_F^2.
\end{aligned}
$$

By applying Proposition I.2, we know that in order to have (I.2), one needs to sample

$$
\frac{4\left(\frac{\frac{\sqrt{2}}{\mu_F}\frac{GR}{(1-\gamma)^2}G^2+\frac{\sqrt{2}GR}{(1-\gamma)^2}}{\sqrt{\mu_F}}\sqrt{d}+\frac{\sqrt{2}G^2R}{\mu_F(1-\gamma)^2}\right)^2}{\mu_F\varepsilon'} = \mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon}\right)
$$

trajectories. By (I.2) we know that

$$
\mathbb{E}[\|\theta_{t+1}^{j+1}-\theta_{t+1,\star}^{j+1}\|^2] \leq \frac{\varepsilon}{3\left(\frac{8G^2\mu_F}{4\eta^2}+\frac{8G^4}{4\eta^2}\right)} \tag{I.3}
$$

where $\theta_{t+1,\star}^{j+1} = \theta_t^{j+1} + \eta F^{-1}(\theta_t^{j+1})u_t^{j+1}$.
On the other hand, we have

$$
\begin{aligned}
J^H(\theta_{t+1}^{j+1}) &\geq J^H(\theta_t^{j+1}) + \langle \nabla J^H(\theta_t^{j+1}), \theta_{t+1}^{j+1}-\theta_t^{j+1}\rangle - \frac{L_J}{2}\|\theta_{t+1}^{j+1}-\theta_t^{j+1}\|^2 \\
&= J^H(\theta_t^{j+1}) + \langle \nabla J^H(\theta_t^{j+1}) - u_t^{j+1}, \theta_{t+1}^{j+1}-\theta_t^{j+1}\rangle \\
&\quad + \langle u_t^{j+1}, \theta_{t+1}^{j+1}-\theta_t^{j+1}\rangle - \frac{L_J}{2}\|\theta_{t+1}^{j+1}-\theta_t^{j+1}\|^2 \\
&\geq J^H(\theta_t^{j+1}) - \frac{\eta}{\mu_F}\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 - \frac{\mu_F}{4\eta}\|\theta_{t+1}^{j+1}-\theta_t^{j+1}\|^2 \\
&\quad + \langle u_t^{j+1}, \theta_{t+1}^{j+1}-\theta_t^{j+1}\rangle - \frac{L_J}{2}\|\theta_{t+1}^{j+1}-\theta_t^{j+1}\|^2
\end{aligned}
$$

where we have applied Lemma B.1 in the first inequality, and Cauchy-Schwartz in the second one.

Rearranging gives

$$
\begin{aligned}
J^H(\theta_{t+1}^{j+1}) \geq\; & J^H(\theta_t^{j+1}) - \frac{\eta}{\mu_F}\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 - \frac{\mu_F}{4\eta}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 \\
& + \frac{1}{2}\langle u_t^{j+1}, \theta_{t+1,\star}^{j+1} - \theta_t^{j+1}\rangle \\
& + \frac{1}{2}\langle u_t^{j+1}, \theta_{t+1}^{j+1} - \theta_{t+1,\star}^{j+1}\rangle + \frac{1}{2}\langle \frac{1}{\eta}F(\theta_t^{j+1})(\theta_{t+1,\star}^{j+1} - \theta_t^{j+1}), \theta_{t+1}^{j+1} - \theta_t^{j+1}\rangle \\
& - \frac{L_J}{2}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 \\
=\; & J^H(\theta_t^{j+1}) - \frac{\eta}{\mu_F}\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 \\
& - \frac{\mu_F}{4\eta}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 + \frac{1}{2}\langle u_t^{j+1}, \eta F^{-1}(\theta_t^{j+1}) u_t^{j+1}\rangle \\
& + \frac{1}{2}\langle u_t^{j+1}, \theta_{t+1}^{j+1} - \theta_{t+1,\star}^{j+1}\rangle + \frac{1}{2}\langle \frac{1}{\eta}F(\theta_t^{j+1})(\theta_{t+1,\star}^{j+1} - \theta_t^{j+1}), \theta_{t+1}^{j+1} - \theta_t^{j+1}\rangle \\
& - \frac{L_J}{2}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2,
\end{aligned}
$$

Applying $F^{-1}(\theta) \succeq \frac{1}{G^2}I$ on the first inner product, and Cauchy-Schwartz on the second inner product term leads to

$$
\begin{aligned}
J^H(\theta_{t+1}^{j+1}) \geq\; & J^H(\theta_t^{j+1}) - \frac{\eta}{\mu_F}\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 \\
& - \frac{\mu_F}{4\eta}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 + \frac{\eta}{2G^2}\|u_t^{j+1}\|^2 \\
& - \frac{\eta}{4G^2}\|u_t^{j+1}\|^2 - \frac{G^2}{4\eta}\|\theta_{t+1}^{j+1} - \theta_{t+1,\star}^{j+1}\|^2 \\
& + \frac{1}{2}\langle \frac{1}{\eta}F(\theta_t^{j+1})(\theta_{t+1,\star}^{j+1} - \theta_t^{j+1}), \theta_{t+1}^{j+1} - \theta_t^{j+1}\rangle - \frac{L_J}{2}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 \\
=\; & J^H(\theta_t^{j+1}) - \frac{\eta}{\mu_F}\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 \\
& - \left(\frac{\mu_F}{4\eta} + \frac{L_J}{2}\right)\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 + \frac{\eta}{4G^2}\|u_t^{j+1}\|^2 \\
& + \frac{1}{2}\langle \frac{1}{\eta}F(\theta_t^{j+1})(\theta_{t+1,\star}^{j+1} - \theta_t^{j+1}), \theta_{t+1}^{j+1} - \theta_t^{j+1}\rangle - \frac{G^2}{4\eta}\|\theta_{t+1}^{j+1} - \theta_{t+1,\star}^{j+1}\|^2 \\
=\; & J^H(\theta_t^{j+1}) \\
& - \frac{\eta}{\mu_F}\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 - \left(\frac{\mu_F}{4\eta} + \frac{L_J}{2}\right)\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 + \frac{\eta}{4G^2}\|u_t^{j+1}\|^2 \\
& + \frac{1}{2}\langle \frac{1}{\eta}F(\theta_t^{j+1})(\theta_{t+1}^{j+1} - \theta_t^{j+1}), \theta_{t+1}^{j+1} - \theta_t^{j+1}\rangle \\
& + \frac{1}{2}\langle \frac{1}{\eta}F(\theta_t^{j+1})(\theta_{t+1,\star}^{j+1} - \theta_{t+1}^{j+1}), \theta_{t+1}^{j+1} - \theta_t^{j+1}\rangle \\
& - \frac{G^2}{4\eta}\|\theta_{t+1}^{j+1} - \theta_{t+1,\star}^{j+1}\|^2.
\end{aligned}
$$

Applying $\|\nabla J^H(\theta_t^{j+1})\|^2 \leq 2\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 + 2\|u_t^{j+1}\|^2$ and $F(\theta_t^{j+1}) \succeq \mu_F I_d$ yields

$$
\begin{aligned}
J^H(\theta_{t+1}^{j+1}) \geq\; & J^H(\theta_t^{j+1}) - \left(\frac{\eta}{\mu_F} + \frac{\eta}{4G^2}\right)\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 \\
& - \left(\frac{\mu_F}{4\eta} + \frac{L_J}{2}\right)\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 + \frac{\eta}{8G^2}\|\nabla J^H(\theta_t^{j+1})\|^2 \\
& + \frac{\mu_F}{2\eta}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 + \frac{1}{2}\langle \frac{1}{\eta}F(\theta_t^{j+1})(\theta_{t+1,\star}^{j+1} - \theta_{t+1}^{j+1}), \theta_{t+1}^{j+1} - \theta_t^{j+1}\rangle \\
& - \frac{G^2}{4\eta}\|\theta_{t+1}^{j+1} - \theta_{t+1,\star}^{j+1}\|^2 \\
\geq\; & J^H(\theta_t^{j+1}) - \left(\frac{\eta}{\mu_F} + \frac{\eta}{4G^2}\right)\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 \\
& - \left(\frac{\mu_F}{4\eta} + \frac{L_J}{2}\right)\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 + \frac{\eta}{8G^2}\|\nabla J^H(\theta_t^{j+1})\|^2 \\
& + \frac{\mu_F}{2\eta}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 - \frac{\mu_F^2}{2\mu_F\eta}\|\theta_{t+1,\star}^{j+1} - \theta_{t+1}^{j+1}\|^2 - \frac{\mu_F}{8\eta}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 \\
& - \frac{G^2}{4\eta}\|\theta_{t+1}^{j+1} - \theta_{t+1,\star}^{j+1}\|^2 \\
=\; & J^H(\theta_t^{j+1}) - \left(\frac{\eta}{\mu_F} + \frac{\eta}{4G^2}\right)\|\nabla J^H(\theta_t^{j+1}) - u_t^{j+1}\|^2 \\
& + \left(\frac{\mu_F}{8\eta} - \frac{L_J}{2}\right)\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 + \frac{\eta}{8G^2}\|\nabla J^H(\theta_t^{j+1})\|^2 \\
& - (\frac{\mu_F}{4\eta} + \frac{G^2}{4\eta})\|\theta_{t+1,\star}^{j+1} - \theta_{t+1}^{j+1}\|^2.
\end{aligned} \tag{I.4}
$$

From Lemma I.1, we further know that

$$
\begin{aligned}
\mathbb{E}[J^H(\theta_{t+1}^{j+1})] \geq\; & \mathbb{E}[J^H(\theta_t^{j+1})] - \left(\frac{\eta}{\mu_F} + \frac{\eta}{4G^2}\right)\left(\frac{C_\gamma}{B}\sum_{t=0}^{m-1}\mathbb{E}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 + \frac{\sigma^2}{N}\right) \\
& + \left(\frac{\mu_F}{8\eta} - \frac{L_J}{2}\right)\mathbb{E}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 \\
& + \frac{\eta}{8G^2}\mathbb{E}\|\nabla J^H(\theta_t^{j+1})\|^2 - (\frac{\mu_F}{4\eta} + \frac{G^2}{4\eta})\mathbb{E}\|\theta_{t+1,\star}^{j+1} - \theta_{t+1}^{j+1}\|^2.
\end{aligned}
$$

Telescoping for $s = 0, 1, ..., S-1$ and $t = 0, 1, ..., m-1$ and dividing by Sm gives

$$
\begin{aligned}
& \frac{8G^2}{\eta}\left(\frac{\mu_F}{8\eta} - \frac{L_J}{2} - \left(\frac{\eta}{\mu_F} + \frac{\eta}{4G^2}\right)\frac{C_\gamma m}{B}\right)\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2 \\
& + \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\nabla J^H(\theta_t^{j+1})\|^2 \\
& \leq \frac{8G^2}{\eta}\frac{J^\star - J^H(\theta_0)}{Sm} + \left(\frac{8G^2}{\mu_F} + 2\right)\frac{\sigma^2}{N} + \left(\frac{8G^2\mu_F}{4\eta^2} + \frac{8G^4}{4\eta^2}\right)\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\theta_{t+1,\star}^{j+1} - \theta_{t+1}^{j+1}\|^2.
\end{aligned} \tag{I.5}
$$

Let us first show that the first term on the left hand side of (I.5) is non-negative. In fact, from $\eta = \frac{\mu_F}{8L_J}$ and $B = \left(\frac{\eta}{\mu_F} + \frac{\eta}{4G^2}\right)\frac{4C_\gamma m}{L_J\varepsilon^{0.25}}$ we have

$$
\frac{8G^2}{\eta}\left(\frac{\mu_F}{8\eta} - \frac{L_J}{2} - \left(\frac{\eta}{\mu_F} + \frac{\eta}{4G^2}\right)\frac{C_\gamma m}{B}\right) \geq \frac{16G^2L_J^2}{\mu_F} > 0.
$$

Therefore, in order to have

$$\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\nabla J^H(\theta_t^{j+1})\|^2 \le \varepsilon,$$

we can set all the three terms on the right hand side of (I.5) to be $\frac{\varepsilon}{3}$, which gives

$$\begin{aligned} Sm &= \frac{24G^2(J^{H,\star} - J^H(\theta_0))}{\eta\varepsilon}, \\ N &= 3\left(\frac{8G^2}{\mu_F} + 2\right)\frac{\sigma^2}{\varepsilon} \\ \mathbb{E}[\|\theta_{t+1}^{j+1} - \theta_{t+1,\star}^{j+1}\|^2] &\le \frac{\varepsilon}{3\left(\frac{8G^2\mu_F}{4\eta^2} + \frac{8G^4}{4\eta^2}\right)}, \end{aligned}$$

where the last requirement is satisfied according to (I.3).

For the parameters S, m, B, and N, we have

$$\begin{aligned} S &= \frac{24G^2(J^{H,\star} - J^H(\theta_0))}{\eta\varepsilon^{0.5}} = \mathcal{O}\left(\frac{1}{(1-\gamma)^2\varepsilon^{0.5}}\right), \\ m &= \frac{1}{\varepsilon^{0.5}}, \\ B &= \left(\frac{\eta}{\mu_F} + \frac{\eta}{4G^2}\right)\frac{4C_\gamma m}{L_J\varepsilon^{0.25}} = \mathcal{O}\left(\frac{W}{(1-\gamma)\varepsilon^{0.75}}\right), \\ N &= 3\left(\frac{8G^2}{\mu_F} + 2\right)\frac{\sigma^2}{\varepsilon} = \mathcal{O}\left(\frac{\sigma^2}{\varepsilon}\right), \end{aligned}$$

where we have applied the definition of C_γ in Lemma I.1 in the third equality.

Therefore in total, the number of trajectories required by SRVR-NPG to reach $\varepsilon-$stationarity is

$$\begin{aligned} &S\left(N + mB + (m+1)\frac{4\left(\left(\frac{2G^2R}{\mu_F(1-\gamma)^2} + \frac{2}{1-\gamma}\right)\sqrt{d} + \frac{2G^2R}{\mu_F(1-\gamma)^2}\right)^2}{\mu_F\frac{\varepsilon}{3\left(\frac{8G^2\mu_F}{4} + \frac{8G^4}{4}\right)}}\right) \\ &= \mathcal{O}\left(\frac{\sigma^2}{(1-\gamma)^2\varepsilon^{1.5}} + \frac{W}{(1-\gamma)^3\varepsilon^{1.75}} + \frac{1}{(1-\gamma)^6\varepsilon^2}\right). \end{aligned}$$

□

J Proof of Proposition 4.5

In this section, we proceed to prove Proposition 4.5, which establishes a general global convergence result on policy gradient methods of the form $\theta^{k+1} = \theta^k + \eta w^k$.

Proof. First, by the $M-$smoothness of score function (see Assumption 4.2), we know that

$$\begin{aligned} &\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^k}(\cdot\,|\,s)\right) - \mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^{k+1}}(\cdot\,|\,s)\right)\right] \\ &= \mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\log\frac{\pi_{\theta^{k+1}}(a\,|\,s)}{\pi_{\theta^k}(a\,|\,s)}\right] \\ &\ge \mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta\log\pi_{\theta^k}(a\,|\,s)\cdot(\theta^{k+1} - \theta^k)] - \frac{M}{2}\|\theta^{k+1} - \theta^k\|^2 \\ &= \eta\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta\log\pi_{\theta^k}(a\,|\,s)\cdot w^k] - \frac{M\eta^2}{2}\|w^k\|^2 \\ &= \eta\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta\log\pi_{\theta^k}(a\,|\,s)\cdot w_\star^k] \\ &\quad + \eta\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta\log\pi_{\theta^k}(a\,|\,s)\cdot(w^k - w_\star^k)] - \frac{M\eta^2}{2}\|w^k\|^2 \end{aligned}$$

where we have applied $\mathrm{KL}(p||q) = \mathbb{E}_{x\sim p}[-\log \frac{q(x)}{p(x)}]$ in the first step.

On the other hand, by the performance difference lemma [59] we know that

$$\mathbb{E}_{s\sim d_\rho^{\pi^\star}} \mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[A^{\pi_{\theta^k}}(s,a)] = (1-\gamma)\left(J^\star - J(\theta^k)\right).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^k}(\cdot\,|\,s)\right) - \mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^{k+1}}(\cdot\,|\,s)\right)\right] \\
&\geq \eta\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\cdot w_\star^k] \\
&\quad + \eta\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\cdot (w^k - w_\star^k)] - \frac{M\eta^2}{2}\|w^k\|^2 \\
&= \eta\left(J^\star - J(\theta^k)\right) + \eta\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\cdot w_\star^k - \frac{1}{1-\gamma}A^{\pi_{\theta^k}}(s,a)] \\
&\quad + \eta\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\cdot (w^k - w_\star^k)] - \frac{M\eta^2}{2}\|w^k\|^2 \\
&= \eta\left(J^\star - J(\theta^k)\right) + \eta\frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\cdot (1-\gamma)w_\star^k - A^{\pi_{\theta^k}}(s,a)] \\
&\quad + \eta\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}[\nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\cdot (w^k - w_\star^k)] - \frac{M\eta^2}{2}\|w^k\|^2.
\end{aligned}$$

Now, let us apply Jensen's inequality and Assumption 4.2 to obtain

$$\begin{aligned}
&\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^k}(\cdot\,|\,s)\right) - \mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^{k+1}}(\cdot\,|\,s)\right)\right] \\
&\geq \eta\left(J^\star - J(\theta^k)\right) \\
&\quad - \eta\frac{1}{1-\gamma}\sqrt{\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\mathbb{E}_{a\sim\pi^\star(\cdot\,|\,s)}\left[\left(\nabla_\theta \log \pi_{\theta^k}(a\,|\,s)\cdot(1-\gamma)w_\star^k - A^{\pi_{\theta^k}}(s,a)\right)^2\right]} \\
&\quad - \eta G\|w^k - w_\star^k\| - \frac{M\eta^2}{2}\|w^k\|^2
\end{aligned}$$

Combining this with Assumption 4.4 yields

$$\begin{aligned}
&\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^k}(\cdot\,|\,s)\right) - \mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^{k+1}}(\cdot\,|\,s)\right)\right] \\
&\geq \eta\left(J^\star - J(\theta^k)\right) - \eta\sqrt{\frac{1}{(1-\gamma)^2}\varepsilon_{\text{bias}}} - \eta G\|w^k - w_\star^k\| - \frac{M\eta^2}{2}\|w^k\|^2.
\end{aligned} \tag{J.1}$$

Finally, let us telescope the above inequality from $k = 0$ to $K-1$, and divide by K, which gives

$$\begin{aligned}
J(\pi^\star) - \frac{1}{K}\sum_{k=0}^{K-1} J(\theta^k) \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} &+ \frac{1}{\eta K}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right] \\
&+ \frac{G}{K}\sum_{k=0}^{K-1}\|w^k - w_\star^k\| + \frac{M\eta}{2K}\sum_{k=0}^{K-1}\|w^k\|^2.
\end{aligned} \tag{J.2}$$

□

On the right hand side of (J.2), the first term reflects the function approximation error due to the possibly imperfect policy parametrization. The second term vanishes as $K \to \infty$.

By looking at the third and fourth term, we know that for an update of the form $\theta^{k+1} = \theta^k + \eta w^k$, its global convergence rate depends crucially on i) the difference between its update directions w^k and the exact NPG update direction $w_\star^k$, and ii) its stationary convergence rate.

In the rest of this paper, we shall see that for stochastic PG, NPG, SRVR-PG, and SRVR-NPG, both the third and fourth terms of (J.2) go to 0 as $K \to \infty$, whose speed lead to different global convergence rates for different algorithms. In order to achieve this, we will apply some intermediate results in the previous proof of stationary convergence.

K Proof of Theorem 4.6

Let us take w^k as the update direction of PG and apply Proposition 4.5. To this end, we need to upper bound $\frac{1}{K}\sum_{k=0}^{K-1}\|w^k - w_\star^k\|$, $\frac{1}{K}\sum_{k=0}^{K-1}\|w^k\|^2$, and $\frac{1}{K}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\mid s)||\pi_{\theta^0}(\cdot\mid s)\right)\right]$, where $w_\star^k = F^{-1}(\theta^k)\nabla J(\theta^k)$ is the exact NPG update direction at θ^k.

- Bounding $\frac{1}{K}\sum_{k=0}^{K-1}\|w^k - w_\star^k\|$.

  We know from Jensen's inequality and $\left(\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|]\right)^2 \leq \mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2]$ that

$$
\begin{aligned}
&\left(\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k - w_\star^k\|]\right)^2\\
&\leq \frac{1}{K}\sum_{k=0}^{K-1}\left(\mathbb{E}[\|w^k - w_\star^k\|]\right)^2\\
&\leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k - w_\star^k\|^2]\\
&\leq \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k - \nabla J(\theta^k)\|^2] + \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J(\theta^k) - F^{-1}(\theta^k)\nabla J(\theta^k)\|^2]
\end{aligned}
\tag{K.1}
$$

  Since

$$
w^k = \frac{1}{N}\sum_{i=1}^{N} g(\tau_i^H \mid \theta^k),
$$

  we have from Lemma B.1 and Assumption 4.1 that

$$
\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k - \nabla J(\theta^k)\|^2]\\
&\leq \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k - \nabla J^H(\theta^k)\|^2] + \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J^H(\theta^k) - \nabla J(\theta^k)\|^2]\\
&\leq 2\frac{\sigma^2}{N} + 2G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H}.
\end{aligned}
\tag{K.2}
$$

  Furthermore, Assumption 2.1 tells us that

$$
\begin{aligned}
&\mathbb{E}[\|\nabla J(\theta^k) - F^{-1}(\theta^k)\nabla J(\theta^k)\|^2]\\
&\leq \left(1 + \frac{1}{\mu_F}\right)^2\mathbb{E}[\|\nabla J(\theta^k)\|^2]\\
&\leq \left(1 + \frac{1}{\mu_F}\right)^2\left(2\mathbb{E}[\|\nabla J^H(\theta^k)\|^2] + 2G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H}\right).
\end{aligned}
\tag{K.3}
$$

Combining (K.2) and (K.3) with (K.1) gives

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k - w_\star^k\|] \\
&\leq \Bigg(2\frac{\sigma^2}{N} + 2G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H} \\
&\quad + 2\left(1+\frac{1}{\mu_F}\right)^2 \\
&\qquad \cdot \frac{1}{K}\sum_{k=0}^{K-1}\left(2\mathbb{E}[\|\nabla J^H(\theta^k)\|^2] + 2G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H}\right)\Bigg)^{0.5}
\end{aligned} \tag{K.4}$$

And recall from (F.2) that

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J^H(\theta^k)\|^2] \leq \frac{\frac{J^{H,\star}-J^H(\theta_0)}{K} + (\frac{\eta}{2}+L_J\eta^2)\frac{\sigma^2}{N}}{\frac{\eta}{2}-L_J\eta^2}.$$

Let us take $\eta = \frac{1}{4L_J}$. In addition, let H, N, and K satisfy

$$\begin{aligned}
\frac{1}{3}(\frac{\varepsilon}{3G})^2 &\geq \left(2+4(1+\frac{1}{\mu_F})^2\right)G^2R^2\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H} \\
N &\geq \frac{\left(2+12(1+\frac{1}{\mu_F})^2\right)\sigma^2}{\frac{1}{3}\left(\frac{\varepsilon}{3G}\right)^2}, \\
K &\geq \frac{\left(1+\frac{1}{\mu_F}\right)^2 64L_J(J^{H,\star}-J^H(\theta_0))}{\frac{1}{3}\left(\frac{\varepsilon}{3G}\right)^2}.
\end{aligned} \tag{K.5}$$

Then, we have

$$\frac{G}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k - w_\star^k\|] \leq \frac{\varepsilon}{3}. \tag{K.6}$$

- Bounding $\frac{1}{K}\sum_{k=0}^{K-1}\|w^k\|^2$.
  We have from (K.2) and (F.2) that

$$\begin{aligned}
&\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}\|w^k\|^2 \\
&\leq \frac{\sigma^2}{N} + \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J^H(\theta^k)\|^2] \\
&\leq \frac{\sigma^2}{N} + \frac{\frac{J^{H,\star}-J^H(\theta_0)}{K} + (\frac{\eta}{2}+L_J\eta^2)\frac{\sigma^2}{N}}{\frac{\eta}{2}-L_J\eta^2}.
\end{aligned}$$

Taking $\eta = \frac{1}{4L_j}$ and

$$\begin{aligned}
N &\geq \frac{12M\eta\sigma^2}{\varepsilon}, \\
K &\geq \frac{48L_JM\eta(J^{H,\star}-J^H(\theta_0))}{\varepsilon},
\end{aligned} \tag{K.7}$$

we arrive at

$$\frac{M\eta}{2K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k\|^2] \leq \frac{\varepsilon}{3}. \tag{K.8}$$

- Bounding $\frac{1}{K}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]$.
  By taking
$$K \geq \frac{3\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]}{\eta\varepsilon} \tag{K.9}$$
  we have
$$\frac{1}{\eta K}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right] \leq \frac{\varepsilon}{3}, \tag{K.10}$$

In summary, we require N and K to satisfy (K.5), (K.7), and (K.9), which leads to
$$N = \mathcal{O}\left(\frac{\sigma^2}{\varepsilon^2}\right), \qquad K = \mathcal{O}\left(\frac{1}{(1-\gamma)^2\varepsilon^2}\right), \qquad H = \mathcal{O}\left(\log((1-\gamma)^{-1}\varepsilon^{-1})\right).$$
By combining (K.6), (K.8), (K.10) and (J.2), we can conclude that
$$J(\pi^\star) - \frac{1}{K}\sum_{k=0}^{K-1} J(\theta^k) \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \varepsilon.$$
In total, stochastic PG requires to sample $KN = \mathcal{O}\left(\frac{\sigma^2}{(1-\gamma)^2\varepsilon^4}\right)$ trajectories.

L Proof of Theorem 4.9

Let us take w^k as the update direction of NPG and apply Proposition 4.5. To this end, we need to upper bound $\frac{1}{K}\sum_{k=0}^{K-1}\|w^k - w_\star^k\|$, $\frac{1}{K}\sum_{k=0}^{K-1}\|w^k\|^2$, and $\frac{1}{K}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]$, where $w_\star^k = F^{-1}(\theta^k)\nabla J(\theta^k)$ is the exact NPG update direction at θ^k.

Let us take $\eta = \frac{\mu_F^2}{4G^2 L_J}$ and apply SGD as in Procedure 1 to obtain a w^k that satisfies
$$\mathbb{E}[\|w^k - w_\star^k\|^2] \leq \min\Big\{\frac{\varepsilon}{12M\eta}, \left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2, \frac{\varepsilon}{12M\eta^3}\frac{\mu_F^2}{2}\frac{\frac{\eta}{2G^2} - \frac{L_J\eta^2}{\mu_F^2}}{\frac{G^2}{2\eta} + L_J}\Big\}. \tag{L.1}$$

From Proposition G.1, we know that this requires sampling $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon^2}\right)$ trajectories at each iteration.

- Bounding $\frac{1}{K}\sum_{k=0}^{K-1}\|w^k - w_\star^k\|$.
  Recall that the update direction $w^k \approx w_\star^k = F^{-1}(\theta^k)\nabla J(\theta^k)$ is obtained by solving the subproblem
$$\begin{aligned} w^k &\approx \operatorname*{argmin}_{w\in\mathbb{R}^{\mathrm{d}}} L_{\nu^{\pi_\theta}}(w;\theta^k) \\ &= \operatorname*{argmin}_{w\in\mathbb{R}^{\mathrm{d}}} \mathbb{E}_{(s,a)\sim\nu^{\pi_{\theta^k}}}\left[A^{\pi_{\theta^k}}(s,a) - (1-\gamma)w^\top\nabla_\theta\log\pi_{\theta^k}(a\,|\,s)\right]^2. \end{aligned}$$
  By (L.1) and Jensen's inequality, we can write
$$\left(\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k - w_\star^k\|]\right)^2 \leq \frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k - w_\star^k\|^2] \leq \left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2. \tag{L.2}$$
  On the other hand, by replacing (G.1) with (L.1), the stationary convergence of NPG stated in (G.3) becomes
$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J(\theta^k)\|^2] \leq \frac{\frac{J^\star - J(\theta_0)}{K} + \frac{\varepsilon}{12M\eta}\frac{\mu_F^2}{2}\left(\frac{\eta}{2G^2} - \frac{L_J\eta^2}{\mu_F^2}\right)}{\frac{\eta}{2G^2} - \frac{L_J\eta^2}{\mu_F^2}}$$

- Bounding $\frac{1}{K}\sum_{k=0}^{K-1}\|w^k\|^2$.

  Taking $\eta = \frac{\mu_F^2}{4G^2 L_J}$ and

$$K \geq \frac{24(J^\star - J(\theta_0))M\eta}{\mu_F^2\left(\frac{\eta}{2G^2} - \frac{L_J\eta^2}{\mu_F^2}\right)\varepsilon} \tag{L.3}$$

  gives us

$$\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J(\theta^k)\|^2] \leq \frac{\mu_F^2\varepsilon}{12M\eta}.$$

  (L.1) and the above inequality yields

$$\begin{aligned}\frac{1}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k\|^2] &\leq \frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|w^k - w_\star^k\|^2] + \frac{1}{\mu_F^2}\frac{2}{K}\sum_{k=0}^{K-1}\mathbb{E}[\|\nabla J(\theta^k)\|^2] \\ &\leq \frac{2\varepsilon}{12M\eta} + \frac{2}{\mu_F^2}\cdot\frac{\mu_F^2\varepsilon}{12M\eta} = \frac{\varepsilon}{3M\eta}.\end{aligned} \tag{L.4}$$

- Bounding $\frac{1}{K}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]$.

  Let us also set

$$K \geq \frac{3\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]}{\eta\varepsilon}, \tag{L.5}$$

  so that

$$\frac{1}{\eta K}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right] \leq \frac{\varepsilon}{3}. \tag{L.6}$$

In summary, we require K to satisfy (L.3) and (L.5), which leads to

$$K = \mathcal{O}\left(\frac{1}{(1-\gamma)^2\varepsilon}\right)$$

By combining (L.2), (L.4), (L.6) and (J.2), we can conclude that

$$J(\pi^\star) - \frac{1}{K}\sum_{k=0}^{K-1}J(\theta^k) \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \varepsilon.$$

Since at each iteration, SGD needs to sample $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon^2}\right)$ trajectories so that (L.1) is satisfied, NPG requires to sample $\mathcal{O}\left(\frac{1}{(1-\gamma)^6\varepsilon^3}\right)$ trajectories in total.

M Proof of Theorem 4.11

Let us take w_t^{j+1} as the update direction of SRVR-PG and apply Proposition 4.5. To this end, we need to upper bound $\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\|w_t^{j+1} - w_{t,\star}^{j+1}\|$, $\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\|w_t^{j+1}\|^2$, and $\frac{1}{Sm}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]$, where $w_{t,\star}^{j+1} = F^{-1}(\theta_t^{j+1})\nabla J(\theta_t^{j+1})$ is the exact NPG update direction at θ_t^{j+1}.

- Bounding $\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\|w_t^{j+1} - w_{t,\star}^{j+1}\|$.

Since $w_t^{j+1} = u_t^{j+1}$ and $w_{t,\star}^{j+1} = F^{-1}(\theta_t^{j+1})\nabla J(\theta_t^{j+1})$, we have from Lemmas I.1 and B.1 that

$$
\begin{aligned}
&\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2] \\
&\le 2\mathbb{E}[\|u_t^{j+1} - \nabla J(\theta_t^{j+1})\|^2] + 2\mathbb{E}[\|\nabla J(\theta_t^{j+1}) - F^{-1}(\theta_t^{j+1})\nabla J(\theta_t^{j+1})\|^2] \\
&\le 2\mathbb{E}[\|u_t^{j+1} - \nabla J^H(\theta_t^{j+1})\|^2] + 2(1+\frac{1}{\mu_F})^2\mathbb{E}[\|\nabla J(\theta_t^{j+1})\|^2] \\
&\quad + 2G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H} \\
&\le 2\left(\frac{C_\gamma}{B}\sum_{l=1}^{t}\mathbb{E}[\|\theta_l^{j+1} - \theta_{l-1}^{j+1}\|^2] + \frac{\sigma^2}{N}\right) + 2(1+\frac{1}{\mu_F})^2\mathbb{E}[\|\nabla J(\theta_t^{j+1})\|^2] \\
&\quad + 2G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H} \\
&\le 2\left(\frac{C_\gamma}{B}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2] + \frac{\sigma^2}{N}\right) + 2(1+\frac{1}{\mu_F})^2\mathbb{E}[\|\nabla J(\theta_t^{j+1})\|^2] \\
&\quad + 2G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H},
\end{aligned}
$$

where we have applied Lemma B.1 and Assumption 2.1 in the second inequality, and Lemma I.1 in the third one.
Telescoping this over $s = 0, 1, .., S-1$, $t = 0, 1, m-1$ and dividing by Sm gives

$$
\begin{aligned}
&\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2] \\
&\le 2(1+\frac{1}{\mu_F})^2\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\nabla J(\theta_t^{j+1})\|^2] \\
&\quad + 2\left(\frac{C_\gamma m}{B}\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2] + \frac{\sigma^2}{N}\right) \\
&\quad + 2G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H}.
\end{aligned}
\tag{M.1}
$$

On the other hand, from Equation (B.14) of [21] we know that

$$
\begin{aligned}
&\left(\frac{2}{\eta^2} - \frac{4L_J}{\eta} - \frac{12mC_\gamma}{B}\right)\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2] \\
&+ \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\nabla J^H(\theta_t^{j+1})\|^2] \\
&\le \frac{8(J^{H,\star} - J^H(\theta_0))}{\eta Sm} + \frac{6\sigma^2}{N}.
\end{aligned}
\tag{M.2}
$$

By the definition of C_γ in Lemma I.1, we have

$$
B = \frac{3\eta C_\gamma m}{L_J} = \frac{72\eta RG(2G^2+M)(W+1)\gamma}{L_J(1-\gamma)^5}m \tag{M.3}
$$

Therefore, (M.2) becomes

$$
\begin{aligned}
\left(\frac{2}{\eta^2} - \frac{8L_J}{\eta}\right)\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2] + \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\nabla J^H(\theta_t^{j+1})\|^2] \\
\le \frac{8(J^{H,\star} - J^H(\theta_0))}{\eta Sm} + \frac{6\sigma^2}{N},
\end{aligned}
$$

Since $\eta = \frac{1}{8L_J}$, we further have

$$\begin{aligned}
\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2] &\leq \frac{J^{H,\star} - J^H(\theta_0)}{L_J Sm} + \frac{6\sigma^2}{64L_J^2 N}, \\
\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\nabla J^H(\theta_t^{j+1})\|^2] &\leq \frac{64L_J(J^{H,\star} - J^H(\theta_0))}{Sm} + \frac{6\sigma^2}{N}.
\end{aligned} \tag{M.4}$$

Putting these inequalities back into (M.1) yields

$$\begin{aligned}
\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2] \leq{} & 2(1+\frac{1}{\mu_F})^2\left(\frac{64L_J(J^{H,\star} - J^H(\theta_0))}{Sm} + \frac{6\sigma^2}{N}\right) \\
& + 2\left(\frac{8L_J^2}{3}\left(\frac{J^{H,\star} - J^H(\theta_0)}{L_J Sm} + \frac{6\sigma^2}{64L_J^2 N}\right) + \frac{\sigma^2}{N}\right) \\
& + 2G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H}.
\end{aligned}$$

Let us set

$$\begin{aligned}
\frac{1}{3}\left(\frac{\varepsilon}{3G}\right)^2 &\geq 2G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H}, \\
N &\geq \frac{\left(12(1+\frac{1}{\mu_F})^2 + 2.5\right)\sigma^2}{\frac{1}{3}\left(\frac{\varepsilon}{3G}\right)^2}, \\
Sm &\geq \frac{\left(128(1+\frac{1}{\mu_F})^2 + \frac{16}{3}\right)L_J(J^{H,\star} - J^H(\theta_0))}{\frac{1}{3}\left(\frac{\varepsilon}{3G}\right)^2},
\end{aligned} \tag{M.5}$$

so that

$$\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2] \leq \left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2$$

By Jensen's inequality, we further have

$$\begin{aligned}
\left(\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|]\right)^2 &\leq \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\left(\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|]\right)^2 \\
. \qquad\qquad &\leq \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2] \\
&\leq \left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2.
\end{aligned} \tag{M.6}$$

where we have also applied $\left(\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|]\right)^2 \leq \mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2]$.

- Bounding $\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\|w_t^{j+1}\|^2$

Since $\mathbb{E}[w_t^{j+1}] = \mathbb{E}[u_t^{j+1}] = \nabla J^H(\theta_t^{j+1})$, by Lemma I.1 we have

$$
\begin{aligned}
&\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1}\|^2] \\
&= \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|u_t^{j+1} - \nabla J^H(\theta_t^{j+1})\|^2] + \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\nabla J^H(\theta_t^{j+1})\|^2] \\
&\leq \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\left(\frac{C_\gamma}{B}\sum_{l=1}^{t}\mathbb{E}[\|\theta_l^{j+1} - \theta_{l-1}^{j+1}\|^2] + \frac{\sigma^2}{N}\right) \\
&\quad + \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\nabla J^H(\theta_t^{j+1})\|^2] \\
&\leq \frac{C_\gamma m}{B}\cdot\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2] + \frac{\sigma^2}{N} + \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\nabla J^H(\theta_t^{j+1})\|^2].
\end{aligned}
$$

By setting $\eta = \frac{1}{8L_J}$ and applying (M.3) and (M.4), we further have

$$
\begin{aligned}
&\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1}\|^2] \\
&\leq \frac{8L_J^2}{3}\cdot\left(\frac{J^{H,\star} - J^H(\theta_0)}{L_J Sm} + \frac{6\sigma^2}{64L_J^2 N}\right) + \frac{\sigma^2}{N} + \frac{64L_J(J^{H,\star} - J^H(\theta_0))}{Sm} + \frac{6\sigma^2}{N}
\end{aligned}
$$

Therefore, we can set

$$
\begin{aligned}
N &\geq \frac{174M\sigma^2}{32L_J\varepsilon}, \\
Sm &\geq \frac{50M(J^{H,\star} - J^H(\theta_0))}{\varepsilon},
\end{aligned}
\tag{M.7}
$$

so that

$$
\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|w_t^{j+1}\|^2 \leq \frac{\varepsilon}{3M\eta}. \tag{M.8}
$$

- Bounding $\frac{1}{Sm}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]$.
  Let us set

$$
Sm \geq \frac{3\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]}{\eta\varepsilon} \tag{M.9}
$$

  so that

$$
\frac{1}{\eta Sm}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right] \leq \frac{\varepsilon}{3}. \tag{M.10}
$$

By combining (M.6), (M.8), (M.10) and (J.2), we can conclude that

$$
J(\pi^\star) - \frac{1}{K}\sum_{k=0}^{K-1}J(\theta^k) \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \varepsilon.
$$

To achieve this, we require Sm and N to satisfy (M.5), (M.7), and (M.9), which leads to

$$
Sm = \mathcal{O}\left(\frac{1}{(1-\gamma)^2\varepsilon^2}\right), \qquad N = \mathcal{O}\left(\frac{\sigma^2}{\varepsilon^2}\right), \qquad H = \mathcal{O}\left(\log(\frac{1}{(1-\gamma)\varepsilon})\right).
$$

By (M.3), we know that $B = \mathcal{O}(W(1-\gamma)^{-1}m)$.

Therefore, by taking $S = \mathcal{O}\left(\frac{1}{(1-\gamma)^{2.5}\varepsilon}\right)$ and $m = \mathcal{O}\left(\frac{1}{(1-\gamma)^{-0.5}\varepsilon}\right)$, the sample complexity of SRVR-PG is

$$\begin{aligned} S(N+mB) &= \mathcal{O}\left(\frac{\sigma^2}{(1-\gamma)^{2.5}\varepsilon^3} + \frac{W}{(1-\gamma)^{2.5}\varepsilon^3}\right) \\ &= \mathcal{O}\left(\frac{W+\sigma^2}{(1-\gamma)^{2.5}\varepsilon^3}\right). \end{aligned}$$

N Proof of Theorem 4.13

Let us take w_t^{j+1} as the update direction of SRVR-NPG and apply Proposition 4.5. To this end, we need to upper bound $\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\|w_t^{j+1} - w_{t,\star}^{j+1}\|$, $\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\|w_t^{j+1}\|^2$, and $\frac{1}{Sm}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]$, where $w_{t,\star}^{j+1} = F^{-1}(\theta_t^{j+1})\nabla J(\theta_t^{j+1})$ is the exact NPG update direction at θ_t^{j+1}.

Let us take $\eta = \frac{\mu_F}{16L_J}$ and apply SGD as in Procedure 2 to obtain a w_t^{j+1} that satisfies

$$\mathbb{E}[\|w_t^{j+1} - F^{-1}(\theta_t^{j+1})u_t^{j+1}\|^2] \le \min\left\{\frac{\frac{1}{4}\left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2}{2 + \frac{G^2\mu_F + G^4}{\mu_F^2}\cdot\frac{\mu_F}{4G^2+\mu_F}}, \frac{64\eta^2L_J^2}{\mu_F(\mu_F+G^2)}\cdot\frac{\varepsilon}{9M\eta}\right\}. \tag{N.1}$$

In order to apply Proposition I.2, let assume the following so that its assumptions are satisfied:

$$\begin{aligned} \frac{\sigma^2}{N} &\le \left(\frac{GR}{(1-\gamma)^2}\right)^2, \\ \min\left\{\frac{\frac{1}{4}\left(\frac{1-\gamma}{G}\frac{\varepsilon}{3}\right)^2}{2 + \frac{G^2\mu_F + G^4}{\mu_F^2}\cdot\frac{\mu_F}{4G^2+\mu_F}}, \frac{64\eta^2L_J^2}{\mu_F(\mu_F+G^2)}\cdot\frac{(1-\gamma)\varepsilon}{9M\eta}\right\} &\le \frac{2}{\mu_F^2}\left(\frac{GR}{(1-\gamma)^2}\right)^2 \\ \frac{C_\gamma m}{B}2\eta^2 &\le \frac{1}{3}\mu_F^2. \end{aligned} \tag{N.2}$$

At the end of this proof, we will see that these assumptions are indeed satisfied for small ε.

From Proposition I.2, we know that this requires sampling $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon^2}\right)$ trajectories at each iteration.

- Bounding $\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\|w_t^{j+1} - w_{t,\star}^{j+1}\|$.
  First of all, we have

$$\begin{aligned} &\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2] \\ &\le 2\mathbb{E}[\|w_t^{j+1} - F^{-1}(\theta_t^{j+1})u_t^{j+1}\|^2] + 2\mathbb{E}[\|F^{-1}(\theta_t^{j+1})u_t^{j+1} - F^{-1}(\theta_t^{j+1})\nabla J(\theta_t^{j+1})\|^2] \\ &\le 2\mathbb{E}[\|w_t^{j+1} - F^{-1}(\theta_t^{j+1})u_t^{j+1}\|^2] + 2\frac{1}{\mu_F^2}\mathbb{E}[\|u_t^{j+1} - \nabla J(\theta_t^{j+1})\|^2] \\ &\le 2\mathbb{E}[\|w_t^{j+1} - F^{-1}(\theta_t^{j+1})u_t^{j+1}\|^2] + 4\frac{1}{\mu_F^2}\left(\frac{C_\gamma}{B}\sum_{l=1}^{t}\mathbb{E}[\|\theta_l^{j+1} - \theta_{l-1}^{j+1}\|^2] + \frac{\sigma^2}{N}\right) \\ &\quad + 4G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H} \\ &\le 2\mathbb{E}[\|w_t^{j+1} - F^{-1}(\theta_t^{j+1})u_t^{j+1}\|^2] + 4\frac{1}{\mu_F^2}\left(\frac{C_\gamma}{B}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2] + \frac{\sigma^2}{N}\right) \\ &\quad + 4G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H}, \end{aligned}$$

  where we have applied Assumption 2.1 in the second inequality, and Lemmas I.1 and B.1 in the third one.

Telescoping this over $s = 0, 1, .., S-1$, $t = 0, 1, m-1$ and dividing by Sm gives

$$
\begin{aligned}
&\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1}-w_{t,\star}^{j+1}\|^2] \\
&\leq 2\frac{\frac{1}{4}\left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2}{2+\frac{G^2\mu_F+G^4}{\mu_F^2}\cdot\frac{\mu_F}{4G^2+\mu_F}} + 4\frac{1}{\mu_F^2}\left(\frac{C_\gamma m}{B}\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1}-\theta_t^{j+1}\|^2]+\frac{\sigma^2}{N}\right) \\
&\quad + 4G^2R^2\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H}.
\end{aligned} \tag{N.3}
$$

On the other hand, from (I.5) we know that

$$
\begin{aligned}
&\frac{8G^2}{\eta}\left(\frac{\mu_F}{8\eta}-\frac{L_J}{2}-\left(\frac{\eta}{\mu_F}+\frac{\eta}{4G^2}\right)\frac{C_\gamma m}{B}\right)\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\theta_{t+1}^{j+1}-\theta_t^{j+1}\|^2 \\
&\quad+\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\nabla J^H(\theta_t^{j+1})\|^2 \\
&\leq \frac{8G^2}{\eta}\frac{J^{H,\star}-J^H(\theta_0)}{Sm} \\
&\quad+\left(\frac{8G^2}{\mu_F}+2\right)\frac{\sigma^2}{N}+\left(\frac{8G^2\mu_F}{4}+\frac{8G^4}{4}\right)\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|F^{-1}(\theta_t^{j+1})u_t^{j+1}-w_t^{j+1}\|^2.
\end{aligned} \tag{N.4}
$$

Let us set

$$
B \geq \left(\frac{\eta}{\mu_F}+\frac{\eta}{4G^2}\right)\frac{2C_\gamma m}{L_J} = \left(\frac{\eta}{\mu_F}+\frac{\eta}{4G^2}\right)\cdot\frac{48RG^2(2G^2+M)(W+1)\gamma}{L_J(1-\gamma)^5}m. \tag{N.5}
$$

Since $\eta = \frac{\mu_F}{16L_J}$, (N.4) becomes

$$
\begin{aligned}
&\frac{8G^2L_J}{\eta}\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\theta_{t+1}^{j+1}-\theta_t^{j+1}\|^2+\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|\nabla J^H(\theta_t^{j+1})\|^2 \\
&\leq \frac{8G^2}{\eta}\frac{J^{H,\star}-J^H(\theta_0)}{Sm}+\left(\frac{8G^2}{\mu_F}+2\right)\frac{\sigma^2}{N} \\
&\quad+\left(\frac{8G^2\mu_F}{4}+\frac{8G^4}{4}\right)\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|F^{-1}(\theta_t^{j+1})u_t^{j+1}-w_t^{j+1}\|^2,
\end{aligned}
$$

from which we have

$$
\begin{aligned}
&\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1}-\theta_t^{j+1}\|^2] \\
&\leq \frac{J^{H,\star}-J^H(\theta_0)}{L_JSm}+\left(\frac{8G^2}{\mu_F}+2\right)\frac{\mu_F}{128G^2L_J^2}\frac{\sigma^2}{N} \\
&\quad+\left(\frac{8G^2\mu_F}{4}+\frac{8G^4}{4}\right)\frac{\mu_F}{128G^2L_J^2}\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|F^{-1}(\theta_t^{j+1})u_t^{j+1}-w_t^{j+1}\|^2.
\end{aligned} \tag{N.6}
$$

Putting these inequalities back into (N.3) and applying (N.1) yields

$$
\begin{aligned}
&\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1}-w_{t,\star}^{j+1}\|^2] \\
&\le 2\frac{\frac{1}{4}\left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2}{2+\frac{G^2\mu_F+G^4}{\mu_F^2}\cdot\frac{\mu_F}{4G^2+\mu_F}} \\
&\quad + 4\frac{1}{\mu_F^2}\left(\frac{C_\gamma m}{B}\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1}-\theta_t^{j+1}\|^2]+\frac{\sigma^2}{N}\right) \\
&\quad + 4G^2R^2\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H} \\
&\le 2\frac{\frac{1}{4}\left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2}{2+\frac{G^2\mu_F+G^4}{\mu_F^2}\cdot\frac{\mu_F}{4G^2+\mu_F}} \\
&\quad + 2\frac{1}{\mu_F^2}\frac{C_\gamma m}{B}\left(\frac{J^\star-J(\theta_0)}{L_J Sm}+\left(\frac{8G^2}{\mu_F}+2\right)\frac{\mu_F}{128G^2L_J^2}\frac{\sigma^2}{N}\right. \\
&\qquad\qquad \left. + \left(G^2\mu_F+G^4\right)\frac{\mu_F}{64G^2L_J^2}\frac{\frac{1}{3}\left(\frac{1-\gamma}{G}\frac{\varepsilon}{3}\right)^2}{1+\frac{G^2\mu_F+G^4}{\mu_F^2}\cdot\frac{\mu_F}{4G^2+\mu_F}}\right) \\
&\quad + 2\frac{1}{\mu_F^2}\frac{\sigma^2}{N}+4G^2R^2\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H}.
\end{aligned}
$$

From (N.5) we know that

$$
\frac{C_\gamma m}{B}\le\frac{32L_J^2}{4+\frac{\mu_F}{G^2}},
$$

which gives us

$$
\begin{aligned}
&\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1}-w_{t,\star}^{j+1}\|^2] \\
&\le \left(2+\frac{G^2\mu_F+G^4}{\mu_F^2}\cdot\frac{\mu_F}{4G^2+\mu_F}\right)\frac{\frac{1}{4}\left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2}{2+\frac{G^2\mu_F+G^4}{\mu_F^2}\cdot\frac{\mu_F}{4G^2+\mu_F}} \\
&\quad + 2\frac{1}{\mu_F^2}\frac{C_\gamma m}{B}\frac{J^{H,\star}-J^H(\theta_0)}{L_J Sm} \\
&\quad + 2\frac{1}{\mu_F^2}\left(1+(\frac{8G^2}{\mu_F}+2)\frac{\mu_F}{4(4G^2+\mu_F)}\right)\frac{\sigma^2}{N} \\
&\quad + 4G^2R^2\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H} \\
&= \frac{1}{4}\left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2+2\frac{1}{\mu_F^2}\frac{C_\gamma m}{B}\frac{J^{H,\star}-J^H(\theta_0)}{L_J Sm} \\
&\quad + 3\frac{1}{\mu_F^2}\frac{\sigma^2}{N}+4G^2R^2\left(\frac{H+1}{1-\gamma}+\frac{\gamma}{(1-\gamma)^2}\right)^2\gamma^{2H},
\end{aligned}
$$

where we have applied (N.1) in the first equality.

Let us set

$$
\begin{aligned}
N &\geq \frac{108G^2\sigma^2}{\mu_F^2\varepsilon^2}, \\
B &\geq \frac{72C_\gamma m}{\mu_F L_J^2 \varepsilon}, \\
Sm &\geq \frac{L_J(J^{H,\star} - J^H(\theta_0))G^2}{\mu_F \varepsilon}, \\
\frac{1}{4}\left(\frac{\varepsilon}{3G}\right)^2 &\geq 4G^2R^2\left(\frac{H+1}{1-\gamma} + \frac{\gamma}{(1-\gamma)^2}\right)^2 \gamma^{2H},
\end{aligned}
\tag{N.7}
$$

so that

$$
\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2] \leq \left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2
$$

By Jensen's inequality, we further have

$$
\begin{aligned}
\left(\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|]\right)^2 &\leq \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\left(\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|]\right)^2 \\
. \qquad &\leq \frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2] \\
&\leq \left(\frac{1}{G}\frac{\varepsilon}{3}\right)^2.
\end{aligned}
\tag{N.8}
$$

where we have also applied $\left(\mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|]\right)^2 \leq \mathbb{E}[\|w_t^{j+1} - w_{t,\star}^{j+1}\|^2]$.

- Bounding $\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\|w_t^{j+1}\|^2$.
  By (N.6) we have

$$
\begin{aligned}
&\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|w_t^{j+1}\|^2] \\
&= \frac{1}{Sm}\frac{1}{\eta^2}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}[\|\theta_{t+1}^{j+1} - \theta_t^{j+1}\|^2] \\
&\leq \frac{J^\star - J(\theta_0)}{L_J\eta^2 Sm} + \left(\frac{8G^2}{\mu_F} + 2\right)\frac{\mu_F}{128G^2\eta^2L_J^2}\frac{\sigma^2}{N} \\
&\quad + \left(\frac{8G^2\mu_F}{4} + \frac{8G^4}{4}\right)\frac{\mu_F}{128\eta^2G^2L_J^2}\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|F^{-1}(\theta_t^{j+1})u_t^{j+1} - w_t^{j+1}\|^2 \\
&\leq \frac{J^\star - J(\theta_0)}{L_J\eta^2 Sm} + \left(\frac{8G^2}{\mu_F} + 2\right)\frac{\mu_F}{128G^2\eta^2L_J^2}\frac{\sigma^2}{N} + \frac{\varepsilon}{9M\eta},
\end{aligned}
$$

  where we have applied (N.6) in the first inequality, and (N.1) in the last step.
  We can set

$$
\begin{aligned}
N &\geq \frac{9M\mu_F(\frac{8G^2}{\mu_F} + 2)\sigma^2}{128G^2\eta L_J^2\varepsilon}, \\
Sm &\geq \frac{9M(J^\star - J(\theta_0))}{L_J\eta\varepsilon},
\end{aligned}
\tag{N.9}
$$

  so that

$$
\frac{1}{Sm}\sum_{s=0}^{S-1}\sum_{t=0}^{m-1}\mathbb{E}\|w_t^{j+1}\|^2 \leq \frac{\varepsilon}{3M\eta}. \tag{N.10}
$$

- Bounding $\frac{1}{Sm}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]$.
  Let us set
$$Sm \geq \frac{3\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right]}{\eta\varepsilon} \tag{N.11}$$
  so that
$$\frac{1}{\eta Sm}\mathbb{E}_{s\sim d_\rho^{\pi^\star}}\left[\mathrm{KL}\left(\pi^\star(\cdot\,|\,s)||\pi_{\theta^0}(\cdot\,|\,s)\right)\right] \leq \frac{\varepsilon}{3}. \tag{N.12}$$

By combining (N.8), (N.10), (N.12) and (J.2), we can conclude that
$$J(\pi^\star) - \frac{1}{K}\sum_{k=0}^{K-1} J(\theta^k) \leq \frac{\sqrt{\varepsilon_{\text{bias}}}}{1-\gamma} + \varepsilon.$$

To achieve this, we require Sm, B, and N to satisfy (N.5), (N.7), (N.9), and (N.11), which leads to
$$\begin{aligned} Sm &= \mathcal{O}\left(\frac{1}{(1-\gamma)^2\varepsilon}\right), \qquad N = \mathcal{O}\left(\frac{\sigma^2}{\varepsilon^2}\right), \\ B &= \mathcal{O}\left(\frac{W}{(1-\gamma)\varepsilon}m\right), \qquad H = \mathcal{O}\left(\log\left(\frac{1}{(1-\gamma)\varepsilon}\right)\right). \end{aligned}$$

By Proposition I.2, we know that in order to achieve (N.1), SGD requires sampling $\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon^2}\right)$ trajectories per iteration.

Therefore, by taking $S = \mathcal{O}\left(\frac{1}{(1-\gamma)^{2.5}\varepsilon^{0.5}}\right)$ and $m = \mathcal{O}\left(\frac{(1-\gamma)^{0.5}}{\varepsilon^{0.5}}\right)$, the amount of trajectories required by SRVR-NPG is
$$\begin{aligned} &S\left(N + mB + (1+m)\mathcal{O}\left(\frac{1}{(1-\gamma)^4\varepsilon^2}\right)\right) \\ &= \mathcal{O}\left(\frac{\sigma^2}{(1-\gamma)^{2.5}\varepsilon^{2.5}} + \frac{W}{(1-\gamma)^{2.5}\varepsilon^{2.5}} + \frac{1}{(1-\gamma)^6\varepsilon^3}\right). \end{aligned}$$

It is straightforward to verify that the requirements listed in (N.2) are also satisfied as long as ε is small enough.

O Implementation Details

In this section, we provide additional details on the implementation of PG, NPG, SRVR-PG and SRVR-NPG.

1. For NPG, we use the default implementation provided by rllab[1], which actually implements the trust region policy optimization(TRPO) algorithm [5]. For cartplole, we sample 200 trajectories at each iteration to solve the subproblem of TRPO. For mountain car, we sample 120 trajectories at each iteration.
2. We found that the naive implementation of PG and SRVR-PG typically do not work for our tests. For example, PG and SRVR-PG often give an average reward around -90 for the mountain-car test, despite of our best efforts.
3. As in [19] and [21], we found that it is necessary to apply Adagrad [62] or Adam [63] type of averaging to improve their performances.
4. In our experiments, we apply Adagrad type of averaging for PG and SRVR-PG, which results in much better performances. As for SRVR-NPG, we apply Adam type of averaging, which gives an approximation of the Fisher information matrix at each iteration (see section 11.2 of [53]). We leave the implementation of a better approximation of the Fisher information matrix to the future work.

[1] `https://github.com/rll/rllab`